PhD Thesis

Investigations on the Role of Expectations and Individual Decisions in Quality Perception

Andreas Sackl

February 14, 2016

Technical University of Berlin
Faculty IV
Institute of Telecommunicationsystems
Assessment of IP-based Applications

Berlin, den February 14, 2016

[Andreas Sackl]

Acknowledgements

First of all, I want to thank Raimund Schatz for having been my FTW-internal PhD supervisor and mentor during the last years. His continuous encouragement, his valuable advice and the fruitful discussions over the years have supported me during the challenging phase of achieving my PhD.

I want to thank Peter Reichl for being my academic supervisor at the University of Vienna at the beginning of my research. Furthermore, I want to express my gratitude to Alexander Raake, who became my academic supervisor after I changed to TU Berlin. He has supported my research in every possible way especially with the opportunity of being a guest researcher at TU Berlin in summer 2015.

I want to thank my FTW colleagues Sebastian Egger, Dietmar Schabus, Patrick Zwickl, Pedro Casas, Bruno Gardlo, Kathrin Masuch, Ronny Fischer, Theresa Stürmer and Janina Gospodarek who provided me with outstanding support and fruitful discussions. I really enjoyed my time at FTW, working together in such an excellent team. I will always benefit from the personal experiences I was able to gain during that time.

I am particularly grateful for the support provided by my colleagues at TU Berlin during my research stay in summer 2015. I would like to thank Michal Solducha, Hagen Wirstorf, Sebastian Möller and Jan-Niklas Antons for integrating me in their daily lives at university and for all the great fun we had outside of work.

I would also like to offer my special thanks to Sebastian Möller and Lea Skorin-Kapov for serving as reviewers of this thesis.

I would also like to express my gratitude to my longstanding friends who provided welcome distractions in tough times, reliable support and words of cheer during the last years, thanks to you guys! Additionally, I thank Hannes Loch for proof-reading this thesis.

And of course I want thank my parents, my sister and my grandmother who supported me with their love, their imperturbable patience and assistance during the last years.

Abstract

In the scientific field of Quality of Experience (QoE) the impact of technical influencing factors — often referred as Quality of Service (Qos) — and additional context factors like applications, usage scenarios, emotional state, etc. on the subjective quality assessment is examined. Nevertheless, two relevant aspects have been neglected so far in QoE related research: user expectations and user decisions. Hence, a literature study is presented which explains how expectations are addressed in related research fields and how individual expectations can be quantified. Based on that, four empirical user studies are discussed which show how to trigger quality-related expectations in a laboratory setup and which methodological aspects need to be considered. Based on the concept of desired and adequate expectations [11], questionnaires have been used to measure individual expectations in the context of video on demand and telecommunication providers. This information is furthermore used to enhance the prediction accuracy of quantitative QoE models up to 12%. A failure to fulfill quality-related user expectations may entail certain user decisions, e.g., the technical video quality of a streamed video is increased, which also includes higher fees. For this reason, three user studies are discussed which have examined the impact of economic user decisions on QoE. In these studies the test participants were able to spend real money for enhanced video quality. According to the discussed findings, the individual decision regarding quality and price has a positive impact on subjective quality assessment. Moreover, recommendations are presented which can be used for further QoE experiments including economic decisions. To investigate the impact of non-economic decisions on QoE assessment, three user studies are discussed which focused on the impact of individual decisions regarding quality impairment, end device and content. So far, the ambiguous results of these three user studies make it impossible to draw final conclusions regarding the impact of non-economic user decisions on quality assessment. Finally, both user related QoE-aspects — user expectations and user decisions — are presented in the framework of the quality formation process [12] which describes the interaction between expectations, decisions and subjective quality assessment.

Kurzfassung

Im Forschungsgebiet Quality of Experience (QoE) wird der Einfluss von technischen Qualitätsaspekten — meistens als Quality of Service (QoS) bezeichnet — und zusätzlichen Einflussfaktoren bezogen auf Anwendungen, Einsatzszenarien, Emotionen, etc. auf das subjektiv wahrgenommene Qualitätsempfinden untersucht. Nichtsdestotrotz, zwei relevante Aspekte wurden bisher vernachlässigt: individuelle Erwartungshaltungen und Entscheidungen. Anhand einer Literaturstudie wird daher aufgezeigt, wie Erwartungshaltungen in benachbarten Forschungsgebieten untersucht und gemessen werden. Darauf aufbauend werden vier Benutzerstudien vorgestellt, in welchen spezifische Erwartungshaltungen ausgelöst wurden und es wird auf methodische Herausforderungen hingewiesen. Basierend auf dem Konzept von gewünschten und angemessenen Erwartungshaltungen [13], werden Fragebögen vorgestellt, welche eingesetzt wurden um individuelle Erwartungshaltungen hinsichtlich Videodienst- und Telekommunikationsanbietern zu erheben. Anhand dieser Informationen kann die Genauigkeit von quantitativen QoE Modellen um bis zu 12% erhöht werden. Werden qualitätsbezogene Erwartungshaltungen nicht erfüllt kann diese zu einer Entscheidung seitens des Benutzers führen, beispielsweise kann die Qualität eines Videostreams erhöht werden was mit zusätzlichen Kosten verbunden sein kann. Um dies näher zu untersuchen, wurden drei Benutzerstudien durchgeführt in welchen die Teilnehmer echtes Geld ausgeben konnten um die Videoqualität eines abgespielten Videos zu erhöhen. Wie gezeigt wird, haben diese Art von Entscheidungen einen positiven Einfluss auf die Qualitätsbeurteilung. Weiters werden methodologische Empfehlungen diskutiert um weiterführende Studie erfolgreich durchführen zu können. Anschließend werden drei Benutzerstudien besprochen, welche den Einfluss von nicht-ökonomischen Entscheidungen auf das Qualitätsempfinden untersuchen. In diesen drei Studien konnten sich die Versuchsteilnehmer hinsichtlich Qualitätsbeeinträchtigung, Endgerät und Inhalt entscheiden. Die mehrdeutigen Studien- ergebnisse ermöglichen jedoch kein eindeutiges Fazit hinsichtlich des Einflusses von nicht-ökonomischen Entscheidungen auf das Qualitätsempfinden. Beide benutzerrelevante QoE-Aspekte — individuelle Erwartungshaltungen und Entscheidungen — können zusätzlich in das bekannte Quality formation process Modell [12] einfügt werden, um die Interaktion zwischen Erwartungshaltungen, Entscheidungen und Qualitätswahrnehmung zu erklären.

Contents

List of Figures

List of Tables

Glossary

ACR Absolute category rating

ADSL Asymmetric digital subscriber line

AQoS Application quality of system

CDF Cumulative distribution function

FPS frames per second

HD High definition (video) quality

HLS HTTP live streaming

IMDb Internet movie database

ISP Internet service provider

ITU International telecommunication union

LTE Long term evolution

MDS-K Münchener Dissonanzskala

MOS Mean opinion score

MSE Mean squared error

PSNR Peak signal-to-noise ratio

QoE Quality of Experience

QoS Quality of Service

RMSE Root mean square error

SD Standard deviation/Standard (video) quality

SSIM Structural similarity

UDP User datagram protocol

UMTS Universal mobile telecommunications system

VBR Video bit rate

VoD Video on demand

WTP Willingness to pay

1 Introduction

During the last years and decades, various IP-based applications and services have penetrated our everyday lives more and more: it is common to perform HD-video conference calls, video sharing websites provide 4k videos with high frame-rates, mobile Internet connections enable high-speed and low-latency access to network functionalities, etc. Related to these digital ecosystems [15], several aspects are being examined, evaluated and enhanced on an ongoing basis, like storage and process-capacities, reliability, usability, etc. In this context, the term *quality* plays a critical role as a highly relevant concept.

In general, the expression *quality* is not a sharply defined term due its usage in several contexts like general human experiencing [16], experiencing of sound [17], technical aspects of a system — in the International Telecommunication Union (ITU) Recommendation E.800 [18] quality is defined as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs" —, quality management systems [19], etc. As discussed in [12], the term *quality* relates to more than just describable, technical aspects of a product or a service. Martens and Martens [20] point out two contrasting views on quality: first, the objective, product-related and rationalistic view and second the subjective, user-oriented and perceptual view. Whereas the objective view on quality deals with measurable and quantifiable characteristics and properties, the subjective view on quality focuses on the experience of the user, which includes the user's evaluation of the experience in terms of "evaluated excellence or goodness" [20]. Apparently, the second view involves apparently more aspects than purely technical ones, e.g., the usage context, the intention of the user, etc.[1].

Within the framework of this thesis, two quality-related concepts are crucial: *Quality of Service* (QoS) and *Quality of Experience* (QoE). Despite its technical focus, the ITU-T definition describes Quality of Service as the "totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service" [18], i.e., the user is explicitly mentioned. Nevertheless, in the concerned research community QoS-parameters relate to technical aspects of the underlying system like network bandwidth [21], video encoding bitrate [22], network packet loss [23], etc. In contrast to this, Quality of Experience relates more to a user-centric view on quality. Nevertheless, also QoE depends strongly on technical aspects, which are integrated in the concept of QoE as the main influence factor causing good or bad user experience. The Qualinet white paper defines Quality of Experience as "the degree of delight or annoyance of the user of an application or service. It results from

[1] For a comprehensive discussion of the term *quality* and its implications please see [12].

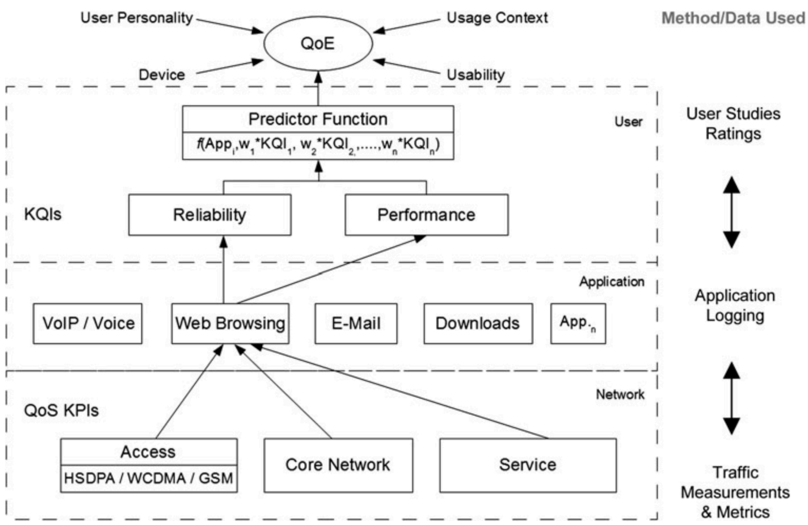


Figure 1.1: QoE/networked service framework from [14].

.

the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the users personality and current state" [24].

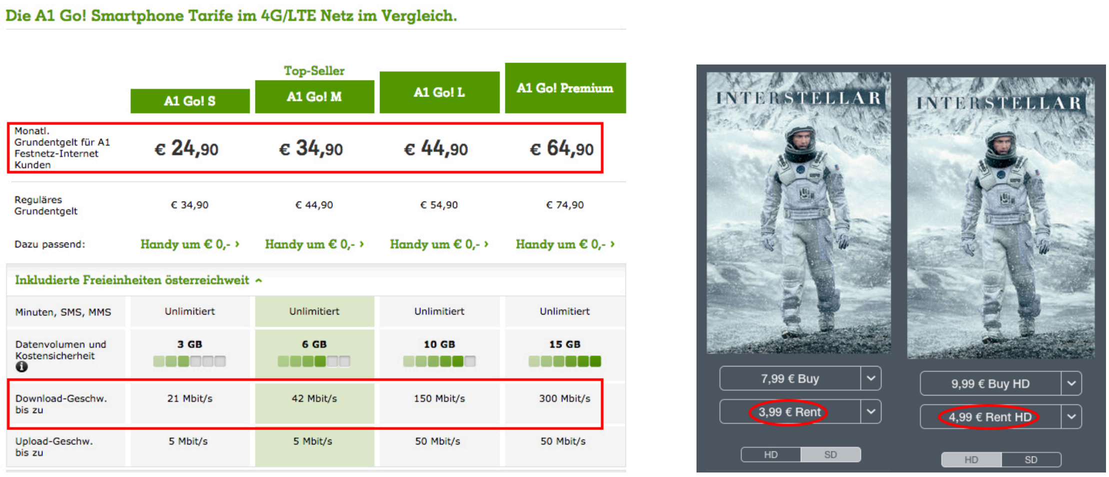
As described in [25], not only technical aspects — like the previously mentioned QoS parameters — impact the subjectively perceived service quality, also user factors (e.g. demographic and socio-demographic user background, emotions, tasks, etc.) and context factors (e.g. time of day, memory effects, usage alone vs. in groups, etc.) are highly relevant. Therefore, to understand, describe and assess Quality of Experience several aspects need to be considered. For example, Figure 1.1 from [14] depicts how various factors influence the quality perception. Nevertheless, two relevant aspects have been neglected so far in QoE related research: (1) user expectations and (2) user decisions.

(1) user expectations: The subjective assessment of quality depends on a comparison between *desired* quality features and *perceived* quality features, see Qualinet White paper definition of QoE [24]: "... fulfillment of his or her expectation ...". In general, a quality feature is the perceived characteristic of an entity "that is relevant to the entity´s quality" [17]. So, certain quality-related expectations are set and the subjective experience depends, among others, on the fulfillment of these expectations. These quality-related expectations refer to a certain context and to subjective attitudes. For example, a discerning, audiophile user would expect high sound quality while listing to a hi-fi system, i.e., a high quality of experience is only achieved if these high expectations are fulfilled. In contrast to this, the same user might accept low sound quality while listing to a simple radio doing kitchen work, i.e, in this particular situation only low expectations regarding audio quality need to be fulfilled. So far, in current QoE-related research the *perceived* experience has been in focus and the *desired* features — which are crucial for the subjective quality judgement — have mostly been implicitly covered.

Hence, it is necessary to examine how information about individual expectations can be systematically gathered and included in QoE research.

(2) user decisions: The subjectively perceived quality is not experienced in a static context, i.e., the affected user usually interacts in several ways with the system. Interactions might be short and occur frequently, e.g., browsing a news site includes clicking on links and scrolling through photo galleries [26]. Besides being more or less casual and frequent, interactions can also be more far-reaching and more conscious. For example, a user has to choose between a smartphone, a tablet or a PC to access a certain website or she has to decide in which video quality a Video on Demand (VoD) movie is consumed, which additionally involves an economic decision (in general a user has to pay more to receive a better video quality, see Figure 1.2(b)). Actually, in many quality-related situations economic aspects are highly relevant and this kind of interaction should not be neglected. For example, the authors of [27] stated that the perceived audio quality of a phone conversation depends on several context-factors, e.g., if a free or a paid telephone plan is used, which also triggers different expectations regarding quality. In [28] the relevance of several aspects like user segments, device characteristics, etc. for the usage of mobile TV are discussed. According to the authors also financial aspects should be included in research, i.e, most of the survey participants were not willing to pay (much) for high quality mobile TV. Also, the authors of [29] stated that costs — paid by the users — influence the quality perception process, but in the aforementioned QoE user study this aspect is neglected. Figure 1.2 shows two examples which demonstrate how technical quality is related to pricing in existing commercial products[2]. Obviously, higher technical quality is more expensive than lower technical quality, e.g., if someone pays 24.90 € per month the maximum downlink bandwidth is limited to 21 Mbit/s and if someone pays 64.90 € per month the maximum downlink bandwidth is limited to 300 Mbit/s, see Figure 1.2(a). To make this explicit, let us assume that a common mobile Web QoE laboratory user study concludes that for a certain task — e.g. downloading a large file — a downlink bandwidth of 21 Mbit/s leads to a MOS (Mean Opinion Score) of 4 (=good). So far, it is unclear if the resulting MOS would differ if pricing aspects were included: maybe users who pay 24.90 € would rate a MOS of 3.8 and users who pay 64.90 € would rate a MOS of 2.5 because of different expectations set by different fees. Only little research exists which empirically correlates economic aspects and quality assessment, hence this gap is addressed in this thesis. One should also consider that a mismatch between user expectations and actual quality perception (see above) could also trigger decisions which subsequently impact the subjective quality assessment. Therefore, both aspects – user decisions and user expectations – are connected and cannot be examined completely separately.

[2] http://www.a1.net/handys-telefonie/sprechen-sms-surfen, last access: 22nd July 2015



(a) Mobile price plans of the Austrian Provider A1 (July 2015)

(b) Renting a movie from Apple's iTunes store (July 2015)

Figure 1.2: Examples of pricing and quality.

1.1 Conceptual Model of User Expectations & Decisions for QoE

At the beginning of this chapter the relevance of expectations and individual user decisions in the context of QoE-related research and the fact that there is a lack of systematical research dealing with these two influencing factors were discussed. To evaluate these two factors empirically via user studies (see Section 1.3) it is necessary to discuss how individual user expectations and decisions could be included in models of mental processes which cover quality perception and assessment.

For this reason the so-called *quality formation process*, which is discussed by the authors of [12] and [17], will be investigated in this section and subsequently be extended by information about individual user expectations and user decisions. Figure 1.3 depicts a reduced model, i.e., only relevant aspects are covered, which are necessary to include further user decisions and user expectations. For a complete and more detailed description please see [12]. Blue elements in Figure 1.3 were added by the author of this thesis. These elements relate to user expectations and decisions and are not discussed in detail in the original quality formation process of [12].

The main input of the whole process depicted in Figure 1.3 is the "signal", which represents any sensory stimulus perceived by a person, for example an audio/video signal. The main output of the whole process in the end is a "quality rating", which is for example the statement of a laboratory test participant that she would accept the perceived technical quality of a certain video at home. Hence, the quality formation process explains which mental steps are involved to transfer an input signal to a subjective quality rating.

In general, the ”perception process” depicted in Figure 1.3 is triggered by ”signals” (for example an audiovisual signal) and impacted by various external factors, e.g., contextual information representing the use case of consuming a video at home. In the perception process ”... physical representations of the stimuli are converted into neural representations that include characteristic electric signals.”[12](p.20). Even at this low-level stage, the perception process is influenced by internal adaptions, e.g., noise suppression if the person is listening to another person. Also so-called explanatory actions may occur due to the presence of stimuli, e.g., reflexively turning the head towards a person who is speaking. The main output of the perception process are ”recognized objects of perception”, which are further processed. For more details about the ”perception process” itself please see [12].

The ”recognized objects of perception” are consciously experienced (”experiencing” in Figure 1.3) which could lead to ”reflection & attribution”, for example if the person is involved in a laboratory QoE user study with the order to evaluate a certain task. Beside artificial laboratory situations, the ”reflection & attribution” process can also be triggered by raised ”quality awareness”, e.g., if the person perceived an inappropriate quality depending on the specific situation and context. In the following ”comparison & judgement” process, the ”perceived quality features” are compared with the ”desired quality features”. For example, in a certain context the person would accept only a single stalling event while consuming a video on YouTube (desired quality feature), but three stallings occurred during playback (perceived quality feature). Hence, this mismatch is translated via the process of ”encoding” to the final output of the quality formation process, the ”quality rating”. In this example the person would state: ”I do not accept this quality”.

Next, the two factors ”user decision” and ”user expectations” are integrated in the quality formation process. In the original quality formation process of [12] user expectations are part of ”assumptions”, but are not described in detail. According to [12], ”assumptions” relate to contextual and/or task-related information, which might be triggered via ”external factors” and the ”perception process”. In Figure 1.3, expectations are included in the quality formation process as ”desired expectations” and ”adequate expectations” [13]. The detailed derivation and discussion of these two kinds of expectations is presented in Section 2.1, i.e., only a short introduction regarding these expectation-types is presented here. In general, adequate expectations are strongly related to a certain context. For example, during a train ride the adequate expectations regarding the stability of a mobile Internet connection are low, i.e., outages and fluctuating bandwidth are expected and tolerated because of the speed of the train, tunnels, etc. In contrast to this, the adequate expectations regarding a mobile Internet connection are notably higher if the connection is used in a coffee shop in the city center, i.e., a high, stable bandwidth is expected and therefore lower quality is not accepted. In contrast to this context-sensitive expectations, desired expectations are more stable and relate to the general attitude of person, e.g., a person could be generally money-saving and would generally prefer lower quality if it was less expensive. Both adequate and desired expectations are involved for determining the ”desired quality features” in the quality formation process, see Figure 1.3. Additionally, ”external factors” can trigger specific adequate expectations via the ”perception process”, e.g., the user is aware of the fact that she is going to consume a movie in high definition quality, which triggers certain quality-related expectations.

After ”comparison & judgement” in the quality formation process in Figure 1.3, not only a ”quality rating”, but also a user decision might be made. For example, the technical quality of a VoD movie is experienced and evaluated as too low and not appropriate for consuming the movie at home via a large TV screen. Hence, the economic decision is made to spend additional money to change the technical quality from standard definition video quality (SD) to high definition video quality (HD). Obviously, this directly impacts the ”signal”. Also, because of the change of the context the related adequate expectations change too, e.g., a high video quality is expected and the ”desired quality features” are adapted. Also non-economic decisions are possible, for example the volume of a music track could be perceived as too low, which leads to the decision to change the volume, which directly impacts the ”signal”. Another example of a non-economic decision could be the choice between several end devices, e.g., the user starts browsing the Internet with a smartphone but decides later to change to a PC.

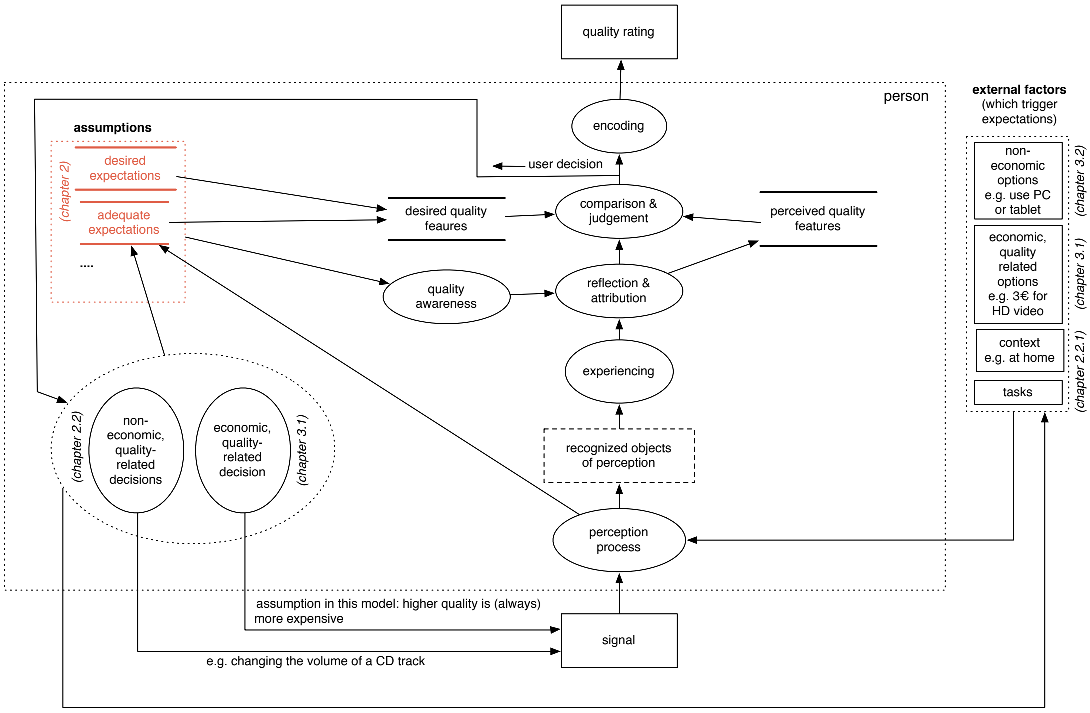


Figure 1.3: QoE Expectations/Decision Model based on the quality formation process from [12]. Blue elements represent novel aspects introduced by the author of this thesis.

1.2 Scientific Contribution

For a systematical investigation of the previously discussed influencing factors *user expectations* and *user decisions*, the following research questions and related research hypotheses are defined. The first research question RQ1 addresses the role of expectations in QoE:

RQ1: How relevant are expectations in QoE?

The first research question focused on the the role of expectations in the context of subjective QoE assessment and QoE modeling. To answer this question a literature overview about related work — both in QoE-related work and in other fields like psychology, consumer satisfaction research, etc. — is presented in Section 2.1. Derived from these findings the concept of *desired* and *adequate* expectations (see Section 2.1.2 for more details about these concepts) is utilized to include expectations in QoE-related research. So far, it is unclear if it is possible to reliably trigger expectations in empirical QoE user studies, which is necessary to investigate their influence on QoE. This assumption is also described in the extended quality formation process (see previous section). Therefore, this assumption is evaluated with the first research hypothesis H1.1:

H1.1: Explicitly communicated information about quality triggers user expectations, which leads to different quality ratings compared to experiments without explicit triggering.

A validation of hypothesis H1.1 would demonstrate that besides technical aspects — downlink bandwidth, video quality, etc. — the triggered expectations impact the Quality of Experience.

To analyze the usefulness of considering user expectations in QoE related research, specially developed questionnaires are discussed and utilized to collect individual information about expectations which are subsequently used to extend quantitative QoE-models for enhanced MOS prediction accuracy. This approach is stated in research hypothesis H1.2:

H1.2: Information about individual quality expectations can be utilized to significantly improve the prediction accuracy of QoE models.

The second influencing factor *user decisions* is the object of the second research question RQ2:

RQ2: What influence do individual user decisions have on QoE assessment?

So far, it is unclear if there is a relevant impact of individual user decisions on subjectively perceived quality. The discussed related work in Section 3.1.1 demonstrates that in general economic decisions impact the assessment of quality features of purchased products. To prove this assumption in the context of QoE, research hypothesis H2.1 is defined as:

H2.1: Individual economic decisions regarding technical quality influence the subjective assessment of quality in a positive way.

In Section 3.2.1 the generally positive impact of having and making a choice is discussed via a literature survey. Hence, research hypothesis H2.2 states a positive impact of non-economic user decisions on QoE.

H2.2: Including individual choices in subjective quality experiments leads to a significant positive difference in quality ratings compared to experiments without any user decisions.

Research question RQ2 deals with the impact of economic/non-economic user decisions on the subjective quality evaluation. Hence, the user decision acts as the independent variable and the resulting quality rating acts as the dependent variable. Nevertheless, in daily life a subjective quality assessment could also impact or trigger subsequent decisions: for example, at home the transmission of a large file from one PC to another PC is started via a WiFi-connection. But because of the connection speed and the file size, the resulting transfer time is perceived as too long and the user switches from WiFi to an Ethernet connection. Hence, the interaction between decisions and quality assessment is rather a permanent interplay than a single, one-way process. Nevertheless, to make this interaction manageable in empirical user studies, the focus in research question RQ2 is on the impact of decision on quality assessment.

Finally, the stated research questions and hypotheses are evaluated in Section 4.1.

1.3 Study Overview & Methodology

To examine the research questions and hypotheses stated in the previous section, not only literature surveys, but also several empirical user studies were conducted. All user studies were laboratory experiments. In each study 26 to 49 users had to fulfill a certain task (e.g. browsing Google Maps on a PC, watching a video via a large TV screen, etc.) under a specific technical Quality of Service (e.g. a low downlink bandwidth, a high video resolution, etc.) with a subsequent quality assessment task in a laboratory setting. Typically, these subjective quality assessments in QoE research consist of a short questionnaire which could contain, for example (see also [30], [31] and [32]):

- The question ”How do you perceive the speed of the Internet connection?” with answering options ranging from 1=bad to 5=excellent via a standard 5-point ACR (absolute category rating) scale [33]. In this case, the user browsed a news site via different downlink bandwidth levels which result in various page loading times.
- The question ”Please evaluate the overall quality of the watched video” with answering options ranging from 1=bad to 5=excellent via a standard 5-point ACR scale. This question is shown after a test participant has watched a video.
- The binary yes/no-question ”Would you accept this quality at home?” This question can be shown after all tasks.

- The question ”How annoying were the interruptions during the video?” with answering options ranging from 5=”not disturbing” to 1=”very disturbing”. Interruptions can occur due to initial loading times before a video is played or by stalling events during the video playback.

Figure 1.4 depicts a schematic representation of an electronic quality assessment questionnaire, which is filled out by the user on a PC after she has fulfilled a certain, quality-related task like browsing a news site with a specific downlink bandwidth. Please note that the user can place the blue ”X” via the mouse cursor anywhere in the grey area between ”excellent” and ”bad”, hence representing a continuous ACR scale. In the depicted example the resulting quality assessment regarding network speed by this user is approx. 3.5 and the user would accept this quality at home.

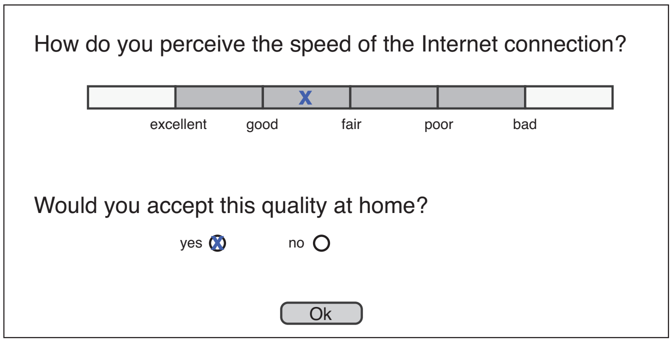


Figure 1.4: Schematic example of an electronic questionnaire to get a subjective quality rating regarding network speed.

Table 1.1 provides an overview about all conducted experiments with information about purpose and subject of study and related publications.

In Table 1.1 studies *Trigger 1, Trigger 2, Trigger 3 und Trigger 4* relate to research hypothesis H1.1 (”Explicit information about quality triggers user expectations, which leads to differences in quality ratings compared to experiments without explicit expectation triggering”). The column ”subject” describes how expectations were triggered, e.g., in study *Trigger 1* the users had to browse the Internet via a mobile Universal Mobile Telecommunication System (UMTS) and a fixed Asymmetric Digital Subscriber Line (ADSL) connection, whereas in study *Trigger 2* the participants used a mobile Long Term Evolution (LTE) and a wireline ADSL Internet connection. In study *Trigger 3* the experiment participants watched videos in standard video quality (SD) and high video quality (HD). In study *Trigger 4* the users had to imagine using certain Video on Demand (VoD) contracts while consuming videos.

The user studies *Measurement 1, Measurement 2, Measurement 3 and Measurement 4* were used to gather individual user expectations via questionnaires and are related to research hypothesis H1.2 (”Information about individual quality expectations can be utilized to significantly improve the prediction accuracy of QoE models”). Information about individual desired and adequate expectations regarding technical quality (for a detailed description of desired and adequate expectations please see Section 2.1) were

purpose of study		name of study	subject of study	chapter	publi-cations
expectations	triggering	Trigger 1	UMTS vs. ADSL	2.2.1.1	[2], [5], [10]
		Trigger 2	LTE vs. ADSL	2.2.1.1	[5], [10]
		Trigger 3	SD vs. HD video quality	2.2.1.2	[5], [10]
		Trigger 4	VoD contracts	2.2.1.3	[5], [10]
	measuring/ modelling	Measuring 1	desired & adequate expectations	2.2.3	[7], [10]
		Measuring 2	desired expectations	2.2.3	[10]
		Measuring 3	adequate expectations	2.2.3	[7], [10]
		Measuring 4	adequate expectations	2.2.3	[10]
Decisions	Eco-nomic	WTP 1	pricing & QoE	3.1.2	[1], [3]
		WTP 2	pricing & QoE	3.1.3	[4], [6]
		WTP 3	pricing & QoE	3.1.4	[8]
	Non-Economic	Decision 1	selectable temporal impairment	3.2	[9]
		Decision 2	selectable end device	3.2	[9]
		Decision 3	selectable web content	3.2	[9]

Table 1.1: Overview of all empirical user studies discussed in this thesis.

gathered and subsequently used to enhance the prediction accuracy of quantitative QoE models.

The user studies *WTP 1*, *WTP 2* and *WTP 3* relate to research hypothesis 2.1 ("Individual economic decisions regarding technical quality influence the subjective assessment of quality in a positive way"). For this purpose, the participants of these studies received real money in advance, which could be used to enhance the quality of an individually selected movie. Afterwards, the users had to evaluate the quality via a QoE questionnaire (please see Section 3.1.2.1 for a detailed description of the experimental setup). To evaluate the research hypothesis 2.1, the findings were compared with user quality ratings which were not affected by individual economic decisions. User study *WTP 1* was conducted in 2011, *WTP 2* in 2012 and *WTP 3* in 2014. Hence, these three user studies also describe the methodological progress over three years which finally culminated in recommendations for conducting QoE experiments including economic user decisions, see Section 3.1.5.

To evaluate research hypothesis 2.2 (Including individual choices in subjective quality experiments leads to a significant positive difference in quality ratings compared to experiments without any user decisions) three user studies *Decision 1, Decision2* and *Decision 3* were conducted. The column "Subject" in Table 1.1 describes the individual user decisions (e.g. in study *Decision 2* the users had to choose between a smartphone, a tablet and a PC for browsing a website).

Additionally, the relevant publications, in which the findings of the related studies were initially presented and discussed are displayed in the column "publications" in Table 1.1 (see also Section 1.4).

In common QoE user studies MOS values are calculated to demonstrate how technical conditions impact subjective quality perception, see [18]. This approach is appropriate for the evaluation of *single* events, e.g., evaluating how browsing a web site with 1 Mbit/s downlink bandwidth (QoS) impacts subjective quality assessment (QoE). In contrast to this, in most of the studies discussed in this thesis *two* ratings from a single person need to be directly compared. For example, one quality rating which is influenced by an economic/non-economic user decision and one unimpaired reference-rating. Therefore, the focus is on the *difference* of these two related ratings. Hence, the so called $DiffRating$ and the concluding $DiffMOS$ are calculated:

For each user i and each quality condition j, the $DiffRating$ can be calculated by subtracting rating A from rating B:

$$DiffRating_{ij} = RatingA_{ij} - RatingB_{ij} \tag{1.1}$$

In the expectation triggering studies (see Section 2.2.1), rating A and rating B relate to two different triggers, e.g., using a wireline Internet connection (A) and wireless Internet connection (B). In contrast to this, for the laboratory studies that investigate the effect of individual user decisions (see Section 3), rating A relates to "pure" quality ratings, i.e., no user decisions were included and rating B relates to quality ratings which include user decisions, for example regarding the used end device.

Based on all related $DiffRating_ij$ values the average $DiffRating_j$ of all users can be calculated which is labeled as $DiffMOS_j$[3]:

$$DiffMOS_j = \frac{\sum^{i} DiffRating_{ij}}{i} \tag{1.2}$$

Depending on the value of $DiffMOS_j$ a direct effect of individual user decisions or triggered expectations on subjective quality assessment ratings can be observed, i.e., if $DiffMOS_j$ is different from zero an effect caused by the experimental setup can be quantified. Nevertheless, a threshold is needed to distinguish between significant differences caused by decisions/expectations and too small differences which are caused by noise, inaccurate rating behavior, etc. To calculate this threshold, the 95% and the 90%-confidence intervals of the $DiffMOS_j$ values are used (specific intervals are stated in each figure).

Apart from bar charts representing MOS values, also CDF (cumulative distribution function) plots depicting the distribution of the $DiffRatings$ are helpful for understanding the rating behavior results. In these CDF-plots, the grey area — defined via

[3] The concept of DiffMOS must not be confused with DMOS (Degradation Mean Opinion Score), which is used for comparison and degradation tests, see [31].

$$(DiffMOS_j - CI) > 0 \quad \Rightarrow \quad \textbf{Positive} \text{ effect of user decisions or expectations for quality condition } j$$

$$(DiffMOS_j + CI) < 0 \quad \Rightarrow \quad \textbf{Negative} \text{ effect of user decisions or expectations for quality condition } j$$

$$|DiffMOS_j| - CI < 0 \quad \Rightarrow \quad \textbf{No effect} \text{ of user decisions or expectations for quality condition } j$$

the confidence intervals — represents the corresponding threshold, i.e., only ratings *outside* this area should be considered as significant. Obviously, depicting this area in a CDF plot is only reasonable if a *single* CDF is plotted. In Figure 1.5 an example of $DiffMOS$ and a CDF based on $DiffRatings$ is provided.

Additionally, to examine the significance of the $DiffMOS$ values — i.e. confidence intervals do not include the value 0 — Wilcoxon rank sum tests are applied [34]. In contrast to other statistical significance tests like ANOVA and t-test which rely on the presumption that the samples follow a normal distribution [35], the Wilcoxon rank sum test is a non-parametric statistical hypothesis test. It tests the null hypotheses that the values from $RatingA_{ij}$ and $RatingB_{ij}$ (see Equation 1.1) are samples from continuous distributions with equal medians. The alternative hypothesis determines that this is not the case. For all Wilcoxon rank sum test results presented in this thesis a 5% significance level is set. In following tables which include information about $DiffMOS$ and Wilcoxon rank sum tests, bold values indicate significant findings: the stated confidence intervals do not include the value 0 and the value of the Wilcoxon rank sum test $p_{ranksum}$ is below 0.05.

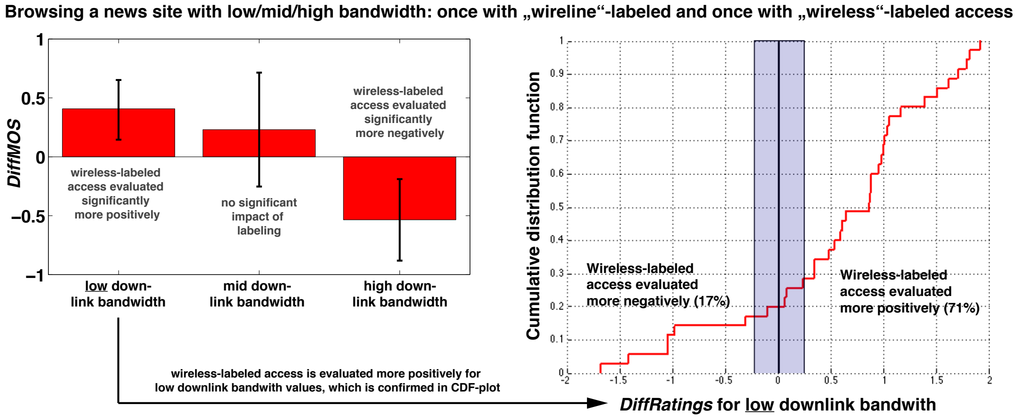


Figure 1.5: Examples of DiffMOS and CDF, which are based on DiffRatings.

1.4 Publications

Parts of this thesis have been previously published via journal papers and conference proceedings, as listed below. All mentioned publications have been evaluated via a peer-review process and conference/workshop submissions were presented personally either orally or via scientific posters by the author of this thesis. The relevant publications are mentioned at the beginning of each section in this thesis.

[1] *Andreas Sackl, Sebastian Egger, Patrick Zwickl, and Peter Reichl. The QoE Alchemy: Turning Quality into Money. Experiences with a Refined Methodology for the Evaluation of Willingness-to-pay for Service Quality. In: Proceedings of the Fourth International Workshop on Quality of the Multimedia Experience (QoMEX) 2012, Yarra Valley, Australia.*
Contribution of the author of this thesis: planning, preparing and analysis of the user study (Study *WTP 1*, see Table 1.1).

[2] *Andreas Sackl, Sebastian Egger, Kathrin Masuch, and Raimund Schatz. Wireless vs. wireline shootout: How user expectations influence quality of experience. In: Proceedings of the Fourth International Workshop on Quality of the Multimedia Experience (QoMEX) 2012, Yarra Valley, Australia.*
Contribution of the author of this thesis: planning, preparing and analysis of the discussed user study (Study *Trigger 1*, see Table 1.1).

[3] *Andreas Sackl, Patrick Zwickl, Sebastian Egger, and Peter Reichl. The Role of Cognitive Dissonance for QoE Evaluation of Multimedia Services. In: Proceedings of the Globecom 2012 Workshop Quality of Experience for Multimedia Communications, Anaheim, California.*
Contribution of the author of this thesis: planning, preparing and analysis of complementary *WTP 1* user study; Theoretical discussion of cognitive dissonance and its implications for QoE).

[4] *Andreas Sackl, Patrick Zwickl, and Peter Reichl. QoE Alchemy 2.0: An Improved Test Setup for the Pecuniary Bias of QoE. In: Proceedings of the Fifth International Workshop on Quality of the Multimedia Experience (QoMEX) 2013, Klagenfurt, Austria.*
Contribution of the author of this thesis: planning, preparing and basic analysis of the discussed user study (Study *WTP 2*, see Table 1.1).

[5] *Andreas Sackl and Raimund Schatz. Evaluating the impact of expectations on end-user quality perception. In: 4th International Workshop on Perceptual Quality of Systems (PQS) 2013, Vienna, Austria.*
Contribution of the author of this thesis: Discussion of expectation-related work; discussion of desired and adequate expectations in the context of QoE; planning, preparation and analysis of user studies (Study *Trigger 2*, *Trigger 3* and *Trigger 4*, see Table 1.1); introduction of a questionnaire to gather information about adequate expectations.

[6] *Andreas Sackl, Patrick Zwickl, and Peter Reichl. The Trouble with Choice: An Empirical Study to Investigate the Influence of Charging Strategies and Content Selection on QoE. In: Advanced Internet Charging and QoS Technology (ICQT) - Associated Workshop of IEEE CNSM 2013, Zurich, Switzerland.*
Contribution of the author of this thesis: planning, preparing and advanced analysis of the discussed user study (Study *WTP 2*, see Table 1.1).

[7] *Andreas Sackl and Raimund Schatz. Got What You Want? Modeling Expectations to Enhance Web QoE Prediction. In: Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX) 2014, Singapore.*
Contribution of the author of this thesis: planning, preparing and advanced analysis of the discussed user study (Study *Measuring 1* and *Measuring 3*, see Table 1.1), discussion of an extended, quantitative QoE/expectations model.

[8] *Andreas Sackl, and Raimund Schatz. "Evaluating the influence of expectations, price and content selection on video quality perception" In: Proceedings of the 6th International Workshop on Quality of Multimedia Experience (QoMEX) 2014, Singapore.*
Contribution of the author of this thesis: Summary of *WTP 1* ,*WTP 2* and *WTP 3* user study (see Table 1.1) and discussion of recommendations for further economic QoE user studies.

[9] *Andreas Sackl and Raimund Schatz. The Influence of User Decisions on Subjective Quality Assessment Ratings. In: Proceedings of the IEEE Workshop on Quality of Experience-based Management for Future Internet Applications and Services (QoE-FI) - in conjunction with IEEE ICC 2015, London, UK.*
Contribution of the author of this thesis: planning, preparing and analysis of the discussed user studies (Study *Decision 1*, *Decision 2* and *Decision 3*, see Table 1.1).

[10] *Andreas Sackl, Raimund Schatz, and Alexander Raake. You Know What You Want? User Expectations and Subjective Quality Perception" In: Quality and User Experience [submitted].*
Contribution of the author of this thesis: Synopsis and discussion of previously published work regarding QoE and expectations.

1.5 Background of the Thesis

The research presented in this thesis was conducted at the FTW Telecommunications Research Center Vienna[4], a Competence Center within the program COMET (Competence Center for Excellent Technologies) supported by the Austrian Government, the city of Vienna and the Austrian Research Promotion Agency (FFG).

[4] http://www.ftw.at, last access: 25^{th} August, 2015.

In the COMET-projects **ACE 2.0** und **ACE 3**[5] the goal was to understand, measure and manage quality in communication networks with a strong focus on the end-users. Hence, several QoE user studies have been conducted with various foci regarding quality evaluation, e.g., high-speed/LTE scenarios, impact of user terminals (smart devices), convergence, QoE for Web and Cloud services, VoiP and video quality (adaptive streaming, IPTV). Methodologically, the ACE projects performed a strictly user-centric cross-layer approach towards QoE by taking into account relevant influencing factors on network, application and user-level. Most of the studies presented in Chapter 2 were conducted within ACE 2.0 or ACE 3.

ETICS (Economics and Technologies for Inter-Carrier Services[6]) was a European research project supported by the European Commission within the 7th Framework Program of the European Union. The goal was to create a new ecosystem of innovative QoS-enabled interconnection models between network service providers allowing for a fair distribution of revenue shares among all the actors of the service delivery value-chain. The three Willingness-to-pay user studies *WTP 1*, *WTP 2* and *WTP 3*, which are discussed in Section 3.1, were conducted within this project.

The strategic project **U-0**[7] of FTW's Area U "User-centered Interaction and Communication Economics" aimed to investigate user experience and interaction in future pervasive communication environments together with their socio-economic context from a holistic and inter-disciplinary perspective. The studies *Decision 1*, *Decision 2* and *Decision 3* regarding user choices and QoE presented in Section 3.2 were conducted within this framework.

1.6 Outline of the Thesis

The remainder of the thesis is structured as follows. At the beginning of Chapter 2 an overview about the current state of the art with respect to expectations is provided, which leads to the utilization of the concept of desired and adequate expectations in the context of QoE. Based on that, the empirical user studies *Trigger 1-4* and *Measuring 1-4* are described and discussed. Finally, it is shown how quantitative QoE models are extended by information about individual expectations to enhance MOS prediction accuracy. The following Chapter 3, which covers user decisions and QoE, is divided into two parts: First, in Section 3.1 an overview about relevant literature regarding economic decisions is given. Additionally, the three user studies *WTP 1-3* are presented and discussed, which finally forms the basis of recommendations for conducting further QoE studies involving economic decisions. In the second part of Chapter 3 the three studies *Decisions 1-3* involving individual user decisions are discussed. Finally, in Chapter 4 of the thesis the research questions and hypotheses are evaluated and an outlook on further research is provided.

[5] http://ace.ftw.at, last access: 4^{th} August, 2015
[6] http://www.ict-etics.eu, last access: 4^{th} August, 2015
[7] http://www.ftw.at/research-innovation/projects/u-0, last access: 4^{th} August, 2015

2 User Expectations and QoE

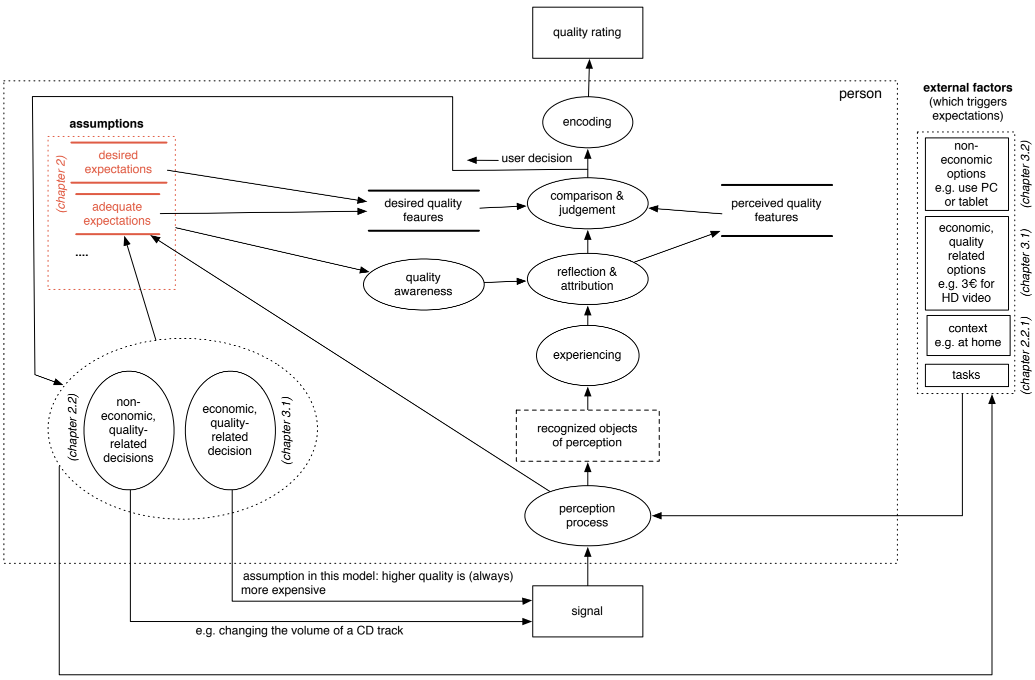


Figure 2.1: QoE expectations/decision model - focus on expectations (red elements)
.

In general, in the context of Quality of Experience (QoE) research [36], the relationship between technical quality parameters like video frame rate, video bit rate, downlink bandwidth, etc. — also referenced as Quality of Service (QoS) — and subjectively perceived quality of a technical system is examined. Various non-technical influencing factors have an effect on this relationship, e.g., user-factors like socio-cultural background or demographic variables and context-factors like individual usage vs. usage within a social context like videoconferencing, see [25] for more information. This chapter of the thesis focuses on one crucial, quality-related aspect which has been neglected so far in empirical, QoE-related research: *expectations*.

Although the term expectations is frequently used in the context of quality perception and QoE related research, there is a lack of applicable concepts and methods enabling the operationalization of expectations and the utilization of related findings in empirical work. Albeit QoE-frameworks and definitions highlight the importance of expectations, clear guidance on how to actually address this influencing factor is missing. For example, in [12](p.19), QoE relates to a "... person's evaluation of the fulfillment of his or her expectations ...", i.e., expectations are described as the perceiving subject's frame of reference. Similarly, in the Qualinet[1] QoE definition white paper expectations are described as key factors determining the end users perceptions and resulting emotional state: "Quality of Experience QoE is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the users personality and current state." [24]. Despite its primarily technical focus, the ITU-T Recommendation P.10 also highlights the relevance of user expectations as an influencing factor. Here, QoE is defined as "the overall acceptability of an application or service, as perceived subjectively by the end user [which] may be influenced by user expectations and context." Also conceptual QoE-models like the one used in [37] explicitly include expectations as a main user influencing factor.

To address this gap in existing research — high relevance of expectations regarding QoE accompanied with low empirical and conceptual research, see also Chapter 2.1.3 for details — this part of the thesis provides a comprehensive analysis of the concept of expectations and makes it methodologically applicable to QoE-research. To this end, the remaining part of this chapter is structured as follows: first, an overview is provided of how other domains — namely consumer satisfaction research and service quality research — define, operationalize and utilize user expectations in various application fields. Based on these findings suitable existing concepts that aid in the operationalization of expectations — namely adequate and desired expectations — are used to extend an existing, conceptual QoE-model. Building on this, an overview about existing QoE literature is given to clarify which aspects of expectations are actually relevant for QoE research. The main part of this chapter focuses on discussing the empirical findings regarding *triggering* and *measuring* expectations in the context of QoE assessment and further describes how to use these findings to *enhance MOS prediction of QoE models.*

Parts of this chapter have been published before in [2], [5], [7], [10].

2.1 Related Work about User Expectations

In this section relevant findings of a literature survey regarding expectations in the context of service quality and consumer satisfaction research are provided. The resulting findings are then used to extend existing QoE-models to finally propose a conceptual QoE/Expectations model, which has been introduced before in Section 1.1. Based on that, existing QoE-related literature is discussed to identify open research challenges in the context of QoE and expectations.

[1] COST Action IC1003, European Cooperation in Science and Technology, http://www.qualinet.eu/, last access: 29th July, 2015

2.1.1 Expectations in Socio-Psychology, Service Quality and Consumer Satisfaction Research

A good starting point for learning about the nature of expectations is the field of psychology in which the concept of expectations is often used to describe the processes of understanding and cognition because "in perception, we consider prior expectations" [38](p.108). For example, when someone reads a piece of written text, she has to make assumptions regarding the content, the context of the work's creation, the author's purpose, etc. Consequently, the "understanding at each point in the [text] was influenced by [...] existing knowledge and expectations based on [...] own experiences within a particular context"[38](p.393).

Additionally, expectations play a critical role in psychological research in the context of decision making, see [39](p.243), and also in motivation and behavioral changes, see [39](p.300). In the field of socio-psychology, expectations are considered as an important aspect which determines how subjects interact with others to fulfill social norms, cf. [39](p.447, 453). Additionally, the well-know concept "self-fulfilling prophecy" describes how triggered expectations can lead to unforeseen output, see [39](p.450).

To obtain a more practical and applicable definition of expectations, research fields like human-computer-interaction, economics, etc. yield more informative results. The authors of [40] define expectations as "a kind of schemata that focuses interpretation processes on specific meanings and functions of communicative action" [40]. As described in [41], expectations are "pre-trial beliefs about a product or service and its performance at some future time" and expectations "form the frame of reference for satisfaction judgments" [41]. Additionally, the authors of [41] divide expectations into four main categories:

1. Forecast (or expected/predictive): user beliefs about what *will* occur in specific forthcoming action regarding a specific provider.
2. Normative (or deserved/desired): consumer perception of what *should* occur based on an assessment of what is realistic and feasible regarding a specific provider.
3. Ideal (or wished): highest level of performance attainable, independent from specific provider or brand.
4. Minimum tolerable (or adequate): minimum baseline performance acceptable, independent from specific provider or brand.

According to the authors of [41], practical implications of expectations have been investigated mostly by two different research traditions:

1. Consumer satisfaction research: here, the primary goal is to understand the user's cognitive processes, which leads to customer's satisfaction.
2. Service quality research: here, the primary object is to understand and to measure quality in service environments[2].

[2] For more information about these two traditions, please see [42], [43] and [44].

In this context, one of the most widely used models regarding perceived service quality and expectations is the GAPS model and, building on that the SERVQUAL model developed by Parasuraman et al. [45]. For example, in the context of information system evaluation, the authors of [46] used the SERVQUAL model to measure existing expectations to compare it with gained experiences: Test users had to indicate their ideal information system and to evaluate 22 questions via a 7-point answering scale with answering options ranging from 1="Strongly disagree" to 7="Strongly agree", e.g., "The employees of these Information Systems units will understand the specific needs of their users". After that, the test participants used a particular information system implementation to perform some tasks. Then, the users had to evaluate their experiences by answering slightly rephrased questions, e.g., "Employees of this Information System understand the specific needs of its users." With this information, it is possible to calculate the expectation gap.

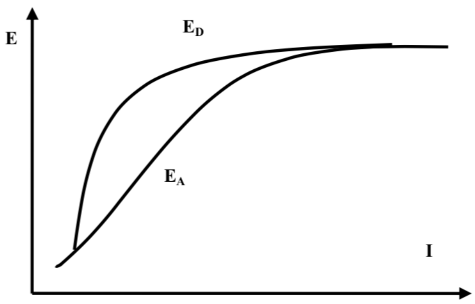
Similar to SERVQUAL but in the context of e-commerce, the authors of [47] used the Expectation Confirmation Theory (ECT), originally developed by the authors of [48], to measure the user-satisfaction of Web services. The satisfaction level was generated by a comparison of post-purchase evaluations of a product or service with pre-purchase expectations. Following the ECT, users generate specific expectations regarding a desired product. After a trial phase, the users form perceptions about its performance. Then the participants determine if their expectations were confirmed regarding the perceived performance. Finally, the users' satisfaction level results from the previous confirmation and the underlying expectations. In the end, a reuse or repurchase is considered or not.

In general, existing literature points out an important difference regarding asking about expectations: in the service quality tradition the subject states expectations about what the service provider *should* offer. In contrast, in the consumer satisfaction literature the subject reveals more about his/her expectations what *will* be offered.

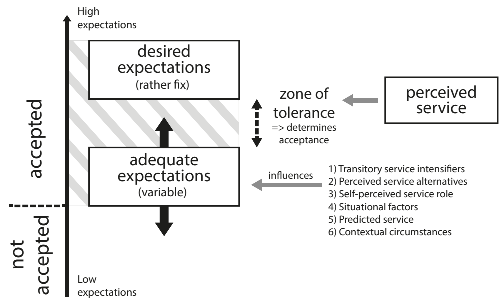
When it comes to QoE assessment, inquiring expectations can be problematic because for the user it is often not easy to *verbalize* expectations. For example, what does a 'fast' Internet connection actually refer to? How should end users quantify the expected technical quality of a video transmission? In this context, the author of [49] focuses on an understudied aspect of service quality and user satisfaction, namely fuzzy, implicit and unrealistic expectations. According to the author, users have fuzzy expectations "when they expect a change but do not have a precise picture of what this change should be.[49]" If these expectations are not fulfilled, for the concerned users the experienced service was unsatisfactory but they do not know why. The opposite of fuzzy expectations are precise expectations. According to the authors of [49], implicit expectations are so self-evident users do not actively think about them. They only become relevant and explicit for the users when these expectations are not fulfilled. Finally, there are also unrealistic expectations which are obviously unable to fulfill. The authors of [49] argue, that fuzzy expectations can be converted in more precise expectations via a dialog between the user and the service provider. Obviously, this qualitative approach is not appropriate in the context of quantifying expectations for QoE related research.

In the field of consumer satisfaction research, *brands* play an important role for expectations. Here, examined expectations — and their fulfillment or non-fulfillment — are based on concrete products or services, cf. [50]. In contrast, assessing the influence of brands or other marketing-related aspects is not very common in QoE research. Here, the focus is mainly on assessing the impact of the quality/performance of the technical system which is typically evaluated by experiment participants without having any background information, e.g., in a video study, the responsible video content provider delivering the streamed video is irrelevant[3].

According to the authors of [50], in the context of service quality research it is not always straightforward to quantify expectations. For example, it is fairly simple to quantify the speed of the service in a restaurant by means of asking about expected seconds or minutes of waiting time. Similarly, menu variety can be determined by the number of food items on the menu. However, it is considerablly more complicated to measure and operationalize employee-friendliness. Obviously, this particular issue can be solved via appropriate questionnaire designs. Nevertheless, quantifying individual expectations is difficult and depends on the context and the evaluated service.



(a) Mathematical relationship between desired and adequate expectations, originally published by [51].



(b) Desired/adequate-expectations model. Own graphical representation of the desired and adequate expectation concept of [13].

Figure 2.2: Mathematical relationship and model of desired/adequate-expectations

In the context of usability research, expectations can be used to identify critical usability problems. Briefly speaking, users of a usability experiment are asked beforehand how difficult a certain task would be, e.g., find item X on website Y. Subsequently, the users have to evaluate the difficulty of this activity after they have fulfilled this particular task. The ensuing comparison shows which issues should be fixed immediately and which are not crucial. For example, if task A was expected by most of the users to be very easy and the ratings after the usage were mostly "very difficult", this issue should be fixed

[3] Most QoE researchers would agree that a broad range of influencing factors impact the perceived quality, but when it comes to empirical research or modeling attempts, softer influencing factors like brands or other marketing factors are mostly neglected.

very quickly, see [52] for more details. Obviously, this approach is not appropriate for QoE because for the users it is often difficult to verbalize quality related expectations, which has been discussed beforehand.

According to Zeithaml [13], there are two different types of expectations: desired and adequate expectations. The authors of [13] stated that the *desired expectations* are rather stable and invariant, e.g., some users are always concerned about high quality or some users are always concerned about low prices, and these basic needs do not change over time. In contrast to this, *adequate expectations* are more flexible and they are strongly influenced by the context. Between these two kinds of expectations, there is the so-called *zone of tolerance*: if the perceived service is in between the invariable desired expectations and the variable adequate expectations, the user/customer accepts the perceived service, see Figure 2.2(b) for a graphical representation. In the context of typical business/customer relations, adequate expectations are influenced by (see also Figure 2.2(b)):

1. Transitory service intensifiers: The urgency of a situation can lower the adequate expectations.
2. Perceived service alternatives: If there are alternatives available in the current situation or if it is possible to solve an issue without external support, the adequate expectations get higher, which leads to a smaller zone of tolerance.
3. Self-perceived service role: The customer tries to fulfill her role in the current process, e.g., it is not always possible to blame others for non-fulfilled expectations. Therefore, the more pretentious the level of a customer's view on the self-perceived service role is, the higher is the level of adequate service.
4. Situational factors: These factors can lower the level of adequate service if the environmental influences are independent from a service provider. Hence, customers realize that this is not the fault of the provider and they accept a lower service level.
5. Predicted service: This is the service quality customers believe they are likely to get.
6. Contextual circumstances: For example, economic aspects are included here. A participant in the study presented in [13] stated that "price increases do not really drive up expectations. But my tolerance level will become more stringent/less flexible with an increase."

There are some attempts to quantify adequate and desired user expectations in current literature. For example, the authors of [51] use the Weber/Fechner-law — which originates from psycho-physics and is also used for QoE modeling — to generate a quantitative expectation measurement model which mathematically describes the relationship between the desired and the adequate expectations of customers regarding service providers. Figure 2.2(a) shows that desired expectations E_D are rather stable even if the stimulus magnitude I of the expectation determinants increases (e.g. personal needs, transitory service intensifiers, perceived service alternatives, customer self-perceived service role, situational factors ... etc., see also [13]). In contrast to this, adequate expectations E_A are more flexible and also increase when the stimulus magnitude I of the expectation determinants increases. Nevertheless, the authors of [51] do not describe *how* to measure or quantify expectation determinants and therefore their contribution to the challenge of quantifying expectations is rather limited.

In common expectation and service quality research approaches, it is usual to capture expectations in the post-consumption phase instead of getting expectation information before a certain action is initiated, cf. [44]. Nevertheless, in the context of QoE it is more relevant to get information about users without conducting a certain experiment, so that this information can be used, for example, for MOS prediction modeling.

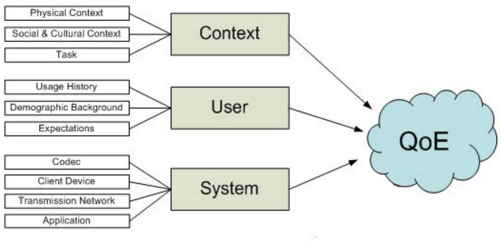
It is also a common assumption that test subjects have had *prior experiences* in a way that they can articulate expectations for current evaluation tasks, see [11]. Nevertheless, some researchers assume that expectations exist even when no prior experience has been gained (see for examples [53] and [54]), whereas the authors of [55] state that expectations can not be generated without prior usage.

To summarize, the nature of expectations has been examined in various research areas, e.g., human-computer-interaction, economics, psychology, etc., with the findings listed below being essential for understanding the role of expectations in the field of QoE:

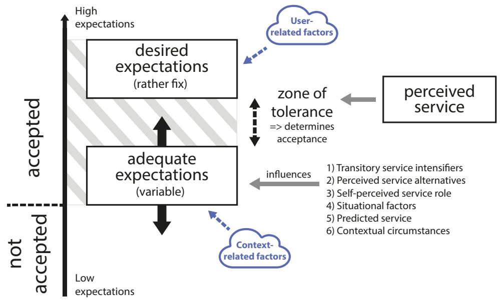
- Expectations depend on a broad variety of influences and understanding how they emerge and how they influence quality perception is not trivial.
- Expectations can have negative effects on perceived quality when being *under-fulfilled*, but also positive ones when being *over-fulfilled*. Therefore, empirical QoE research should also include test conditions in which expectations are over-fulfilled (currently, the focus of QoE research is on situations in which expectations are not fulfilled).
- In research fields dealing with service quality expectations are generally considered as measurable, e.g., by means of questionnaires. But this approach requires expectations which can be *verbalized* and which are *quantifiable*, which is not always possible in the field of QoE. For example, it might be challenging for an average end user to verbalize her expectations regarding Internet connection speed in Mbit/s.
- It is essential to distinguish between relatively stable, higher *desired expectations* and variable, *adequate* expectations, which both together influence the acceptance of a certain service.

2.1.2 Conceptual QoE/Expectations Model

In the previous section, the current state of the art regarding expectation classification and measurement in the field of service quality and customer satisfaction research has been discussed. For the next steps it is necessary to include expectations in common



(a) The term expectations as an important user-related influencing factor on QoE, published in [37]



(b) Adapted desired/adequate-expectations model. Own graphical representation of the desired/adequate expectation concept of [13] extend with user- and context-related factors.

Figure 2.3: Generic QoE framework with influence factors and QoE/Expectations model based on Zeithaml

conceptual QoE models in an appropriate way, which finally leads to the model already presented at the beginning of the thesis in Figure 1.3 in Section 1.1.

Zeithaml's concept of *desired* and *adequate* expectations [13] (see also previous section) is used to integrate expectations into QoE research. As depicted in Figure 2.3(a) current QoE models (e.g. see [37]) often consider the specific influences of *context*-related and *user*-intrinsic variables on QoE. In terms of Zeithaml's expectations concept, the phrase "user" can be seen as related to the rather stable "desired expectations", whereas the phrase "context" can be seen as related to the context-sensitive term "adequate expectations". Hence, a combination of the two influencing factors with Zeithaml's model can be conceived, see Figure 2.3(b).

As already mentioned in Section 1.1 the quality-formation process — based on Jekosch [17] and Raake [56] — is extended by desired and adequate expectations. The detailed description of the *quality formation process* can be found in [12](p.23), also an introduction was given in Section 1.1. Hence, the following explication focuses on how expectations are included in the process.

In the model depicted in Figure 2.1, the "signal" represents any audiovisual information which is perceived by the user, e.g., a consumed video, a Web site interaction, etc. The fine-grained "process of perception" — which is described in detail in [12](p.20) — results in "recognized objects of perception" which furthermore influences the "experience"-character of the situation, e.g., a disturbance is recognized by the user, for example a too long page loading time during Web browsing. The so called "quality awareness" is triggered[4], which results in a ”reflection & attribution” process, i.e., the user is aware of a quality problem which results in a ”comparison & judgement” phase. During ”comparison & judgement”, the ”perceived quality features” and the ”desired quality features” are compared which results via ”encoding” in a ”quality Rating”, e.g., the user could state ”I would not acceptance this video quality at home” or a rating could be made on a 5 point evaluation scale, etc. The ”desired quality features” highly depend on the desired and adequate expectations. It is assumed, that desired expectations are rather stable and do not change over time[5]. Also ”external factors” influence the ”perception process” like the used end devices (e.g. is a video consumed via Laptop or via Smartphone?), situational context (e.g. is a Web site accessed at home or via a mobile connection during a train ride?), is there a specific task to fulfill (e.g. booking a flight via a Web site) or not (e.g. relaxing while listing to music), etc. Hence, via the ”perception process” these external factors shape the adequate expectations/assumptions — e.g. a user might be more tolerant regarding network outages during a train ride in contrast to surfing at home via a fixed-line ADSL connection — which finally determine the ”desired quality features” for ”comparison & judgement”. One output of the ”comparison & judgement” could be that the user is not satisfied with the current (quality) situation and something has to change. Hence, a ”decision” is made, e.g., changing from WiFi to Ethernet connection if Web page loading times are too long (non-economic decision) or changing from SD to HD quality during a Video on Demand Session which implies additional fees (economic decision). More information about economic/non-economic user decisions and QoE can be found in Chapter 3. Of course, these user decisions influence the adequate expectations, e.g., a switch from HD to SD video streaming lowers the adequate expectations regarding video quality, but the adequate expectations regarding interruptions are higher, i.e., the user is less tolerant regarding stallings[6]. Additionally, some decisions also influence the ”signal”: if a user raises the bitrate of a music streaming service, the perceived ”signal” also changes.

2.1.3 Practical Inclusions of Expectations in QoE Related Research

In this section, a review is presented about the current state of the art of QoE-research in the light of the model outlined in the previous section. Based on the literature survey, open challenges are identified which will be empirically addressed in subsequent sections.

[4] Additionally to disturbance situations, quality awareness can also be triggered *externally* in a laboratory setup, i.e., a researcher asks an experiment participant to evaluate the perceived quality in a particular setup, and *internally*, e.g., if the user wants to buy a new TV set and compares the picture quality of several screens.

[5] Obviously, even desired expectations adopt over time but it is assumed that desired expectations evolve over long-time periods like years or decades. Hence, in this model it is supposed that they remain constant.

[6] Assuming that the network connection remains constant, less data is transferred during SD streaming which should result in smoother playback without interruptions.

2.1.3.1 The Relevance of Expectations in QoE-Related Research

Many QoE researchers are aware of the existence and relevance of expectations, see also QoE-definitions at the beginning of this chapter. However, they do not explicitly assess them. In principle, when optimal quality is being achieved in a given context, we can assume that at least *adequate* expectations are being met. In turn, when quality is not perceived as optimal, expectations are obviously not met. Hence, quality assessment can be considered as a way to indirectly assess expectations. In the conceptual model discussed in the previous section, quality assessment results from a comparison of *desired* and *perceived* quality features. In principle, perceived features can be obtained from a multidimensional analysis of quality, following approaches as used, for example, by [57] or [58]. Here, a given percept is considered to be related with a multidimensional set of perceptual (quality) features, and can be represented in a multidimensional feature space. In case of optimal quality, it can be assumed that the desired features are met by the perceived features; hence, at such an operation point, the perceived features can be seen as a measure of the desired features in that given context. Obviously, with such indirect assessment, adequate and desired expectations cannot readily be distinguished.

The authors of [59] discuss some common QoE models and methods like peak signal-to-noise ratio (PSNR), mean squared error (MSE), application Quality of System (AQoS) etc., but in the end they state that ”these models still lack the understanding of Human Visual System (HVS), and even more so, none of the models takes into account the expectations of the viewers” [59].

In the QoE community there is increasing awareness that expectations play a key role and that expectations change and adapt over time. Nevertheless, no quantifiable results are available. For example, the authors of [60] state that network providers are facing major changes in user expectations, e.g., higher awareness of the provided network quality. They also state that ”an increase of usage of online services can be caused either by heavier use by existing users or an increase of the number of users. Anyway, both lead to higher expectations for performance and reliability of the services, thus increasing the demand for QoE mechanisms within the network.” [60]. Additionally, the authors of [61] state: ”Due to the growing number of new handsets and smartphones which increases the user QoE expectations, it is important for the operators to know and to measure the UEs [User equipment] performance”. Also the authors of [62] point out, that ”It can thus help in providing personalized services such as selecting a proper codec or by selecting a network interface which provides QoE based on user’s expectations.”

In the context of speech quality in telecommunication systems, the author of [32] also states that the term expectation is a rather diffused one and it is not used in a unified way in telecommunications. Three components influence expectations: the user´s general experience with a service, the price (e.g. more expensive often correlates with higher quality) and the nature of the connection, e.g., private call vs. informative call. In the work of [32], expectations are discussed in the context of diffusion theory [63], i.e., how expectations change during the introduction and establishment of new technologies. For example, it is stated that after a new speech transmission technology has been introduced, the users start to use the new technology for other, different purposes, which leads to an decreased demand for transmission quality. While the users are getting used to the new technology, the demand for transmission quality is increased. Also, customers can be separated into user groups, e.g., innovator or early adapter, which additionally influences the expectations regarding transmission quality during the phases of establishment.

2.1.3.2 The Controllability of Expectations in QoE-Related Experiments

Although many researchers are aware of the influence of user expectations on their conducted experiments, the controllability of triggered expectations is rather neglected: For example, the authors of [64] investigated the difference between QoE-experiments carried out in standardized environments compared to experiments which were carried out in more realistic living room environments. The different contexts and the different expectations related to these contexts were described as the main influencing factors regarding the different quality assessment outcomes. Nevertheless, the authors did not measure, describe or quantify end-user expectations. Additionally, in [65] the authors evaluated distorted videos with different video resolutions. There, it was found that a different set of expectations seems to apply when comparing HD against SD viewing cases. Nevertheless, neither the direct influence of these expectations nor the controlled triggering of them were part of this study. In [66] the authors state that poor quality rating results might be explained by the fact that a certain scenario came after another scenario, so that the user expectations might have been high and they expected the same quality.

Additionally, some researchers state that expectations are explicitly excluded in their experiments, e.g., ”We remark that we do not consider other situational factors such as the users’ expectation (e.g., free vs. paid call)” [67] and ”However, from a cloud service provider’s perspective, it is challenging to gain insight into the users’ expectations and experiences” [68]. Hence, establishing controllability over evoked or triggered expectations in empirical QoE-research is highly relevant. The authors of [69] developed a QoE-driven bandwidth allocation method based on user characteristics. Nevertheless, the user classification approach neglects some aspects like application, situation and also expectations regarding psychological effects. Hence, also this work demonstrates that the authors are aware of the existence of expectations and their possible impact on user experience, but specific methods to control or trigger expectations are still missing[7].

So far, there are only a few expectation related experiments in the context of telecommunication services. In [32], the author describes a user study — similar to the one discussed in [71] — in which pairs of test participants had conversations via a portable

[7] The challenge of triggering and controlling expectations in the context of QoE is different compared to other research fields, i.e., in some research fields this kind of problem is non-existent or is perceived as non-existent: in a field information system evaluation study the participants of [70] only read descriptions of three information systems and — according to the authors — expectations for the follow-up experiments were set.

headset and a wireline one degraded by time-invariant impairments. For all evaluated test conditions there are no significant differences between the connection types. A separation between user groups based on the user's previous experience with mobile phones lead to small differences between the ratings, but also here the differences are not significant. Also in [32], a study by [72] is discussed which investigated the influence of expectations regarding making a telephone call from an Internet terminal compared to making a call from a standard wireline terminal. Similar to the previously presented study, only not-statistically significant tendencies regarding the impact of different expectations of connection types were found.

2.1.3.3 Assessment of Expectations

In the context of service quality and customer satisfaction research, the authors of [73] stated that customers can *articulate* how well a product or service meets their expectations. This might be true for some kinds of products and services, e.g., cars or restaurants, but in the context of QoE it is at least doubtful that users are always able to *articulate* expectations. For example, in the context of Web QoE users might agree that they expect a fast Internet connection. But it might be hard for the users to articulate and to define what fast exactly means for them. In [41], the authors examine the expectations and satisfaction of an art museum with an adapted SERVQUAL-questionnaire. Obviously, it is possible to verbalize specific expectations in the context of galleries and museum experiences, e.g., range of appropriate souvenirs, employees' willingness to help, minimize waiting lines/ticketing queues, etc. In [74], the authors present a study about the evaluation of a spoken dialogue system via SERVQUAL questionnaires. Five service quality dimensions were evaluated by the test participants: tangibles, reliability, responsiveness, assurance and empathy. Overall, the participants had to answer 22 items which cover these quality dimensions. Before the dialogue system was used by the participants the expectation-related part of the SERVQUAL-questionnaire was filled out, i.e., the users had to state an "accepted level" and a "desired level" regarding all items, for example "Service is fast". After the usage of the system, a questionnaire about the perception of the system was filled out which covered the same quality dimensions. Although the discussed experiment had led to interesting insights, the authors of [74] critically note that the large amount of 66 questions had a negative impact on the motivation of the participants to provide correct answers. Unfortunately, no common QoE related questionnaires were used in the experiment, i.e., it is not possible to combine MOS values directly with SERVQUAL data.

Nevertheless, in the QoE research community there is a general awareness that it is necessary to get information about user expectations. For example, in [75] the authors assume that the different quality expectations of the test participants should be considered because users with a high-speed Internet connection at home might be less tolerant regarding long page loading times compared to users with a slow Internet connection. Hence, it is necessary to get information about existing user expectations to explain gained quality assessment ratings.

Nonetheless, there are also attempts to assess quality expectations in related work. The authors of [76] assume that user behavior can be utilized to derive information about user expectations, i.e., according to the authors actively pausing a video stream reflects certain user expectations and changing the video resolution while consuming a video stream also indicates user expectations. Hence, in the quoted paper expectations are derived from user behavior. It seems that sometimes QoE and expectations are more or less interchangeable, e.g., in [77] the authors state that user expectations can be analyzed by measuring MOS, PSNR and structural similarity (SSIM). In this case, expectations are a result of experience and not vice versa.

Expectations and the desired features they relate to may undergo adaptation. Hence, in the context of quality assessments, such adaptation may be reflected via certain biases. For example, when a set of stimuli is presented that has a specific quality range, the usage of the quality scale will be different for an individual stimulus than when that stimulus is presented with a different set of stimuli. Such effects like the range equalization bias [78] may partially be related to the specific focus on individual degradations or the mapping of features to an overall quality judgment. In the present work, the interaction of the test paradigm with the topic of research, namely to assess expectations and their role in quality evaluation, cannot completely be avoided. In [12], the term "Schrödinger's cat problem of QoE research" was coined to describe this effect. Since the goal of this work is not to exactly quantify expectations but rather to a) trigger them and b) assess them in different contexts, the remaining influence of the test situation on expectations is considered to be acceptable for the research presented in this thesis.

In [79] the authors stated: "The goal of these studies will be to determine user expectations (e.g. by interviewing users before the gaming session is started), to attempt to quantify user satisfaction (e.g. through observation during play) and to collect user feedback (for instance through a questionnaire or post-gaming interviews)." Hence, it might also be a promising approach to gather information about user expectations via qualitative approaches. However, to evaluate the usefulness of operationalized expectations via quantitative models, a qualitative approach like interviewing is inappropriate.

The authors of [80] state that participants in experiments who are used to consuming video content in low resolution will rate differently than those who regularly consume video content in high resolution. Hence, their expectations are different, which makes it necessary to get information about existing user expectations. Additionally, the authors of [80] also presume that expectations are closely related to the country of the users, i.e., users from different regions may have different expectations about the provided content quality. Also here, no explicit empirical evidence is presented regarding this statement.

2.1.3.4 Extending QoE-models with Expectations

So far, there are only few attempts to include expectations in MOS-predictive QoE-models[8]. For example, the E-model defined in [81] and applied e.g. in [82] includes among others user expectations as an additional factor to calculate the resulting transmission quality rating of an audio transmission. Equation 2.1 describes how the quality rating R is defined by five terms:

$$R = R_0 - Is - Id - Ie + A \tag{2.1}$$

According to [32], R_0 describes "... the transmission rating for the basic signalto noise ratio at the virtual 0 dBr point of the connection". The terms Is, Id and Ie describe impairment factors, e.g., quantizing distortions, echoes, device impairment factors like low-bitrate codecs etc. [32]. In the original E-model[9] the expectations factor A stands for 'advantage of access'. Hence, lower technical quality, for example caused by a mobile connection, is compensated by the fact that the user takes full advantage of being able to make a call from various locations. Therefore, the expectation factor A acts as a compensation factor for technical impairments. Nevertheless, the expectation factor A does not cover all expectation-related aspects because "... it does not take into account the special situation of the user which it pretends to model." [32](p. 101). Hence, further effort is required to cover more expectation-related aspects in quantitative QoE models.

To summarize Section 2.1.3, it can be stated that in QoE research literature:

1. There is a strong consensus that expectations play a major role in quality perception and assessment.
2. Expectations may influence the outcome of empirical user studies but control over these influencing factors tends to be limited.
3. Direct measuring of expectations is rather difficult and thus only assumptions or inferences have been made so far, e.g., via MOS [77] or via qualitative interviews [79].

Hence, according to these findings the following research challenges should be met in order to investigate the role of expectations in QoE:

1. Is it possible to trigger expectations in empirical user QoE studies?
2. Is it possible to assess expectations in the context of QoE in a quantitative way?

[8] In the context of service quality and consumer satisfaction research it is more common to use quantified data about user expectation to predict user behavior and/or user satisfaction. For more details please see for example [**?**].

[9] An extended E-Model which has been adapted for Voice over IP services is introduced and discussed in [56].

3. Is it possible to extend quantitative QoE model with expectation information to enhance MOS prediction accuracy?

Obviously, user expectations regarding quality relate to many aspects, like technical aspects (e.g. bandwidth of an Internet connection, video resolution), economic aspects (e.g. how expensive a LTE Internet connection should be), content-related aspects (e.g. amount of available movies in a VoD database), etc. Due to its inherent complexity it is necessary to focus on specific aspects of quality related expectations. Hence, the experiments in the following section focus on the triggering of *technical-related expectations*. The assessment of adequate, quality-related expectations in Section 2.2.2.2 also focuses on technical quality-aspects, e.g., how fast an Internet connection has to be. In contrast to this, in Section 2.2.2.1 several quality-related expectations are covered, including individual expectations regarding contract duration, support modalities, etc. Most of the quantitative QoE-models in Section 2.2.3 include technical-related expectation information to enhance MOS prediction. Nevertheless, in general expectations regarding quality contain usually more than "just" technical aspects, but the focus of this thesis is on the technical ones.

2.2 Expectations in Quality Assessment & Modeling

In this section the conducted empirical user studies and their results regarding triggering, assessment and modeling of individual user expectations are discussed. To this end, first the results of four QoE studies with active triggering — i.e. explicitly triggering of expectations of experiement participants — are presented. Then, a novel method to get individual information about desired and adequate expectations via questionnaires is discussed. Finally, it is explained how to utilize this information to enhance the accuracy of predictive QoE/MOS-models.

2.2.1 Triggering of Expectations in Laboratory Setups

In Section 2.1.3 it is stated that QoE researchers are aware of the influence of expectations on subjects' responses in empirical user studies. Nevertheless, so far it has not been proven that explicit triggering of individual expectations is possible and reasonable. Hence, in this section the results of four expectations triggering experiments are discussed, see Table 2.1 for an overview.

ID	year	expectation trigger	QoE rating	study design	n
Trigger 1	2011	ADSL vs. 3G Internet access	Internet speed	within subjects	49
Trigger 2	2013	ADSL vs. LTE Internet access	Internet speed	within subjects	45
Trigger 3	2012	SD/HD video quality hint	video quality	within Subjects	26
Trigger 4	2011	VoD contract type	video quality	between subjects	44

Table 2.1: Overview of expectation trigger QoE studies (n=number of participants)

In studies *Trigger 1-3* a within-subject design was applied. In this section $DiffRating$ refers to a rating-difference by a certain user and a certain condition, whereas the average difference $DiffMOS$ includes all users of a certain condition (the concept of $DiffRating$ and $DiffMOS$ is described in detail in Section 1.3). In study *Trigger 4* a between-subject design was applied, so here the $DiffRating$ and $DiffMOS$ approach was not possible.

2.2.1.1 Wireline vs. Wireless Internet Access - Study Trigger 1 & 2

In 2011 (study *Trigger 1*) and 2013 (study *Trigger 2*) two empirical user studies were conducted with the goal to assess how the type of Internet access (wireline via ADSL vs. wireless via UMTS and LTE), as *assumed* by participants, and thus the different expectations triggered, influences their QoE ratings. The test participants had to browse several Web sites using a PC, which was connected via an ADSL (2011 & 2013), via a 3G UMTS (2011) and via a LTE (2013) Internet connection. Most importantly, the users had to manually switch the connection type several times during the test procedure by themselves via a physical device called *ConnectionSwitcher*, see Figure 2.4(a) and 2.4(b). In fact, from a technical point of view both connection types *wireline* (ADSL) and *wireless* (3G/LTE) were *identical*, i.e., during the whole test the participants used a dedicated line to connect to the Internet. The ConnectionSwitcher was only a non-functional, but realistic mock-up, e.g., LEDs indicated the current connection mode and the 3G/LTE modem had built-in LEDs to indicate connection build-up and data transfer phases. Hence, the *labeling effect* regarding Internet connection types was evaluated.

Several conditions with different downlink bandwidth levels were tested, i.e., the user experienced different page load times according to the set QoS levels. The management of the applied downlink and/or uplink bandwidth was done via the well-known network emulation tool *netem*[10], which was implemented in the laboratory environment in FTW's i:Lab. After each condition, which lasted for 2 minutes and included browsing a Web site via the wireless- or wireline-labeled connection, the test users had to evaluate the subjective experience regarding network speed via a standard 5-point ACR scale, ranging from "excellent" to "bad", see the technical recommendation [83].

In study *Trigger 1* (2011) the users browsed a custom made news site, photos on Facebook, Google Maps and they consumed animation videos. For browsing the news site the downlink bandwidth levels 512 Mbit/s (low) and 2048 Mbit/s (high) were applied. For browsing Google Maps, the downlink bandwidth levels 512 Mbit/s, 2048 Mbit/s and 8192 Mbit/s were applied. The video named "low"(note: experiment participants did not know the ID of the video) had a resolution of 640 x 360 pixels, a frame-rate of 24 frames-per-second (FPS), a video bit rate (VBR) of 2.69 kBit/s and was encoded with h.264/profile High L4.0. The video named "mid" had a resolution of 1280 x 720 pixels, 24 fps, a VBR of 8.08 kBit/s and was encoded with h.264/profile High L3.1. Finally the video named "high" had a resolution of 1920 x 1080 pixels, 24 fps, a VBR

[10] http://www.linuxfoundation.org/collaborate/workgroups/networking/netem, last access: 16th September, 2015

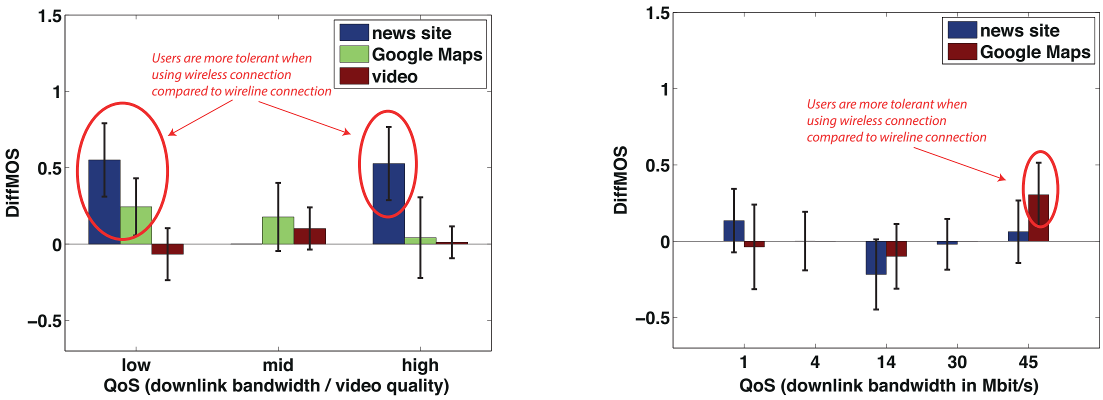
of 22.2 Mbit/s and was encoded with h.64/High L4.0. In study *Trigger 2* (2013), the users browsed a news site (http://www.cnn.com) and Google Maps. The corresponding downlink bandwidth levels were 1, 4 14, 30 and 45 Mbit/s.

(a) Scheme of ConnectionSwitcher

(b) Usage of Connection-Switcher

(c) Desktop Icon which indicates current connection mode

Figure 2.4: Study *Trigger 1* and *Trigger 2*: Details about the ConnectionSwitcher

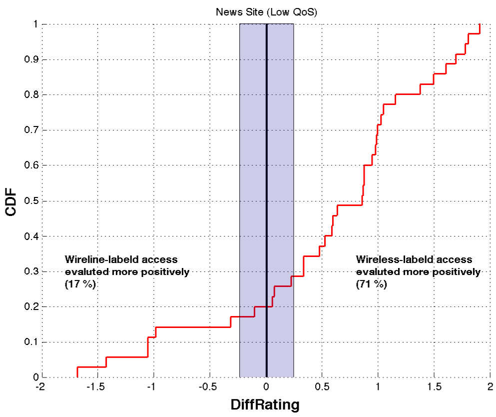


(a) Study Trigger 1: $DiffMOS$ results for 2011 Internet connection labeling experiment (ADSL vs. 3G, 90% confidence interval)

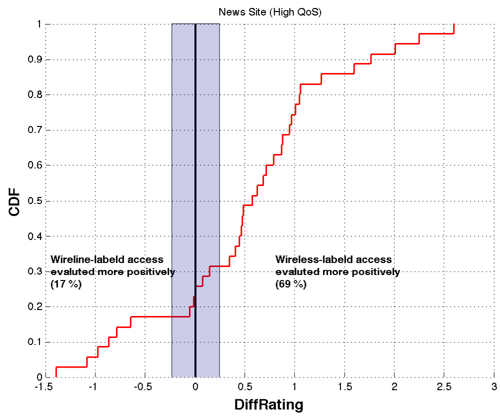
(b) Study Trigger 2: $DiffMOS$ results for 2013 Internet connection labeling labeling experiment (ADSL vs. LTE, 90% confidence interval)

Figure 2.5: Study *Trigger 1* and *Trigger 2*: DiffMOS results

As depicted in Figure 2.5(a) there was a significant labeling effect in study *Trigger 1* for Web usage — browsing a news site or Google Maps via a labeled 3G or ADSL connection — for *low QoS* scenarios (Note: only the significances of the $DiffMOS$ values are considered). Here, users were more tolerant if a wireless-labeled connection was used in contrast to a wireline-labeled connection (Note: both connections were technically identical). Figure 2.6(a) shows that 71% of the $DiffRatings$ were significantly positive, i.e., wireless-labeled connections were evaluated more positively whereas only 17% of the ratings indicate an opposite effect. Table 2.2 additionally shows the results of the Wilcoxon rank sum test (see Section 1.3 for more details). Here, the significance of the impact for the labeling effect for low QoS and browsing Google Maps is *not* confirmed, i.e., the p-value of the Wilcoxon rank sum test is not < 0.05. For mid/high QoS condition — i.e. users experienced smaller page load times compared to low QoS conditions— the labelling effect was relatively weak: Only for browsing a News Site with high downlink bandwidth levels the users were more tolerant when a wireless connection was used, see Figure 2.5(a). For browsing Google Maps, Figure 2.7(b) shows that a positive effect only occurred for approx. 33% of the users when a high downlink bandwidth was applied. Also, according to Figure 2.5(a)(red bars), the differently labeled Internet connections had no impact on the subjectively perceived video quality, i.e., the $DiffMOS$ values are not significantly different from 0.

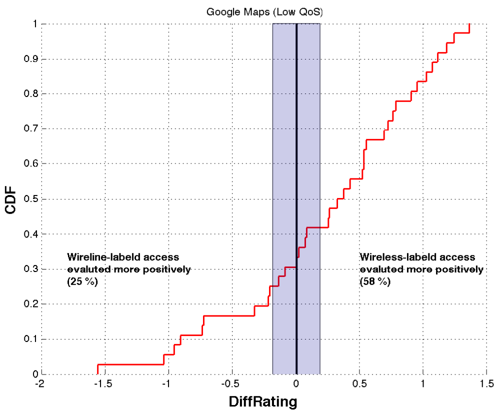


(a) News site - low QoS

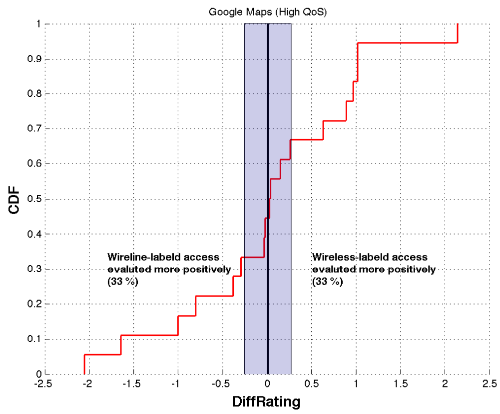


(b) News site - high QoS

Figure 2.6: Study *Trigger 1*: CDF-plots for browsing a news site



(a) Google Maps - low QoS



(b) Google Maps - high QoS

Figure 2.7: Study *Trigger 1*: CDF-plots for browsing Google Maps

In contrast to this, in the second labeling study *Trigger 2* in 2013, no significant labeling effect occured except for browsing Google Maps and high QoS conditions, see Fig-

		$DiffMOS$	median	standard deviation	confidence interval	$p_{ranksum}$
Low QoS	news site	0.5511	0.8500	0.9258	**0.3108 - 0.7914**	**0.0345**
	Google Maps	0.2436	0.3450	0.7215	**0.0563 - 0.4309**	0.3133
	video	-0.0663	-0.0350	0.6559	-0.2365 - 0.1039	0.7574
Mid QoS	news site	NaN	NaN	NaN	NaN	NaN
	Google Maps	0.1778	0.1600	0.8596	-0.0453 - 0.4009	0.2891
	video	0.1022	0.0150	0.5320	-0.0359 - 0.2403	0.6524
High QoS	news site	0.5274	0.5700	0.9243	**0.2875 - 0.7673**	**0.0121**
	Google Maps	0.0422	0.0250	1.0174	-0.2218 - 0.3062	0.8867
	video	0.0116	0.0100	0.4023	-0.0928 - 0.1160	0.9678

Table 2.2: Study *Trigger 1*: Overview about ratings (90% confidence interval)

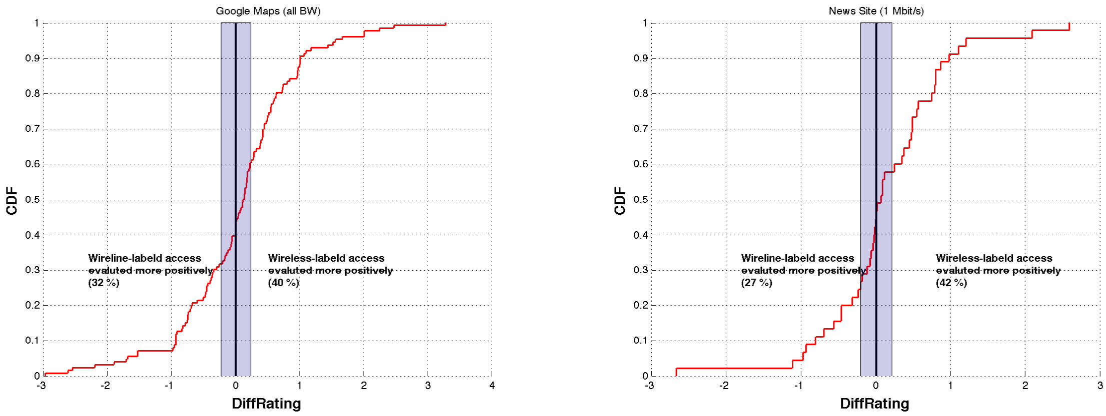
		$DiffMOS$	median	standard deviation	confidence interval	$p_{ranksum}$
1 Mbit/s	news site	0.1353	0.0600	0.8410	-0.0683 - 0.3389	0.3813
	Google Maps	-0.0371	0	1.1203	-0.3083 - 0.2341	0.8527
4 Mbit/s	news site	0.0011	0.0200	0.7760	-0.1868 - 0.1890	0.9775
	Google Maps	NaN	NaN	NaN	NaN	
14 Mbit/s	news site	-0.2178	-0.0700	0.9280	-0.4425 - 0.0069	0.3794
	Google Maps	-0.0988	-0.0100	0.8567	-0.3062 - 0.1086	0.4242
30 Mbit/s	news site	-0.0200	-0.0550	0.6730	-0.1829 - 0.1429	0.3794
	Google Maps	NaN	NaN	NaN	NaN	
45 Mbit/s	news site	0.0624	0.0600	0.8279	-0.1380 - 0.2628	0.8244
	Google Maps	0.3045	0.2950	0.8498	**0.0988 - 0.5102**	**0.0319**

Table 2.3: Study *Trigger 2*: Overview about ratings (90% confidence interval)

ure 2.5(b) and last line of Table 2.3. Also the CDF plots of study *Trigger 2* indicate no significant influence (cf. Figure 2.8).

Hence, there is a difference regarding the Internet connection labeling-effect between the studies *Trigger 1* (2011) and the study *Trigger 2* (2013). The main difference between these two experiments is the naming of the wireless connection: In 2011, the participants thought they used a *3G/UMTS* connection, whereas in 2013 the participants thought they used a *4G/LTE* connection. To get more information about user expectations, the participants were asked about their expectations regarding Internet access via a questionnaire[11], the results are shown in Figure 2.9: In 2011, there was a clear difference between the user expectations regarding high-speed wireline and wireless Internet access, i.e., for wireless connections like 3G high speed was not absolutely mandatory. In contrast to this, in 2013 there were no differences regarding the expectations of high speed Internet connections between wireline and wireless access, see blue circle/arrow in Figure 2.9. Additionally, there was also a change regarding connection reliability expectations from 2011 to 2013: Whereas both connection types should be very reliable in 2011, in 2013 the users did not expect the same reliability for LTE, see red circle/arrow in Figure 2.9. One possible explanation is the varying private usage duration, which was also evaluated via a questionnaire: in 2011, the test participants had used a mobile

[11] The questionnaire can be found in Annex A.



(a) Google Maps - All BW

(b) News Site - low QoS

Figure 2.8: Study *Trigger 2*: CDF-plots Google Maps and News Site

Internet connection for 15 months on average, whereas in 2013 the average duration of mobile Internet access was 28 months.

Also, specific previous Internet usage behavior affects the assessment of specific expectations: regarding the study *Trigger 2*, there is a significant correlation between the used download bandwidth at home and the consent to the question ”Fixed high-speed Internet access is very important” (Kendall's tau coefficient=0.46 with significance level<0.05; Spearman's rank correlation coefficient=0.53 with significance level<0.05). Also the increased usage of YouTube positively correlates with the consent regarding the question ”Mobile high-speed Internet access is very important.” (Pearson's correlation coefficient=0.44 with significance level<0.05). And the duration of usage of mobile Internet access via 3G-modems negatively correlates with the consent to the question ”I expect 100% stability from my fixed Internet access.”, i.e., more experience with mobile Internet usage reduces the expectations regarding stability of Internet connections at home (Pearson's correlation coefficient=-0.67 with significance level<0.05). Nevertheless, the presented findings can only be seen as preliminary, initial attempts to examine the emergence and adaption of expectations. In the context of Zeithaml's concept of desired and adequate expectations, these questionnaires relate to adequate expectations.

In [41] it is stated that advertising is a main influencing factor to shape expectations. Hence, to examine this specific aspect of emergence of expectations, the test participants of study *Trigger 2* were asked about their media consumption behavior and their Internet usage background. There is a small, significant (Pearson's correlation coefficient=0.39, p<0.05) correlation between browsing a news site via labeled LTE access and the consumption of LTE TV commercials: users who remembered LTE TV commercials rated the perceived connection quality more critically than users who did not

[12] First question: ”Mobile/fixed internet access is standard.” Second question: ”I expect 100% stability from my mobile/fixed internet access.” Third Question: ”Mobile/fixed high-speed internet access is very important.”

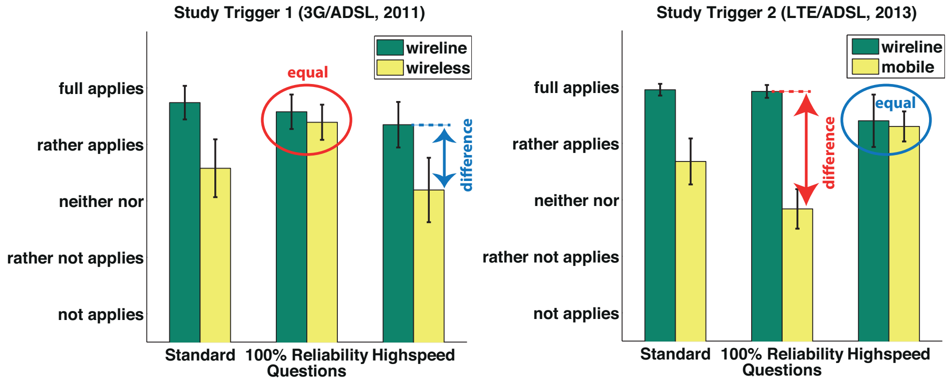


Figure 2.9: Study *Trigger 1* and *Trigger 2*: Results from the assessment of general expectations w.r.t. fixed and mobile Internet access[12]

remembered any LTE TV commercial. Consequently, the consumed TV advertisements and their promises regarding the advantages of LTE possibly lead to higher expectations, which finally results in lower quality ratings. The users were also asked about media consumption to indirectly measure advertisement views. It is assumed that higher media and higher advertisement consumption led to higher expectations and finally to lower quality ratings. This assumption could be proven for some scenarios, e.g., for the ADSL scenarios a higher newspaper consumption led to more critical ratings (Pearson's correlation coefficient=0.54, $p<0.05$). Also for the ADSL download scenario higher radio consumption lead to more critical ratings (Pearson's correlation coefficient=0.31, $p<0.05$). However, also the opposing effect has been observed: for the upload LTE scenario and the Google Maps LTE scenario, the users rated more tolerantly if they also listened to the radio more often (Pearson's correlation coefficient=0.48, $p<0.05$ / Pearson's correlation coefficient=0.41, $p<0.05$). Hence, it is rather not feasible to draw final conclusions about the influence of media/advertisement consumption on the emergence and adaption of quality-related expectations.

To sum up, it can be verified that it is possible and reasonable to trigger expectations in the context of Web QoE and pretended various types of Internet access. In study *Trigger 1*, the labeling effect occurs for *low bandwidth settings* for browsing a news site and Google Maps[13], i.e., the test participants were more tolerant regarding low QoS if a pretended 3G connection was used instead of an ADSL connection. This is in line with the findings from the expectation questionnaire (see Figure 2.9). In 2011, high-speed Internet access was expected from wireline rather than from wireless connections. This effect had vanished in Study *Trigger 2*, which took place in 2013: There are no significant differences between pretended wireline and wireless access regarding low QoS scenarios. Additionally, expectations change over time and previous Internet usage behavior and media consumption seems to have some impact on Internet connection quality expectations.

[13] $DiffMOS$ values are significant for both contents, rank sum test is only significant for browsing a news site

2.2.1.2 Presented Evidence (SD/HD video) - Study Trigger 3

In 2012, a laboratory user study was conducted in which the participants had to evaluate the video quality of short video snippets via questionnaires containing a question about perceived video quality with answering-options ranging from ”bad” to ”excellent”, see also [30]. The goal of this study was to trigger certain expectations regarding the video quality. For this purpose short textual hints were presented on a large TV screen which was also used for playing the videos. Text 1 was: ”The following video is presented in normal TV quality” and Text 2 was: ”The following video is presented in HD quality”. Each user evaluated 8 short videos: 4 videos without any previously displayed text and 4 videos with additional text. Hence, four $DiffRatings$ were calculated per user. Table 2.4 summarizes the study design. Overall, 26 users (13 female, 13 male) participated in the study with a mean age of 39.23 years (median=38.5 years).

case	reference video	displayed text	impaired video	resulting DiffRatings
A	SD	"... next video is in standard TV quality ..."	SD	$Rating_{SD} - Rating_{SD(label:SD)}$
B	SD	"... next video is in high definition quality ..."	SD	$Rating_{SD} - Rating_{SD(label:HD)}$
C	HD	"... next video is in standard TV quality ..."	HD	$Rating_{HD} - Rating_{HD(label:SD)}$
D	HD	"... next video is in high definition quality ..."	HD	$Rating_{HD} - Rating_{HD(label:HD)}$

Table 2.4: Study *Trigger 3*: experimental design

The following results were expected:

1. Case A: the hint and the presented video both relate to SD, i.e., no ”incorrect” expectations were set. Hence, it is assumed that the $DiffMOS$ is not significantly different from 0.

2. Case D: the hint and the presented video both relate to HD, i.e., no ”incorrect” expectations were set. Hence, it is assumed that the $DiffMOS$ is not significantly different from 0.

3. Case B: the hint shows ”next Video is in HD”, i.e., high video quality expectations are set. However, the following video is presented in SD, therefore the expectations are clearly not fulfilled and the $DiffMOS$ should be significantly positive.

4. Case C: the hint shows ”next Video is in SD”, i.e., low video quality expectations are set. Yet, the video is presented in HD, therefore the expectations are clearly over-fulfilled and the $DiffMOS$ should be significantly negative.

Figure 2.10 shows the resulting $DiffMOS$ values for all four cases A, B, C and D. Additionally, Figure 2.11 depicts the CDF plots for Case D and C, which support the findings presented in the following enumeration.

1. Case B, D: As described above, the empirical results confirm the assumptions how expectations should be triggered.
2. Case A: A $DiffMOS$ around 0 was expected, but the gained results show that the hint ”next video is in SD” has a negative impact on the following video quality rating.
3. Case C: A significant negative $DiffMOS$ was expected, but the result is definitely positive. The hint ”the next video is in SD” leads to lower quality ratings of HD videos compared to the HD reference ratings.

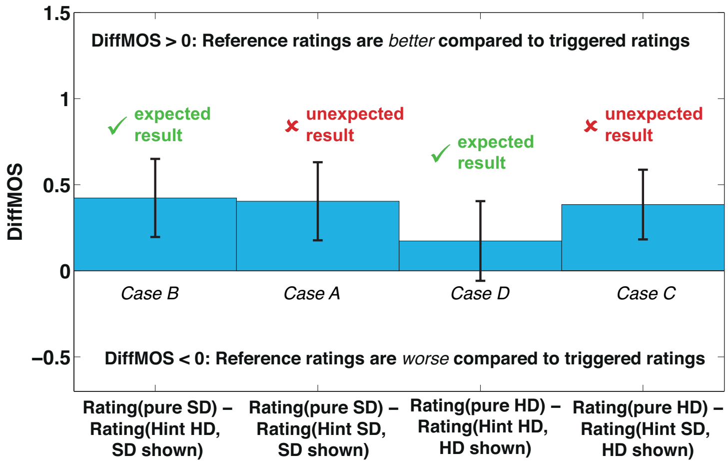


Figure 2.10: *Study 3*: DiffMOS results (90% confidence interval).

In both cases A and C with unexpected results, the hint ”the next video is presented in SD” was shown and influenced the subsequent quality ratings. When the hint ”the next video is presented in HD” was shown the resulting $DiffMOS$ values were as anticipated. Table 2.5 also shows that according to the rank sum test only the $DiffMOS$ values regarding case A and C — the ones with unexpected results — are significant. One possible explanation could be that the hint ”The next video is in SD” triggers the expectation that the following video will be presented in bad quality. This ”bad quality”-labeling subsequently impacts the actual quality rating in a negative way, i.e., the subjective quality rating is also low.

Additionally, focusing on the pure mean $DiffMOS$ values — excluding confidence intervals especially for case D — it seems that presenting a textual label about the presented video quality (SD or HD) *generally leads to lower video quality ratings*, compared to a situation in which no hints are presented. There are many possible explanations: the experimental setup is unusual/unrealistic, test participants are not able to process the video quality hints in the intended way, etc. Hence, there *might* be a general negative impact of labeling on video quality perception: if the user is *not* aware of the presented video quality — i.e. no SD/HD information is provided — the subjective quality assessment ratings will be more positive compared to situations in which the user is informed about the upcoming video quality with a textual hint beforehand.

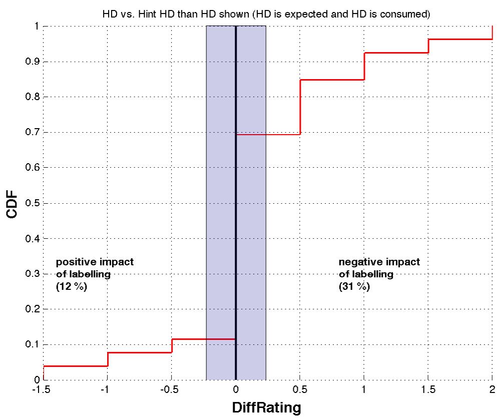
case	$DiffMOS$	median	standard deviation	confidence interval	$p_{ranksum}$
A	0.4038	0.5000	0.6785	**0.1768 - 0.6308**	**0.0342**
B	0.4231	0.2500	1.0459	**0.1961 - 0.6501**	0.0892
C	0.3846	0.5000		**0.1822 - 0.5870**	**0.0079**
D	0.1731	0	0.6920	-0.0584 - 0.4046	0.3472

Table 2.5: Study Trigger 3: Overview about ratings (90% confidence interval)

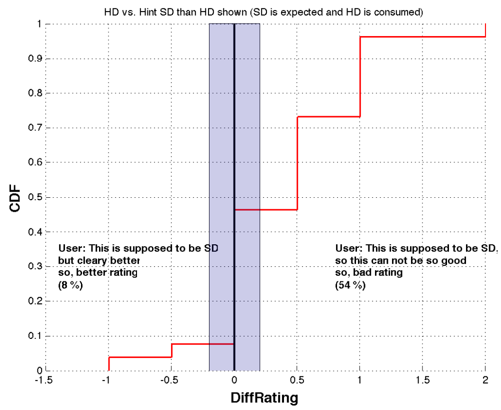
In contrast to studies *Trigger 1* and *Trigger 2*, study *Trigger 3* demonstrates that triggering expectations is complex and the gained results are not always as expected. So far, it is not possible to draw final conclusions if the methodological approach of showing textual information about the presented video quality is a fruitful strategy or not.

In study *Trigger 3* a discrete scale for video quality ratings was used, i.e., "0", "0.5", "1" etc. Hence, it is not possible to get differences smaller than 0.5. For further studies it is highly recommended to use continuous scales to get also smaller differences. Additionally, not all possible combinations of reference-videos, textual video quality hints and impaired video were evaluated. For example, the combination of a SD-reference video, a hint ”the next video is in SD quality” and a following HD video was not evaluated in the experiment. Of course, this increases the amount of conditions but additional conclusions might be possible.

The authors of [84] state, that customers evaluate the quality of a product or service more favorably if their initial expectations are high. This would mean that resulting $DiffMOS$ values would be negative if a HD hint is presented before consumption and evaluation. Nevertheless, all resulting $DiffMOS$ values are positive. Hence, the assumption of [84] cannot be validated by study *Trigger 3* in the context of Video QoE.



(a) Case D: reference video in HD; hint about HD; impaired Video in HD



(b) Case C: reference video in HD; hint about SD; impaired video in HD

Figure 2.11: Study *Trigger 3*: CDF-plots for video QoE and presented evidence.

2.2.1.3 Video on Demand Contract Classes - Study Trigger 4

In this experiment, which took place in 2011, the influence of differently priced Video-on-Demand contracts on subjective video quality perception was evaluated, i.e., it was evaluated if high-priced contracts shift quality expectations. The test participants were randomly assigned to one of three Video-on-Demand contracts (gold, silver and bronze) differing in available movies, support levels, placement of commercials, and mainly the hypothetical price the user was charged, see Figure 2.12. The three contract types were presented to the users on a large TV screen and afterwards every user was assigned to one contract with the hint to "use" this contract type while watching the following video snippets. Hence, the participants had to *imagine* to use this contract while a video was consumed.

Figure 2.12: Study *Trigger 4*: Description of the different Video-on-Demand contracts

After the participants were assigned to a VoD contract, each test user watched three short video clips from the genres Action, Documentary and Sport in three different technical quality levels on a flat screen television (h.264 encoded 1080p/i videos with average bitrates of 1000, 5500 and 8000 Kbit/s). After each video clip had been presented in a particular quality level, the user had to immediately rate the video quality using an ACR video quality evaluation scale, see [30].

Overall, 44 users (22 male and 22 female) participated in the study. The mean age was 36.8 years. Classified in age groups, approximately 40% of our users were between 18 and 29 years old, 32% were between 30 and 44 years old and 28% were older than 45. Most of our users where employed (48%) while 28% were students. More than 93% of the test users were familiar with YouTube, more than 75% of them used this service once a week or more frequently. Most of the YouTube users consumed music videos (67%) while movies and fun videos were not consumed very intensively (20%). Only 5 users had made experiences with video on demand platforms (2x iTunes, 2x A1 Video store, 1x UPC on demand; reminder: study took place in 2011). Each month, they spent between 2,5 € and 9,9 € on such services, resulting in a mean of 5,48 €.

Figure 2.13 depicts the rating results for all three content classes (action, soccer, documentary). For higher video bitrates (1000 and 5500 kBit/s), there were no clear differences regarding the assigned VoD contract. There is a small, but not significant tendency that users with a silver contract evaluate the presented video quality more critically than users with a gold or a bronze contract. It seems that the experimental setup (i.e. test participants had to imagine to use the VoD-contract while evaluating the presented video quality) might not be severe enough. On the other hand, in other research fields it is common to ask participants to only assume a certain context: in [**?**], the subjects were asked to assume they were to stay at a specific hotel during a business trip. But in the context of QoE this approach seems to be less conducive. Instead of imagining a certain contract, a more realistic approach should be used, which could be, for example a field study with real VoD services.

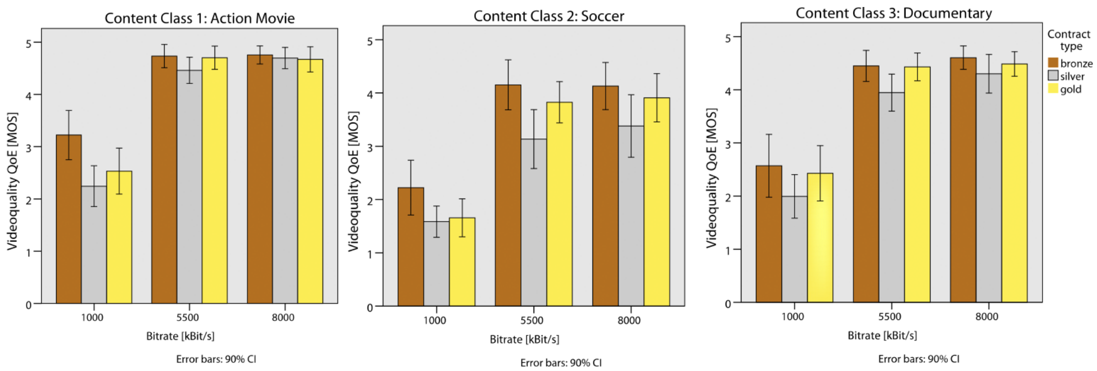


Figure 2.13: Study *Trigger 4*: MOS video quality assessment results for different contract types and quality levels (90% confidence interval).

2.2.1.4 Conclusions about Triggering

This section demonstrated that triggering expectations is a complex task and not all experimental test setups lead to satisfying results. Hence, it is recommended to utilize sophisticated triggering mechanism like the ConnectionSwitcher used in studies *Trigger 1* and *Trigger 2*. Beside its realistic functionality (e.g. blinking LEDs), the context of the switch was plausible, i.e., accessing the Internet via different connection technologies. In contrast to this, the trigger in study *Trigger 3* — presenting a hint on a TV screen before a video is shown — might be too artificial and users might get confused by it. Also, asking participants to just *imagine* a certain situation — like using a certain VoD contract like in study *Trigger 4* — might not be sufficient to trigger expectations in a valid way.

2.2.2 Assessment of Quality Expectations

In Section 2.1.1 the concept of desired and adequate expectations [13] was discussed. The following section describes a novel approach how to *quantify* individual expectations regarding quality based on the differentiation between desired and adequate expectations. As already stated in Chapter 2.1, quantifying quality expectations is a significant challenge. In this regard, the presented method is a first approach to address this challenge. Table 2.6 provides an overview about the user studies which included expectation assessment questionnaires.

Assessment of expectations has been done in QoE related research before, e.g., in the experiments described in [68] the users of the QoE study had to describe their actual Dropbox usage experience (e.g., "Much worse than I expected" or "Much better than I expected"). Also, in [85] the users were asked: "How consistent or inconsistent were the functions of the interface elements with your expectation?". Nevertheless, the approach described in this thesis aims to get information about expectations *before* any specific test condition or task has been executed. Hence, a questionnaire design is needed which supports this research approach. In general, users fill out these questionnaires *before* a particular evaluation task is conducted.

ID	year	type of expectation	expectation ratings about	study objective	n
Measuring 1	2012	desired	telecommunication provider	Web QoE	41
Measuring 2[14]	2014	desired	VoD provider	video QoE	35
Measuring 3	2013	adequate	specific Internet connections	Web QoE	45

Table 2.6: Overview about expectation assessment studies.

2.2.2.1 Assessment of Desired Expectations

According to [13], desired expectations are rather stable and are fairly independent from context. For example, some users are *generally* more economy-driven than quality-driven, i.e., for them it is more important to save money than to spend more money in exchange for higher technical quality. On the other hand, there are generous users who *generally* prefer higher technical quality, which is of course more expensive. If they were asked directly, users would clearly state that both aspects *quality* and *price* are relevant for them. Therefore, *directly* asking might not be the best approach to obtain information about desired quality-related expectations. Hence, one way of indirect questioning is to use *ranking questions*, which are common in other research fields, e.g., consumer research, see [86]. The concept of ranking questions was firstly applied by Rokeach [87], who examined the importance of individual values. In his surveys participants had to arrange 18 values (true friendship, mature love, self-respect ... etc.) into an order of individual importance to them. Obviously, since not all worthwhile values can be evaluated as most important, a trade-off is needed.

Hence, to get information about desired expectations, the test participants of the studies *Measuring 1* and *Measuring 2* (see Table 2.6) were asked to rank features of service providers, i.e., telecommunication providers or Video on Demand (VoD) vendors. In the Web QoE study *Measuring 1* the test users had to rank preferable features of Internet service providers regarding their individual importance: high network speed,

[14] Study Measuring 2 is identical with the study WTP 2, which is discussed in Section 3.1.4, but for reasons of lucidity in the context of the expectations-related section the study is also labeled as Measuring 2.

item	ranked position 1st	2nd	3rd	4th	5th	∅	pos.
high network speed	17	17	5	2	0	1.80	1
low monthly fee	18	9	11	3	0	1.98	2
short contract commitment	0	4	7	18	12	3.92	4
good support via E-Mail & telephone	0	1	6	13	21	4.31	5
unlimited download volume	6	10	12	5	8	2.98	3

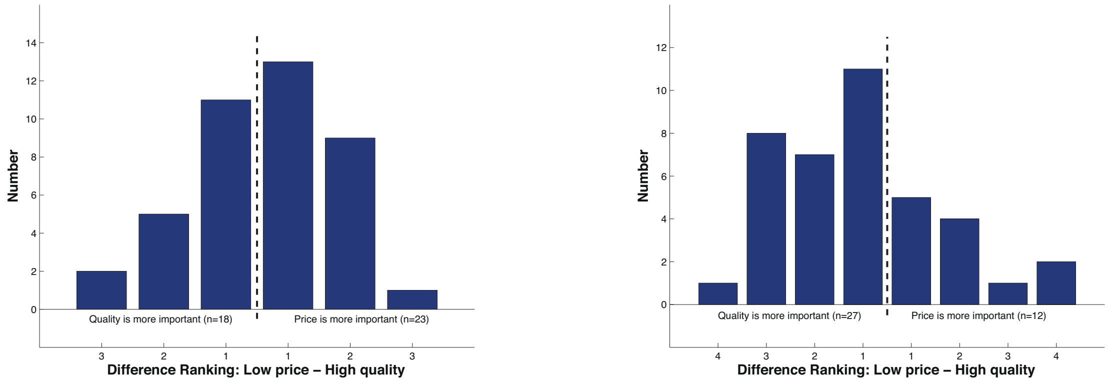
Table 2.7: Study *Measuring 1*: Desired expectations regarding Internet service providers

low monthly fees, short contract commitment, good support via email and telephone & unlimited download volume. The ranking of the item "high network speed" has been used as proxy for gauging the individual desired expectations regarding the desired, technical quality of an Internet connection at home. For example, a generous, quality-aware person would rank the item "high network speed" on top (=1) and the item "low monthly fee" somewhere below, e.g., rank 3 or 4. In contrast to that, a money-saving person would rank the item "low monthly fee" on first position. In the video QoE study *Measuring 2* information about desired expectation of the experiment participants regarding VoD providers was gathered. The rank items were: large amount of available movies, excellent support, low costs, excellent video quality and short contract duration.

item	ranked position 1st	2nd	3rd	4th	5th	∅	pos.
high video quality	12	3	9	12	3	2.77	3
low costs	14	16	5	2	2	2.03	1
short contract commitment	2	9	11	10	7	3.28	4
good support via E-Mail & telephone	1	0	4	11	23	4.41	5
large amount of videos	10	11	11	3	4	2.49	2

Table 2.8: Study *Measuring 2*: Desired expectations regarding video on demand providers

Tables 2.7 and 2.8 show the resulting ranking distributions for study *Measuring 1* and *Measuring 2*: For desired expectations in the field of telecommunication providers the item "high network speed" and "low monthly fee" are similarly important for the user (positions 1 and 2, red circles in Figure 2.7). High network speed is mostly ranked on 1st and 2nd positions, whereas low monthly fee is ranked mostly on 1st and 3rd position. In contrast to this, desired expectations in the field of Video on Demand Service providers are different: low costs (position 1) are more important than video quality (position 3). Figure 2.14 shows histograms of the differences between the ranking positions of the items "low costs" and "high quality" for both studies *Measuring 1* and *Measuring 2*. Interestingly, regarding the difference distribution for study Measuring 2 (Figure 2.14(b)), for many users the quality is more important than the costs. Hence, to get valid results regarding desired expectations the calculation of the average rank is not sufficient, but also the distribution of the differences should be considered.



(a) Study Measuring 1: Difference between price and quality desired expectations

(b) Study Measuring 2: Difference between price and quality desired expectations

Figure 2.14: Study *Measuring 1* and *Measuring 2*: Histogram of differences between ranking positions of the aspects ”high quality” and ”low price”

To evaluate the validity of the information about individual desired expectations gathered in study *Measuring 2* via questionnaires, this data is used in the context of *preference mapping*: In [88], two ways are presented how quality features (e.g. the nosiness of a speech signal, the loudness of a speech signal, etc.) can be described, i.e., via a *vector model* or via an *ideal point model.* Whereas quality features outlined via the vector model are described as ”the more, the better” (e.g. naturalness of a speech signal), quality features described via the ideal point model should be close to this ideal point for the users. To evaluate the collected data about desired expectations the ideal point model is utilized. For more information about preference mapping and its application in the context of QoE please see [56], [89] and [90].

In Section 3.1.4 the empirical quality assessment part of study *Measuring 2* will be discussed in detail, for applying the ideal point model only a short description of the experiment is provided: In study *Measuring 2* — besides filling out questionnaires about desired expectations regarding VoD providers — the test participants had to spend real money to increase the technical video quality of a selected VoD movie. Afterwards, each user had to evaluate the perceived quality via a quality questionnaire. Hence, for each user the following information is available:

- Desired expectations about technical video quality (via questionnaire)
- Desired expectations about VoD fees (via questionnaire)
- Selected technical video quality of the consumed VoD movie
- Price of this selected technical video quality
- Rating of the video quality (5 point scale ranging from ”bad” to ”excellent”)

Figure 2.15(a) visualizes the ideal point model applied on the data of study *Measuring 2*: For each test participant k information about desired costs C_{ki} and desired quality Q_{ki} regarding VoD is used to define the *ideal point*. During the *Measuring 2* user study each user selected a certain video quality Q_{ks} including a specific cost C_{ks}, which is represented as *stimulus* in Figure 2.15(a). The distance d_{kis} represents the theoretical assumption how the participant evaluates the QoE: smaller distances should correlate with higher QoE assessment ratings, greater distances should correlate with lower QoE assessment ratings. Before the experiment results can be applied, a normalization of the data is required: In the *Measuring 2* user study, the information about desired expectations regarding VoD was collected via ranking questions, e.g., for a specific user the desired video quality is stored as ”1” and desired cost is stored as ”3”. Therefore, in this example good video quality has highest priority for this user (position 1 of 5) and low costs only has average priority (position 3 of 5). In contrast to this, the experienced *stimulus* is represented by specific values, for example a movie is consumed with a video bit rate of 1448 kBit/s related to a price of 1.4€. For a better representation of the selected video quality the video bit rate values are converted to R-Ratings [81]. Similar to MOS values, R-Ratings also represent subjective quality assessment via values between 0 and 100 (instead of values between 1 and 5 for MOS). To translate the technical video bit rate, the QoS/QoE-model from Section 3.1.4.2 is used, a detailed description can be found there. Equation 2.2 describes the model:

$$R = 100 - 80.54 * e^{-0.00104 * VideoBitRate} \tag{2.2}$$

Furthermore, to combine the ideal point related values with the stimulus related values, all figures are normalized between 0 and 1, Table 2.9 depicts the normalization. Hence, the distance d_{kis} for user k between the stimulus and the ideal point can be calculated:

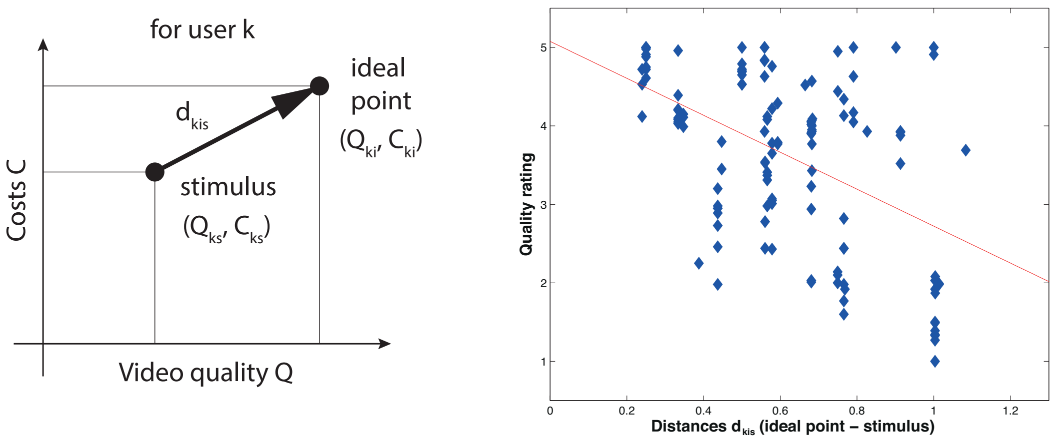
$$d_{kis} = \sqrt{(Q_{ki} - Q_{ks})^2 + (C_{ki} - C_{ks})^2} \tag{2.3}$$

Figure 2.15(b) depicts the correlation between the calculated distances d_{kis} and the related quality ratings (Pearson correlation coefficient r=-0.4612, p<0.0000). If the distance is small (the quality features of the stimulus are close to the quality features of the ideal point), the resulting quality ratings are closer to 5 (=excellent quality). For larger distances, the resulting quality ratings are closer to 1 (=bad quality). Hence, the applied questionnaire is appropriate for collecting information about the desired expectations[15].

[15] In the ideal point model introduced in [88] and discussed in [56], additional weighting coefficients for the quality features are used. Because of missing additional individual information, these weighting coefficients are not included in the used ideal point model in this thesis.

variable	quality feature	empirical values (min,max)	normalized (min, max)
Q_{ki}	desired expectation: importance of video quality	5 (not important) 1 (most important)	0, 1
C_{ki}	desired expectation: importance of low costs	5 (not important) 1 (most important)	0, 1
Q_{ks}	stimulus: R-ratings calculated from the video bitrate	0, 100	0, 1
C_{ki}	stimulus: costs [EUR]	0, 2	0, 1

Table 2.9: Study *Measuring 2*: Normalization of stimulus and ideal point values.



(a) Ideal point model

(b) Correlation between quality ratings and distances d_{kis}

Figure 2.15: Evaluation of desired expectation questionnaires via the ideal point model.

2.2.2.2 Assessment of Adequate Expectations

According to the authors of [13], adequate expectations are — in contrast to the more stable desired expectations — more flexible and are influenced by the context. To quantify adequate expectations the test participants in study *Measurement 3* were asked several questions which included specific details about expected quality. For example, in the context of Web QoE the users were asked about specific tasks (e.g. browsing a news site) in a specific context (e.g. accessing the Internet at home). For all questions there were 5 answering options, e.g., for the question regarding the download duration there were the 5 answering options "10 seconds", "30 seconds", "1 minute", "1 minute 30 seconds" and "2 minutes". For the sake of comparability, for calculating the results presented in Table 2.10 the position of the item was used — 1, 2, 3, 4 and 5 — instead of dedicated values, e.g., the duration in seconds or minutes.

- "How fast should your {home|mobile} Internet access be when you browse a news site (answer: Mbit/s)"

- ”How fast should your {home|mobile} Internet access be when you download a 50 MByte file (answer: Mbit/s)”
- ”How fast should a Web site be loaded at {home|mobile} via your Internet access (answer: seconds)”
- ”How long should it take to download a 50 MByte file when you use your home/-mobile Internet access (answer:seconds)”

These resulting 8 questions were accompanied by 8 questions regarding the answering difficulty (5 answering options ranging from ”very easy to answer” to ”very difficult to answer [I could not answer it]”). Hence, each test participant had to answer 16 questions. Table 2.10 depicts the resulting mean and standard deviation (SD) values. When asked about *durations* — e.g. ”How long should it take to download a 50 MB file?” — the users stated that these questions were easier to answer than the questions about specific technical quality features (downlink throughput in Mbit/s). Hence, to get information about adequate expectations from users regarding a certain situation (e.g. downloading a 50 MB file at home) it seems that it is more expedient to ask about directly visible quality features like waiting/downloading time.

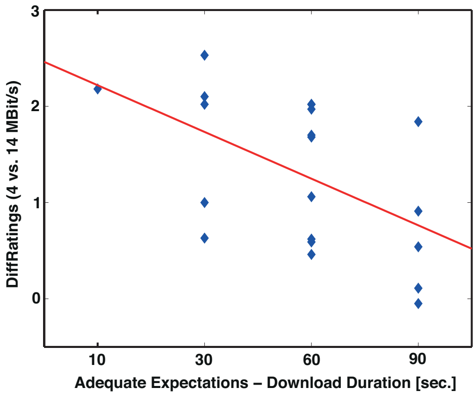
To evaluate the validity of the used questionnaire, the gathered information about individual adequate expectations was combined with user quality ratings, which were also gathered in study *Measuring 3*: the test participants had — besides filling out questionnaires about adequate expectations — to download a 50 MB file via various downlink bandwidth levels (4, 14 and 45 Mbit/s). Afterwards, the users had to evaluate the perceived quality via a 5-point rating scale ranging from ”excellent” to ”bad”. For each user the $DiffRating$ of 4 and 14 Mbit/s downlink bandwidth has been calculated. Figure 2.16(a) depicts the correlation between decreasing adequate expectations (”How long should it take to download a 50 MB file via LTE”) and a decreasing $DiffRating$ (Pearson correlation coefficient r=-0.52, p=0.0173). The correlation indicates that if users state a longer download duration (lower adequate expectations), the $DiffRatings$ are smaller, which shows that users with low adequate expectations are less sensitive regarding low bandwidth levels.

Figure 2.16(b) depicts the boxplot of the adequate expectations separated by the binary acceptance question ”Would you accept this quality at home” which was asked after the test participants had downloaded a file with 4 Mbit/s downlink bandwidth.[16] Test participants who accepted the evaluated bandwidth also stated a higher acceptable downlink duration (average adequate expectation = ca. 60 seconds) compared to users who did not accept the bandwidth (average adequate expectation = ca. 30 seconds). Also the Wilcoxon rank sum test indicates a significant difference between users who accept and those who do not accept 4 Mbit/s as downlink bandwidth in the case of file downloading (p=0.0052). Hence, the proposed method of using questionnaires to receive individual information about adequate expectations is a valid approach.

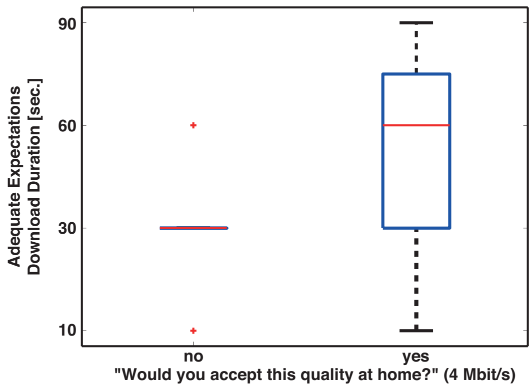
[16] In the study *Measuring 3* also the bandwidth levels 14 and 45 Mbit/s were evaluated by the users but regarding the resulting short download durations all test participants accepted the downlink bandwidth.

		fixed Internet access at home				mobile Internet access			
		expectation		difficulty		expectation		difficulty	
		mean	SD	mean	SD	mean	SD	mean	SD
site	dur.	1.48	0.63	1.61	0.79	2.09	0.74	1.93	0.97
	TP	3.98	1.52	2.70	1.53	2.55	1.50	2.89	1.35
DL	dur.	2.16	0.94	2.05	1.08	2.73	0.95	2.10	0.94
	TP	2.82	1.78	2.82	1.50	3.61	1.62	2.80	1.50

Table 2.10: Study *Measuring 3*: Results regarding assessment of adequate expectations. Red circles indicate avg. most difficult question for the users, green ones indicate avg. easiest question. Site=Browsing a Website, DL=Download of a 50 MB file, dur.=Duration of download/page load. TP=Throughput in Mbit/s



(a) Correlation between DiffRatings and stated adequate expectations



(b) Boxplot of stated adequate expectations separated by acceptance.

Figure 2.16: Study *Measuring 3*: Correlations between adequate expectations and user ratings.

2.2.2.3 Conclusions about Expectation Assessment

This section demonstrated how to get information about individual desired and adequate expectations via dedicated questionnaires. Since this is an initial attempt to get quantified information about expectation in the context of QoE, an evaluation of this approach is rather difficult. Hence, to determine if this approach is valuable the gathered information is used for QoE modeling in the following section.

2.2.3 Extending Quantitative QoE Models with Information about Expectations

Whereas in the previous section ways if collecting information about desired and adequate quality expectations were analyzed, in the following section it will be demonstrated how this information can be utilized to extend QoE models in order to enhance MOS prediction accuracy. Table 2.11 provides an overview about the involved user studies.

ID	year	type of expectations used for modeling	study objective	n
Measuring 1	2012	desired	Web QoE (Google Maps)	41
Measuring 3	2013	adequate	Web QoE (file download)	45
Measuring 4	2014	adequate	Web QoE (Google Maps)	29
Measuring 1	2012	adequate & desired	Web QoE (news site)	41

Table 2.11: Overview of studies for modeling

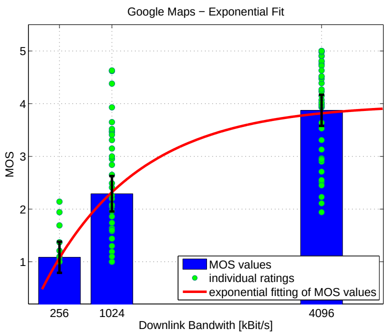
In general, there are several methods to generate quantitative QoS/QoE-models, e.g., via machine learning techniques like decision trees [91] or neuronal networks [92]. It is also common to gain less complex solutions to describe the relationship between technical and perceived quality, e.g., curve fitting, see [93] for an example. In [93], the relationship between various initial delay lengths in music and video streaming scenarios and the perception of these delays (e.g. how annoying was it for the user?) is modeled via a logarithmic relation, whereas in [94] the authors demonstrate that exponential functions are an appropriate way to model the relation between bandwidth and MOS for Web applications. However, both approaches — analytical approach via curve fitting and machine learning — have their justifications, but to demonstrate how quantified information can be included in QoE-models the transparent approach of curve fitting is more adequate than a machine learning based, black box approach.

Hence, in the following subsections an exponential fitting approach for the specific user studies is presented and evaluated. Subsequently, individual information about desired and/or adequate expectations is added and the extended model is evaluated to determine if this additional information enhances the MOS-prediction of the model.

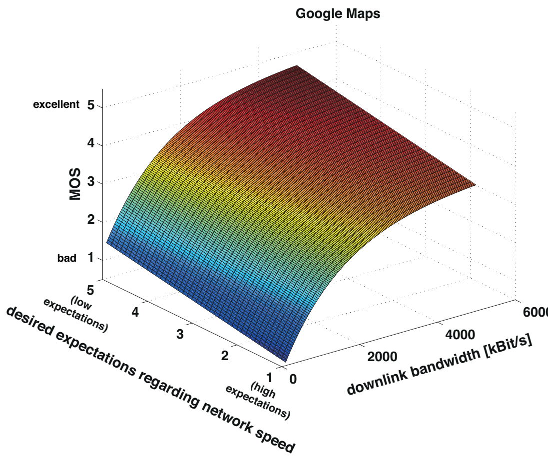
In general, the questionnaires about adequate and desired expectations were filled out by the users *before* the evaluation and quality assessment task was executed.

2.2.3.1 Modeling of Desired Expectations - Study Measuring 1, Google Maps

In study *Measuring 1* (see Table 2.11) the test participants had to browse Google Maps in satellite view mode via three different downlink bandwidth values: 256, 1024, 4096 Kbit/s. After the tasks were completed the users filled out a questionnaire to get their quality impression regarding the perceived speed of the Internet connection ("How do you perceive the speed of the Internet connection?" with answering options ranging from 1=bad to 5=excellent). Before the test participants evaluated various Internet connection bandwidth levels, information had been gathered about desired expectations via a ranking-items-questionnaire with the answering options "high network speed, low monthly fees, short contract commitment, good support via E-Mail and telephone & unlimited download volume"(see Section 2.2.2.1 for more details). Finally, the position of the ranked element "high network speed", whose value was between "1" (generally very important) and "5" (generally not important), was used to determine the desired expectation regarding network quality.



(a) Study Measuring 1: Pure Fitting



(b) Study Measuring 1: Pure and desired expectations model

Figure 2.17: Study *Measuring 1*: Modeling for browsing Google Maps

Figure 2.17(a) depicts the individual quality ratings (green dots), the resulting MOS values (blue bars) and the resulting exponential fitting curve (red line). The first line of Table 2.12 shows the resulting, pure QoS/QoE model with the resulting adjusted R^2 value (0=no fitting of the model with the underlying data; 1=perfect fitting) and the root-mean-square error RMSE.

Please note for this study and the following studies *Measuring 3* and *Measuring 4*: the fitting curves are based on *individual ratings*, which is contrary to common QoS/QoE-fitting approaches, which are based on the resulting *MOS* values (i.e. a single average value over all ratings). Hence, the resulting adjusted R^2-values (based on many individual ratings) are lower compared to adjusted R^2-values in other studies, which are mostly based on only a few MOS-values.

model type	**Model**	**adj. R^2**	**RMSE**	**add. expl. Value**
QoS	$MOS_{Maps} = 4 - 3.52 * e^{-0.00072 * DLBW}$	0.6483	0.8386	-
QoS & Exp_{des}	$MOS_{Maps} = 3.73 - 3.59 * e{-}0.0007 * DLBW$ $+0.21142 * exp_{des}$	0.6591	0.8256	+ 4.44 %

Table 2.12: Study *Measuring 1*: QoS/QoE-models evaluation

Next, the pure QoS/QoE-model was extended with the additional additive, linear factor exp_{des}, which represents individual, quantified desired expectations. The resulting model is displayed in Figure 2.17(b). This extended model has two input parameters: the technical quality via downlink bandwidth in Kbit/s and information about the individual desired expectation. Obviously, a lower desired expectation (”5”) results in a higher MOS score compared to higher expectations (”1”) (assuming that the technical quality is identical). The second line of Table 2.12 provides some additional information: the adjusted R^2 value for the extended model is higher compared to the pure QoS/QoE model and the RMSE is lower. To get information about the added explanatory value of the factor exp_{des}, the squared Pearson correlation coefficient is calculated between the residuals of the pure QoS/QoE model and the factor exp_{des}. Hence, by including information about adequate expectations the MOS prediction accuracy was enhanced by 4.44%.

2.2.3.2 Modeling of Adequate Expectations - Study Measuring 3, File Download

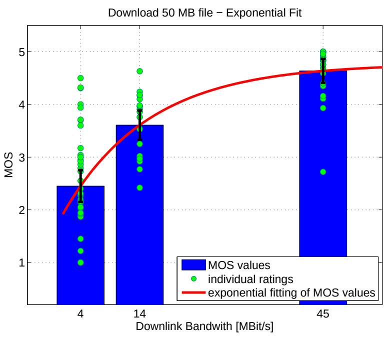
In study *Measuring 3* the test participants had to download a 50 MB file from a website (see Figure 2.18(a)) via three different downlink bandwidth values: 4, 14 and 45 Mbit/s. After the tasks were completed, the users filled out a questionnaire to get their quality assessment regarding the perceived speed of the Internet connection (”How do you perceive the speed of the Internet connection?” with answering options ranging from 1=bad to 5=excellent). Additionally, information was gathered about adequate expectations with a questionnaire which included the question ”How long should it take to download a 50 MB file at home?” with the answering options ”10 seconds”, ”30 seconds”, ”60 seconds”, ”1 minute 30 seconds” and ”2 minutes”, see Section 2.2.2.2 for details.

(a) Download Task ***(b)*** Browsing Google Maps ***(c)*** Browsing a News Site

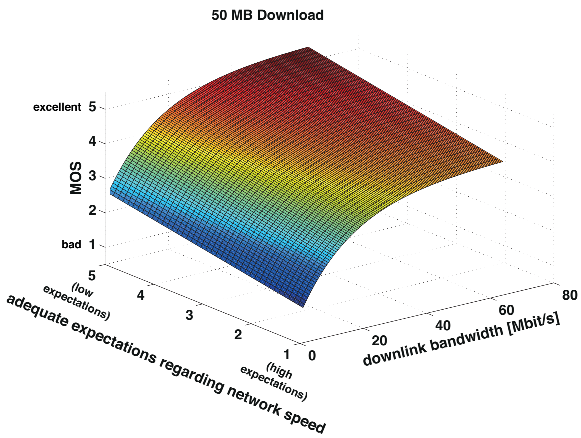
Figure 2.18: User tasks in studies *Measuring 1*, *Measuring 3* and *Measuring 4*.

Figure 2.19(a) depicts the individual quality ratings (green dots), the resulting MOS values (blue bars) and the resulting exponential fitting curve. The first line of Table 2.13 shows the pure QoS/QoE-model with the resulting adjusted R^2 value (0=no fitting of the model with the underlying data; 1=perfect fitting) and the root-mean-square error.

Next, the pure QoS/QoE-model was extended with the additional additive, linear factor exp_{ade}, which represents the individual, quantified adequate expectations. The resulting model is displayed in Figure 2.19(b). This extended model has two input parameters: the technical quality via downlink bandwidth in Mbit/s and the individual adequate expectation. Obviously, a lower adequate expectation ("5") results in a higher MOS score compared to higher adequate expectations ("1"). The second line of Table 2.13 provides some additional modeling information: the adjusted R^2 value is higher compared to the pure QoS/QoE-model and the RMSE is lower. Hence, by including the information about adequate expectations the MOS prediction accuracy was enhanced by 9.41%.



(a) Study Measuring 3: Pure Fitting



(b) Study Measuring 3: Pure and adequate expectations model

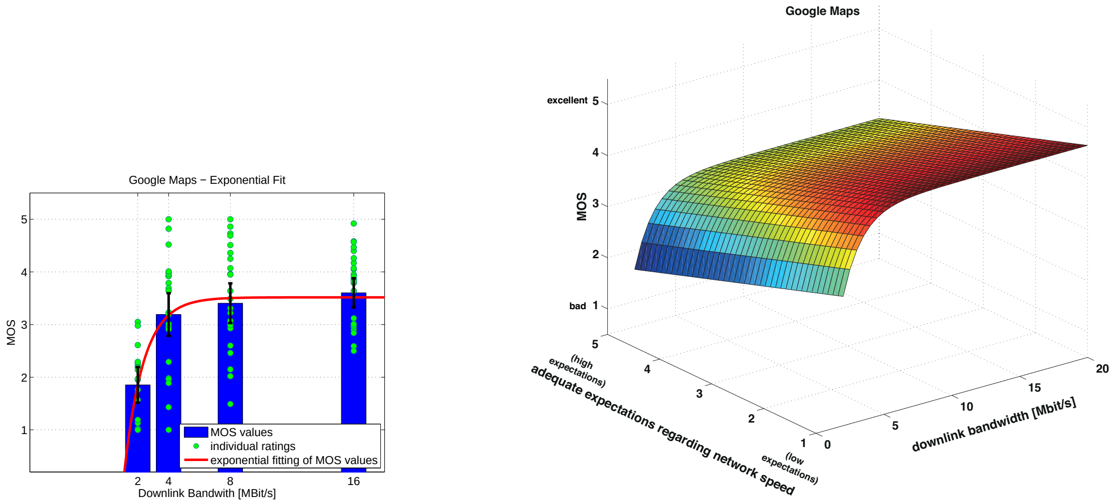
Figure 2.19: Study *Measuring 3*: Modeling for download task

Model Type	**Model**	**adj.** R^2	**RMSE**	**add. expl. Value**
QoS	$MOS_{DL} = 4.77 - 3.06 * e^{-0.07*DLBW}$	0.57	0.80	-
QoS & Exp_{ade}	$MOS_{DL} = 4.17 - 3.09 * e^{-0.06*DLBW} + 0.26 * exp_{ade}$	0.61	0.77	+9.41%

Table 2.13: Study *Measuring 3*: QoS/QoE-models evaluation

2.2.3.3 Modeling of Adequate Expectations - Study Measuring 4, Google Maps

In study *Measuring 4* (cf. Table 2.11) the test participants had to browse Google Maps (satellite view, see Figure 2.18(b)) via four different downlink bandwidth values: 2, 4, 8 and 16 Mbit/s. After the tasks were completed, the users filled out a questionnaire to get their quality impression regarding the perceived speed of the Internet connection ("How do you perceive the speed of the Internet connection?" with answering options ranging from 1=bad to 5=excellent). Additionally, information was collected about adequate expectations with a questionnaire which included the question ”How fast should your Internet connection be at home for browsing the web e.g. using Google Maps” with the answering options ”2 Mbit/s”, ”4 Mbit/s”, ”8 Mbit/s” and ”16 Mbit/s”. Note: In contrast to the previous Section 2.2.3.2, the adequate expectations relate to the expected downlink bandwidth and not to a certain duration, e.g., how long it should take to download/render a Web site.



(a) Study *Measuring 4*: Pure Fitting

(b) Study *Measuring 4*: Pure and adequate expectations model

Figure 2.20: Study *Measuring 4*: Modeling for browsing Google Maps

Figure 2.20(a) depicts the individual quality ratings (green dots), the resulting MOS values (blue bars) and the resulting exponential fitting curve (red line). The first line of Table 2.14 shows the pure QoS/QoE-model with the resulting adjusted R^2 value (0=no fitting of the model with the underlying data; 1=perfect fitting) and the root-mean-square error.

model type	**model**	**adj. R^2**	**RMSE**	**add. expl. value**
QoS	$MOS_{Maps} = 3.52 - 8.13 * e^{-0.79*DLBW}$	0.3319	0.8593	-
QoS & Exp_{ade}	$MOS_{Maps} = 5 - 6.83* e^{-0.72*DLBW} - 0.35 * exp_{ade}$	0.3924	0.8195	+9.34%

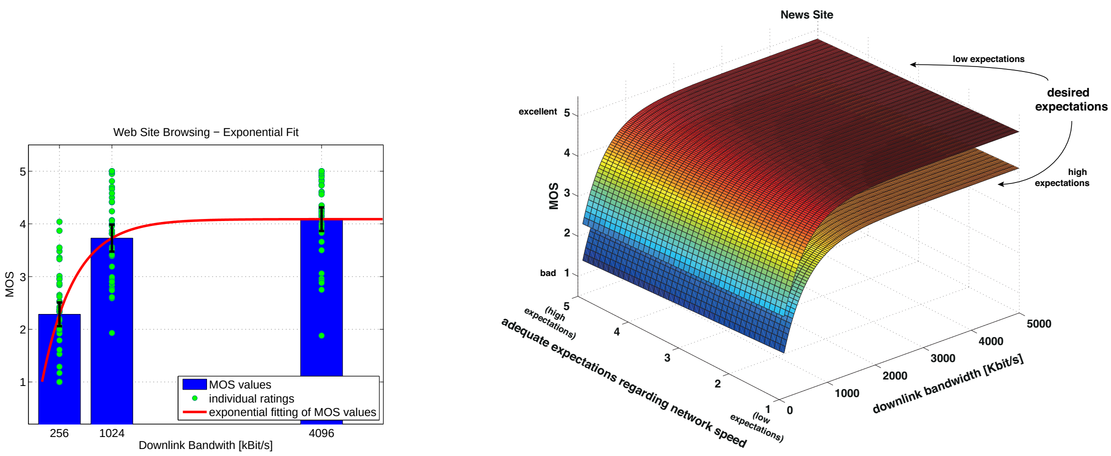
Table 2.14: Study *Measuring 4*: QoS/QoE-models evaluation

Next, the pure QoS/QoE-model was extended with the additional additive, linear factor exp_{ade} which represents the individual, quantified adequate expectations. The resulting model is displayed in Figure 2.20(b). This extended model has two input parameters: the technical quality via downlink bandwidth in Mbit/s and information about the individual adequate expectations. Obviously, a lower adequate expectation (”5”) results in a higher MOS score compared to higher expectations (”1”). The second line of Table 2.14

provides some additional information: the adjusted R^2 value is higher compared to the pure QoS/QoE-model and the RMSE is lower. Hence, by including the information about adequate expectations the MOS prediction accuracy was enhanced by 9.34%.

2.2.3.4 Modeling of Adequate & Desired Expectations - Study Measuring 1, News Site

In study *Measuring 1* the test participants had — besides using Google Maps, see Section 2.2.3.1 — to browse a news site (http://www. nachrichten.yahoo.de, see Figure 2.18(c)) via three different downlink bandwidth values: 256, 1024, 4096 Kbit/s. After the tasks were completed, the users filled out a questionnaire to get their quality impression regarding the perceived speed of the internet connection (”How do you perceive the speed of the Internet connection?” with answering options ranging from 1=bad to 5=excellent). Additionally, information was gathered about desired expectations with ranking questions (see Section 2.2.2.1 for more details). Finally, the position of the ranked element ”Importance of Network Speed” was used — the values were between 1 (generally very important) and 5 (generally not important) — to determine the desired expectation regarding connection quality. Furthermore, information about adequate expectations was collected with a questionnaire which included the question ”How fast should your Internet connection be at home for browsing the web e.g. a news site" with the answering options ”0.256 Mbit/s”, ”0.512 Mbit/s”, ”1 Mbit/s”, ”4 Mbit/s” and ”8 Mbit/s”. Note: In contrast to the previous subsection 2.2.3.2, the adequate expectations relate to the expected downlink bandwidth and not to a certain duration e.g. how long it should take to download and render a Web site.



(a) Study *Measuring 1*: Pure Fitting

(b) Study *Measuring 1*: Pure, desired and adequate expections model

Figure 2.21: Study *Measuring 1*: Modeling for browsing a News Site

Figure 2.21(a) depicts the individual quality ratings (green dots), the resulting MOS values (blue bars) and the resulting exponential fitting curve (red line). The first line of

Table 2.15 shows the pure QoS/QoE-model with the resulting adjusted R^2 value (0=no fitting of the model with the underlying data; 1=perfect fitting) and the root-mean-square error.

Next, the pure QoS/QoE-model was extended with the additional additive, linear factor exp_{des}, which represents the individual, quantified desired expectations and the additional additive, linear factor exp_{ade}, which represents the individual, quantified adequate expectations. The resulting model is displayed in Figure 2.21(b). This extended model has three input parameters: the technical quality, i.e., downlink bandwidth in Kbit/s, the individual adequate expectations and the individual desired expectations. For reasons of legibility only two desired expectation values — the lowest and highest possible values — are displayed in Figure 2.21(b). In contrast to the previous models, the overall added explanatory value +12.42% is the sum of two values: the squared Pearson correlation coefficient between the residuals of the the pure QoS/QoE-model and the factor exp_{des}, and the squared Pearson correlation coefficient between the residuals of the the pure QoS/QoE model and the factor exp_{ade}.

model type	model	adj. R^2	RMSE	add. expl. value
QoS	$MOS_{News} = 4.09 - 3.09 * e^{-0.91016*DLBW}$	0.497	0.779	-
QoS & Exp_{des} & Exp_{ade}	$MOS_{News} = 4.06 - 3.09 * e^{-0.0021*DLBW}$ $-0.066 * exp_{ade} + 0.228 * exp_{des}$	0.553	0.734	+ 12.42%

Table 2.15: Study *Measuring 1*: QoS/QoE-models evaluation

2.2.3.5 Conclusions about Expectation Modeling

In this section it has been demonstrated that it is expedient to include information about individual quality expectations in quantitative QoE models to enhance the accuracy of MOS prediction. Depending on the expectation information — desired, adequate or both — the prediction enhancement is between 4.44 % and 12.42%.

2.3 Conclusions

In this chapter of the thesis it has been shown how expectations can be systematically integrated in QoE-related research regarding subjective assessment and modeling by utilizing dedicated questionnaires for expectation assessment to improve quantitative QoE-models of MOS prediction. Additionally, based on the literature study regarding expectations in the fields of psychology, service quality and consumer satisfaction theory, an existing fine-grained QoE-model has been extended by the inclusion of desired and adequate expectations in the quality perception process. Additionally, it has been proven that triggering expectations in laboratory settings is reasonable, altthough triggering specific expectations is complex and requires precise and special experiment setups.

2.3.1 Discussion

The conducted empirical work and the literature survey have confirmed that expectations play a major role in the context of subjective quality perception related research, but have not been adequately addressed or operationalized yet. In this regard, the described methods are a first approach how to utilize the somewhat fuzzy concept of expectations in the context of QoE assessment and modeling in a systematical way. It has to be mentioned that the presented results and methods followed a consistent engineering-approach, i.e., the goal is to provide practical tools and methods for including expectations in QoE research. Hence, it might be necessary to develop and evaluate more psychology-driven models to get a better understanding how expectations are mentally set, adapted over time and how these expectations interact with quality perception in a more detailed way.

It cannot be ruled out that the questionnaires regarding desired and adequate expectations need to be revised to get more precise and valid information. Maybe there are more exact ways to gather data about desired expectations than the stated approach with ranking questions. Furthermore, fitting-based models were extended with linear factors representing expectation-related information to demonstrate how the prediction of QoS/QoE-models can be enhanced. Other integrations, however, could also lead to promising results, e.g., multiplicative instead of additive factors. Additionally, other modeling techniques extended with expectation information could lead to useful results, like machine learning, see Section 2.2.3.

The questionnaires dealing with adequate expectations used in the studies included questions about expected downlink bandwidth in Kbit/s and questions about expected download duration in seconds or minutes. Whereas questions about download duration seem to be easier for the participants, both questions are useful for modeling. Hence, correlations or interactions between the related answers should be examined to clarify if it is really necessary to include both types of questions in adequate expectation questionnaires.

3 User Decisions and QoE

This chapter describes the interaction of individual, economic/non-economic user decisions and subjective quality assessment. Three empirical user studies (WTP 1-3) in the context of pricing and video quality have been conducted to investigate the influence of individual monetary decisions on quality perception, see Section 3.1. Furthermore, three additional user studies (Decision 1-3) have been executed to explore the impact of non-economic user decisions on QoE assessment, see Section 3.2.

Figure 3.1 depicts which sections of the conceptual QoE-decision-expectation — discussed in Chapter 1.1 — are concerned in this chapter.

Parts of this chapter have been published before in [1], [3], [4], [6], [8], [9].

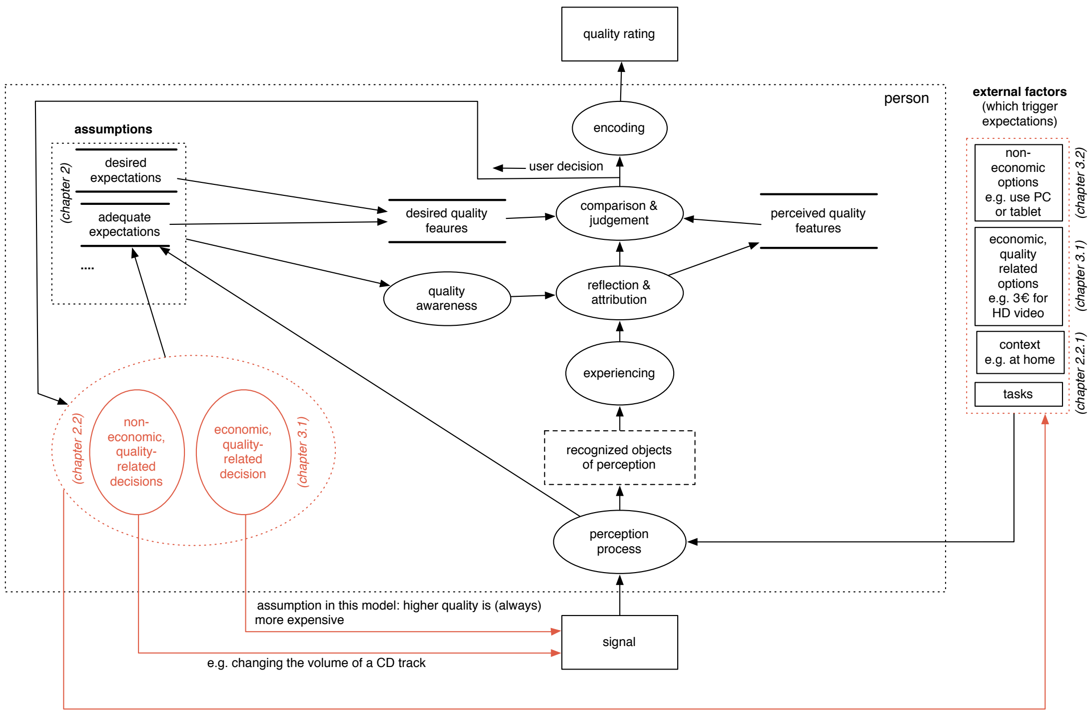


Figure 3.1: Positioning of economic/non-economic decisions in the conceptual QoE model.

3.1 Economic Decisions and QoE

In the introduction of this thesis the relevance of economic user decisions in the context of QoE was explained via the examples of differently priced mobile data plans and increasing fees for improved video quality in the context of VoD services, see Figure 1.2. To provide a theoretical basis, an overview about economic theories dealing with purchasing behavior and related work about the interaction between price and perceived quality is given in this chapter. Additionally, literature research results are presented about methods dealing with the Willingness-to-pay (WTP) of customers. Then, the first user study *WTP 1* — which was conducted in 2011 — is discussed. This experiment delivered initial insights in the correlation of purchasing and quality assessment, but also revealed methodological limitations of the used setup. Hence, a second WTP user study *WTP 2* — conducted in 2012 — is presented, which provides additional insights in pricing and quality assessment. Nevertheless, at the end of the second WTP user study it was not possible to draw final conclusions about the correlation between purchasing and quality assessment. So, a third WTP user study *WTP 3* — conducted in 2014 — focused on fewer aspects but delivered more valid results. At the end of this section recommendations are presented which can be applied in further QoE user studies including economic decisions.

3.1.1 Related Work regarding Economic Decisions

In this section the relevant literature regarding economic consumer decisions is discussed. Although non-economic user decisions are also relevant, the corresponding literature is discussed in Section 3.2.

3.1.1.1 Purchasing Decisions Theories

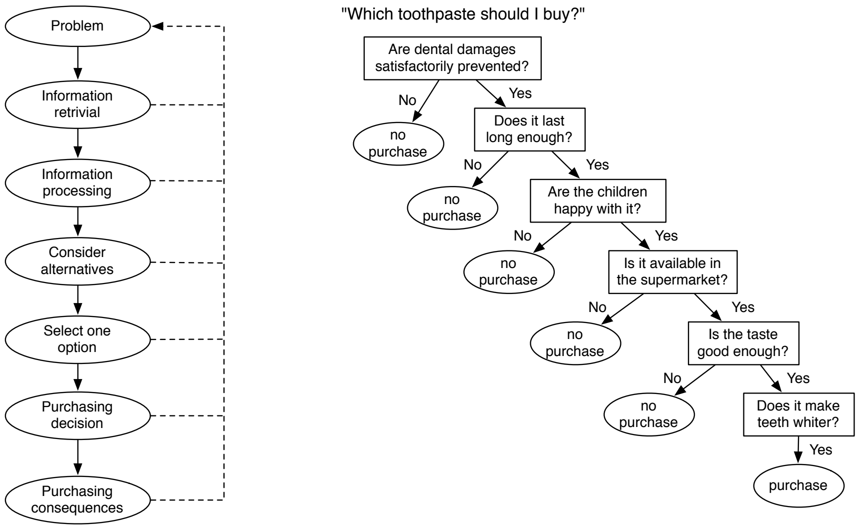
In the field of economic psychology — see for example [95], [96] and [97] — the former model of the rational homo economicus, who acts totally rationally, has been replaced by a more irrational, impulsive approach in which consumers neither assess all possible alternatives nor are totally rational. Also, decisions are made and the justification is made ex post, see [95](p.26). The so called *bounded rationality approach* — introduced by [98], see also [99] — assumes that not all alternatives are considered and not all features of a product are used for making a decision. If an option meets a certain standard, this option is selected, hence the ordering of the assessment of the available options is critical, i.e., the last alternatives are less likely to be chosen. The *implicit favorite model* — introduced by the authors of [100] — is even less rational: this model assumes that decision makers *spontaneously* select an available option. This selection is then unconsciously compared with the remaining alternatives to legitimate the made decision. Obviously, this approach requires less cognitive load than a full-rational approach which includes all available options and a comprehensive comparison process.

The authors of [101] also show that full-rational behavioral models are insufficient to describe economical decisions. According to their findings persons tend to choose the option which promised gains/risk-aversion over the option which promises losses/risk-taking, even if both options are identical regarding the outcome. This relates to the well-known *prospect theory* by Kahneman [102], which describes the human perception in the light of the odd relationship between gains and losses in a decision situation and how probabilities of these outcomes influence the decision process. Despite its relevance in economical research, the prospect theory and its implications are not fully appropriate regarding monetary decisions and QoE. For further reading please see [103].

In research dealing with purchasing behavior the concept of *involvement* describes "a person's perceived relevance of the object based on inherent needs, values and interests" [104]. In general, a high involvement of a person in a purchasing decision situation indicates that the person is willing to perform elaborate cognitive and emotional efforts to come to a decision. In contrast to this, low involvement situations correlate with low cognitive load but with moderate/high emotional activity. For example, if a person wants to buy a car she is probably in a high involvement mode, i.e., lots of information is used to make a reasonable decisions. In contrast to this, when buying a basic inexpensive item like yoghurt, the color of the packing is relevant as it triggers emotional procedures. For more information about involvement please see [105] and [106].

The authors of [107] demonstrate that there is a difference between what a test participant would *choose* and what a test participant would *state* as the best experience, which is described as *lay economism.* This means that economic factors are overvalued and other experience factors are neglected in the purchasing decision. The experiments in [107] are based on specific consumer situations like buying a stereo system or buying a dinner. Another example of irrational buying behavior is presented by the authors of [108]: The evaluation of a purchasing decision is less rational if the consumer uses available, but in all objectivity irrelevant features of the product to assess the bargain. In the stated example, an overfilled ice cream waffle serving with 200 grams of ice cream was evaluated better than an under filled ice cream waffle serving with 230 grams of ice cream.

Several models exists to describe purchasing processes. For example, Figure 3.2(a) depicts a simple approach how a potential customer made a purchasing decision: The dotted line indicates that during each step a fallback to a pervious step is possible. Nevertheless, the stated model oversimplifies the purchasing process and its practicability in research is limited, see [105]. Another approach is the model depicted in Figure 3.2(b), originally introduced by the authors of [109]. Nevertheless, these kinds of cognitive models simplify the complex processes of decision making too much. Hence, current research tries to explain only *partial aspects* of the purchasing decision. According to the authors of [105] it is reasonable to distinguish between models which consider decisions with strong cognitive aspects and models which consider decisions with weak cognitive aspects (for a complete overview please see [105]). In the context of pricing and purchasing decisions, the author of [110] describes the following approach: If a consumer has to choose from a range of similar-priced products, the price is irrelevant and the consumer selects the product with the highest quality. If the prices are different, the price of the selected product should not exceed a certain subjective threshold. If the quality and the prices differ widely and the price is relevant for the customer, the customer will select the product for which the difference between perceived utility and expenses is maximal. These assumptions have been empirically validated, please see [105] for more details. In general, demanding cognitive processes can be simplified by applying *heuristic selection rules*, cf. [105]. For example, if the *subjunctive rule* is applied several features of a product have to meet a certain standard, i.e., the product is discarded even if only a single feature does not meet the requirement. The *disjunctive rule* is easier to apply: If one feature of a product is perceived as outstanding other features are neglected and the product is purchased. For example, one heuristic disjunctive rule could be: ”Always pay the cheapest brand”. Discerning readers will have realized that this heuristic is connected with the concept of desired expectations, see Section 2.



(a) Classic Phase-model of purchasing decisions (based on [94](p.416))

(b) Example of a decision process according to [98]

Figure 3.2: Models of purchasing decisions

3.1.1.2 The Relationship between Price, Quality and Purchasing

In economical research the relationship between price, quality and purchasing has been examined to a certain extent. In common customer/vendor situations a *price-quality effect* may occur: If a potential customer is not aware of certain product-related quality aspects — e.g. an unknown restaurant in a foreign city — the customer is less price sensitive and the price is utilized as a direct quality indicator: in this case a higher price automatically indicates a higher quality. However, in situations in which prior knowledge is available or reference products are obtainable, the price-quality effect vanishes. Please see [111] for further details.

In the experiments conducted by the authors of [112] it has been shown that perceived quality correlates with higher prices, i.e., constant quality is evaluated better if prices get higher: Students had to evaluate the quality of three differently priced, but apart from that *equivalent* beers. The results show that the quality assessment directly relates to the price: the most expensive beer gets higher ratings. Nevertheless, also opposing findings are available: in [113] the author found out that the intermediate-priced product gets the highest quality ratings and in [114] the authors emphasize the role of prior familiarity with the evaluated product, demographic characteristics and product categories on price/quality relationships. Nevertheless, such experimental setups neglect several aspects which could also impact the perception of quality, e.g., the brand of a product or the brand image of the store in which a product is bought [115].

Additionally, the correlation of price and quality — the degree to which price is used as a basis for quality — is reduced if: (1) the desire for a definite opinion is low[1], (2) the information load is high and processing is difficult and (3) information is presented randomly, see [116]. To sum up, the authors of [105] point out that the price is perceived as a less important quality indicator if (1) the price and the quality of the product are used as basis for the purchasing decision, (2) the involvement is low, (3) the buying experience is high, (4) the knowledge of the product is high, (5) other quality indicators are available, e.g., brands, (6) the range of variation is small regarding quality and price in the product category, and (7) the prestige of the product is low. Hence, only in particular purchasing situations the customer uses the prices as an indicator for product quality.

In economic psychology literature it is common to categorize consumers into segments to handle different purchasing strategies. The authors of [111] propose to avoid too simple segmentations, e.g., only distinguish between price-sensitive and the quality-sensitive segments. Hence, in [111] four types of customers are described: (1) *Price buyers* seek for the lowest price which is consistent with some minimum level of acceptable quality, but no further comparisons with other vendors are made. (2) *Relationship buyers* have a strong relationship to a certain brand and its products are bought without any further comparison if the price does not exceed a certain, reasonable range. In contrast to this, (3) *Value buyers* are concerned about the price and the related quality, i.e., after carefully checking prices and features, also relatively high-priced brands or products are bought. (4) *Convenience buyers* are not brand-sensitive and also extensive price comparison is avoided, i.e., they buy whatever is available without spending too much effort in the buying/comparison process. However, it is difficult to derive conclusions from the three WTP experiments discussed later in this thesis because of their artificial setting neglecting real world aspects like additional competitors, the possibility of interrupt/abort activities, etc. In [111] the authors describe strategies how to sell products for the four types of customers. For example, price buyers are very price-sensitive and any additional feature is hard to sell. Hence, lots of additional information and/or conversations are needed to emphasize the advantage of additional features. Nevertheless, most price buyers will not change their behavior and any additional effort is useless and

[1] The desire for a definite opinion is often referred as NFC (Need for closure), see also [105]

a waste of time and money. From a company's point of view, it is crucial to understand the relationship between pricing and perceived quality to understand the user buying behavior and to derive strategies to find an optimal mix for all parties concerned, e.g., customers, vendors, network providers, etc. Additionally, not only the consumer-type is relevant, but also the context should be considered: The authors of [111] mention that the value of an offer strongly depends on the context and on the customer type, e.g., at a beach-restaurant a soda is a more expensive compared to a supermarket 300 meters away. So, for a certain customer segment a higher price is less inconvenient than a walk to the supermarket and vice versa.

3.1.1.3 Experimental Evaluation of Consumer Behavior

In the field of economics various types of customer-related experiments have been conducted to examine purchasing behavior of relevant buyer segments. These methodologies provide relevant input for planning and conducting pricing/QoE experiments.

According to the authors of [111], there are two types of experiments which examine the influence of prices on consumer behavior and the actual purchasing process: *in-store purchase experiments* and *laboratory purchase experiments*. During in-store purchase experiments, prices are changed over a certain amount of time to examine the influence of pricing on buying behavior. Additional factors can be included, for example the influence of marketing techniques. Usually, a control group is necessary which makes this kind of experiments very expensive[2]. In contrast to this, a typical laboratory experiment is conducted in a research facility, mostly located directly in a shopping mall. As an advantage it is possible to decide who participates in the experiment and many factors can directly be controlled, e.g., which products are exposed to which kind of participant. Often, this kind of research is considered to produce valid results: the experiment participant makes a real purchase – or can choose not to buy at all — and real money is used. Nevertheless, this kind of experiment is also confronted with several issues: The participant spends a lot of time on and pays attention to a particular purchasing option, which does not automatically happen in daily-life shopping situations (high involvement vs. low involvement, see previous section). In many real-world purchasing situations the price is neglected by the potential buyer but in the experimental context the participant does not want to appear careless, i.e., experiment participants are mostly aware of the fact that they are observed. Hence, some customers may try to appear smart by choosing the low-priced brand or they choose the high-price brand to avoid appearing thrifty. Also, techniques which examine purchasing intentions may evoke a "treatment effect", i.e., the process of asking triggers a purchasing action which would not have occurred without asking [117].

In the WTP/QoE user studies discussed later in this chapter the users received 10 € in advanced which could be used to increase the quality of the selected videos. One could state that this is not realistic, i.e., in real life scenarios customers have to spend their

[2] Quaker Oats conducted an in-store experiment — focusing only on pricing — running in 120 stores for three months. The effort was declared as "several million dollars" [111].

own money. But giving real money as a deposit to evaluate participants while buying real products in experimental settings is common in economic-related research, see [111]. Nevertheless, several drawbacks reduce the external validity: the participants are too rational regarding their purchasing behavior, the setup is too artificial and of course the test participants do not use their own money, see [118] for more details. Nonetheless, providing a deposit during an experiment is advantageous because test participants are more willing to spend money from a given deposit than their own money. On the other hand, a *Framing Effect* might occur — see [111] — which means that customers are price-sensitive if a buying decision is perceived as a *loss* instead of a *gain*. For example, for a private university it is better to set high tuition fees as standard which are then reduced for students via scholarships instead of starting with low fees for all with additional fees for some students (even if both ways would lead to same fees). In the WTP/QoE user studies the test participants received 10 € at the beginning, but then this amount can be *reduced* — a *loss* occurs. To avoid the situation of spending real money, it is also common to conduct so-called *constant-sum-techniques* to get information about purchasing intentions: Test participants have to divide a certain amount of money over several brands. The amount of money for each brand should represent the intention of buying this brand in an anticipated shopping situation, see [105] for more details.

To get information about purchasing behavior and price sensitivity it is also common to conduct surveys, see [111]. This kind of research is less expensive than elaborate laboratory experiments, it is applicable for large goods like cars and it can be applied even before a certain good is available. Nevertheless, the gained results are often not very reliable, i.e., answers regarding "How much would you pay for product X" cannot be validated by real purchasing behavior. This kind of determining the price for a certain quality has also been applied in QoE-related research, see [119] for details. Another example for simple asking: In [120] the author conducted a representative survey about mobile TV and payment methods. 33.7% of the participants would accept commercial breaks if additional fees were avoided. 44.9% of the participants would pay a fix amount per month (85% would pay 5 € per month or less) and 20.4% of the participants would pay per view (89% of them would pay 0.5 € or less per view). Nevertheless, these values are based on imagined situations and were not validated by real user behavior.

Also qualitative research methods can be applied to get information about buying and purchasing behavior from customers, e.g., with semistructured depth interviews, cf. [111]. Although such interviews reveal valuable information, e.g., which product feature is import for the customer and why, such techniques are costly and the output cannot always directly be transferred to quantitative models. Another qualitative approach is the usage of diaries, i.e., participants have to keep a diary about their purchasing behavior which, ideally includes lots of detailed information about feelings and thoughts which, helps to explain the purchasing decisions. Nevertheless, this approach is inappropriate for a "simple" quality-related decision in the later discussed WTP/QoE experiments, for further details please see [121] and [122].

It is also possible to artificially generate the purchasing situation during laboratory experiments, i.e., the researcher asks the participant to *imagine* that she is on a shopping trip and pictorial representations or samples are shown to the customer. For more information see [111]. Nevertheless, the process of *imagination* during a laboratory situation must be questioned. In the experiment described in Section 2.2.1 in this thesis the test participants had to imagine to using certain VoD contract which finally led to vague results.

Hence, in economic literature several methodologies are utilized to examine the Willingness-to-pay and subsequent purchasing decisions. Mostly, the output of these methods is used to plan marketing and selling strategies. In the context of QoE assessment, the term WTP is used for experiments which involve real money in quality-related situations. For a broad overview about Willingness-to-pay in the economic context please see [123] and [124].

3.1.1.4 Economical Aspects and QoE

In the context of QoE assessment some attempts have been conducted to combine economic aspects with subjective quality assessments. For example, the authors of [119] combined the concept of traditional video quality studies — test participants watched short video clips with subjective quality evaluation afterwards — with asking the participants after video consumption the binary question: "If you should pay for this video sequence, would you be satisfied with the video quality?". According to their findings, there is a linear relationship between increasing MOS values and an increasing percentage of approval regarding this question. Nevertheless, no real consumer behavior regarding video quality was examined and the video duration of 10 seconds was rather short and artificial regarding purchasing decisions. In the context of cloud gaming QoE, the authors of [125] asked the participants of a user study about their general willingness to pay for the evaluated network quality levels during gaming sessions: only 15% of the users were willing to pay a monthly fee.

Another experiment which investigated the relation between price and quality is described in the work of [32]. Here, the users were asked how much they would pay for a certain telephony connection with a specific transmission quality after the users had conducted some conversation tests. Answers about the willingness to pay were given in the currency "Deutsche Mark" and in a more abstract "points per time unit". The findings reveal that there is a strong correlation between the resulting MOS ratings and the willingness to pay, i.e., for better evaluated quality the concerned users would pay more. Nevertheless, also here the willingness to pay was investigated via questionnaires, i.e., no real purchasing behavior was observed.

In 2001, during the European FP5 project M3I[3], the authors of [126] conducted experiments to link different quality settings with consumers' monetary decisions. During these experiments, the test participants received real money (£10) as a deposit with the possibility to purchase video quality enhancements — the video bit rate of a played out video could be in- or decreased — during the test or to take the money home.

[3] http://www.m3i.org/, last access: 14th August 2015

The participants had two control options: setting a bitrate preference or setting a price preference. Both options could be altered during the test where those option choices influenced each other, e.g., if a certain price was selected and a different profile was automatically applied the bitrate changed and the costs remained constant. The test subjects watched a set of MPEG-4 videos on a computer monitor. Each of the subjects had to choose two videos of interest for the remainder of the entire test. While consuming the selected videos the prices for different video bitrates varied every 30 seconds, so the costs for watching the video in a certain quality altered. Six profiles with varying costs for certain video qualities were used. For example, in Profile 1 the lowest video bitrate (64 Kbit/s) cost £0 and the highest video bitrate (2048 Kbit/s) cost £2 while in Profile 2 the highest video bitrate cost £4. On average £2.80 were spent on increasing the video qualities. This may be explained by the low costs for consuming 512 Kbit/s coded videos (a reasonable video quality) in most of the cost profiles. Overall, users were price- and QoS-sensitive. Qualities lower than 384 kBit/s were not accepted when bitrate preferences were used. Conversely, too high prices led to user induced quality reductions. On a content level there were no differences between low and high motion video clips regarding the spent money and the selected quality. None of the subjects were willing to pay for the lowest two video qualities (64 and 128 Kbit/s). The participants also declared that they would prefer a constant video quality level instead of a constant price level. Unfortunately, the very concise result presentation in [126] omits detailed discussions regarding the users' behavior throughout the test. Especially the usage of available choices, the distribution of user types and the influence of content classes on purchasing decisions would have been of high interest from a research methodological point of view. Also, no subjective quality assessment was done during the experiments, i.e., it is not possible to draw any conclusions regarding pricing, content/quality decision and subjective quality perception. Nevertheless, some findings of the M3I experiments were used to prepare the WTP user studies discussed in this thesis.

The author of [32] states that in general there is only limited information published about findings regarding the correlation between quality and price. For example, studies have been executed in which telephone calls were free of charge but interrupted by advertisements, but the results have never been published.

3.1.1.5 QoE-based Charging

Besides the examples provided at the beginning of this section — the influence of differently priced contracts on subjective quality assessment — the correlation between pricing and subjective quality perception is also important for other QoS/QoE application fields. For example, in the context of Internet economics a shift from pricing models which only include QoS aspects towards pricing models which are based on the perceived quality can be observed. For example, the authors of [127] discuss several pricing schemes for IP-based services: "edge pricing" in which the user is only charged by the first Internet service provider for a certain quality (see [128]), the concept of "congestion pricing" includes increasing prices for congested resources link a link (see [129]) and the concept of ”resource pricing” which tries to achieve economic efficiency of the smart market by a simple packet marking scheme (see [130]). Furthermore, the authors of [127] propose a mechanism which allows the user to spend money for enhanced perceived network quality. Nevertheless, an empirical validation was not conducted, i.e., the impact of charging on QoE was not examined. Hence, for advanced QoE-sensitive pricing mechanisms the interaction between charging, individual economic decisions and subjectively perceived quality needs to be empirically investigated. For more details about QoE-based charging please see [131], [132] and [133].

3.1.1.6 Implications of Related Work for WTP-QoE-Experiments

According to the related work discussed in the previous section the following implications related to the following empirical WTP/QoE user studies can be derived:

- In Section 3.1.1.2 it was shown that the price can be used as a quality indicator. In the following WTP user studies the price *might* be used as a quality indicator because the buying experience is low, no quality indicators like brands are available, etc.
- The purchasing decision theories discussed in Section 3.1.1.1 imply the presence of several vendors providing differently priced products, i.e., a typical customer is able to choose from a broad range of options. In the subsequently discussed WTP/QoE user studies the alternatives are limited: it is not possible to switch to another Video on Demand vendor who offers a different pricing strategy and/or different content. Hence, common purchasing theories cannot be directly applied to the findings of the following studies. Nevertheless, each potential buyer has to evaluate the *value* of a particular offer, e.g., ”Is it worth for me to consume this particular video in this context via this technical quality for this price?”
- In economic research it is common to categorize consumers in customer segmentations depending on their purchasing behavior. Hence, this aspect should be considered in economic QoE experiments.
- If economic decisions concern more than one person, additional aspects of decision making are used to describe these social interactions. Nevertheless, in the following WTP/QoE experiments the focus is on single user decisions without concerning other individuals. For further information about non-individual purchasing decisions[4] please see [134], [135], [136], [137] and [138].
- It is common to provide real money in laboratory experiments examining purchasing decisions. This approach has also been used in the QoS-related video study M3I. Hence, using real money as a deposit is a promising approach in WTP/QoE user studies.

[4] For example, if a couple or a family sit in front of a TV in a VoD setting and a purchasing decision has to made regarding video quality, the decision maker has not only to include his/her own considerations, but also how his/her decision is perceived by the others.

- Survey-based methods about intended purchasing decisions are less complex and cheaper than sophisticated laboratory experiments, however, the validity of the gained results is rather limited. Hence, it is inevitable to conduct user studies to examine the influence of purchasing decisions on subjective quality assessment.
- Economic quality decisions enquired in laboratory settings evoke high involvement scenarios, i.e., the economic decision gains centre stage even if in real world this particular decision might only be experienced as low involvement. Hence, it is not easily possible to transfer the gained results of experimental user studies to real world scenarios. Additionally, in contrast to the real world, experiment participants cannot decide to just switch off the TV, i.e., it is not possible to avoid making an economic decision.
- Laboratory participants are aware of the monitoring situation, i.e., they could feel compelled to make too reasonable, unrealistic purchasing discussions.

3.1.2 WTP 1 User Study

As described in the beginning of this chapter, three Willingness-to-pay experiments were conducted in the years 2011, 2012 and 2014. Despite the different foci and technical details of the experiments, the main concept and procedure of the trials remained constant: the experiment participants received 10 € at the beginning of the study which *could* be used to increase the technical quality of individually selected videos, i.e., the participants had to make a trade-off between quality and expense. Additionally, during the studies several subjective quality evaluations were executed.

Parts of this subsection have been published before in [1], [3], [139], [4], [6] and [8].

3.1.2.1 Setup

For a better understanding, the test setup of study *WTP 1* can be divided in four segments: procedure, economic aspects, technical setup and video content.

Procedure: Overall, each test participant selected three videos of her interest with 20 minutes duration each, entailing an overall test duration of approximately one hour. Via a tablet (see Figure 3.4(a)), the user was able to browse the video content library and to choose one of the available videos. Immediately after the video had started, there was the first of four *quality selection phases* (QSP, cf. Figure 3.3). In these phases, the participant was able to try different video qualities *for free* by changing the network packet loss rate of the video transmission. Any change of the packet loss rate was immediately applied on the video stream and its effect was therefore instantly visible. At the end of each QSP (the first lasted 4 minutes, the others 2 minutes each), the last selected quality level was locked and applied during the following *no interaction phase (NIP)*, see Figure 3.3). In each of these NIP segments, the subjects were not able to change the packet loss rate, i.e., the video quality remained constant. Each QSP was visually and acoustically announced via the tablet in order to remind the subjects of the possibility to change the video quality. Moreover, there was no return policy for already purchased quality enhancements during the QSPs, i.e., the quality could only be enhanced. For example, if the user finally selected quality level 2 (out of four) after the first QSP there was no possibility to select quality level 1 in a subsequent QSP. It was only possible to increase the quality during the video duration. In Figure 3.4 the respective tablet screenshots of the content gallery from which a user could select a video (a), the video description providing details on each video (b), and the quality selection interface allowing the quality purchases (c) are depicted. At the end of each video, the user had to fill out a video quality questionnaire containing a question about perceived video quality with answering-options ranging from ”bad” to ”excellent” and the binary yes/no-question ”Would you watch this video in this quality at home?” [30].

Economic aspects: At the beginning of the test every user received a deposit of 10 €, which could be used for upgrading the video quality during the experiment. The user was aware of the fact that the remaining deposit was paid out in cash at the end of the experiment. During the QSPs of each video, the user was able to try out different quality settings for free. At the end of each QSP, the last selected quality level was locked and applied. Thereafter, the quality could not be changed during the following 3 minutes NIP until the next QSP appeared. Table 3.2 depicts the costs for watching a video in each quality as costs per minute and costs per movie. To keep the unit cost for changing the quality constant, the costs were adapted to the elapsed time based on the costs per movie. The current charge was displayed on the tablet in each QSP, see Figure 3.4(c). The prices per quality level related to the pricing scheme of the Apple iTunes store, where a movie in SD quality could be rented for 4 € (for 24 hours, fees according to 2011 apple iTunes price list). Hence, considering the video content durations a fee of 1.50 € is a reasonable price to consume the content in best SD quality. If the user spent for all four videos the maximum amount of their deposit for enhanced quality, they would at least have 5.50 € at the end of the test (10 €-3x1.5 €). Hence, it was not possible to spend the entire deposit during the experiment. With this attempt user interactions should be encouraged and a ”rip-off” situation should be avoided. The user could also watch every selected movie in the worst quality *for free* with a network packet loss rate of 1%. In other words it was possible to obtain 10 € when finishing the test. So, in contrast to the M3I experiment even participants spending the maximum amount for quality purchases received some money at the end of the test.

Technical setup: The technical details are depicted in Figure 3.5(a). The video library was hosted on a Linux server which was accessed by the test participants via a tablet and its inbuilt Web browser, see Figure 3.4(a,b). A VLC-Server streamed the selected video over Ethernet to a thin client (Mac Mini), which was connected to a 40 inch LCD TV. According to [30], the viewing distance was set to 2.8 meters, which is the appropriate viewing distance for SD content with the given screen diagonal. After selecting a movie, the user was able to modify the packet loss rate during each QSP via a simple Web interface on the Tablet (cf. Figure 3.4 (c)). The experiment was conducted in a pleasant lounge atmosphere hiding technical details from the users, i.e., the user only saw the TV set and the Tablet. As a result of some pretests and according

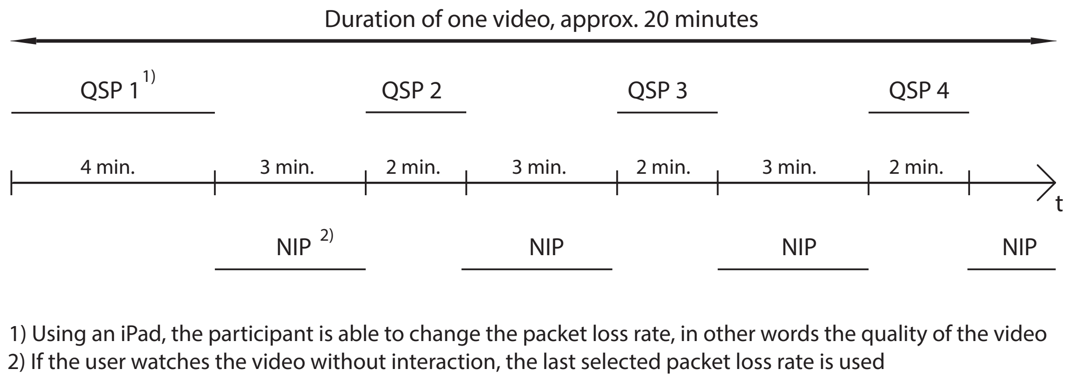


Figure 3.3: Study WTP 1: Timeline of consuming a video with marked QSP(quality selection phases) and NIP(no interaction phases)

ID	**title**	**year**	**genre**	**IMDb rating**
1	Ghost Ship	2002	movie	5.4
2	Sweeney Todd	2007	movie	7.5
3	Sweeney Todd (engl.)	2007	movie	7.5
4	Love and other drugs	2010	movie	6.7
5	Pirates of the Caribbean: At World's End	2007	movie	7.1
6	Moulin Rouge	2001	movie	7.7
7	Pulp Fiction	1994	movie	8.9
8	(500) Days of Summer	2009	movie	7.8
9	Echt fett - Best of	2005	TV show	8.0
10	Big Bang Theory	2008	TV show	8.5
11	Big Bang Theory (engl.)	2008	TV show	8.5
12	How I Met your Mother	2005	TV show	8.5
13	How I Met your Mother (engl.)	2005	TV show	8.5
14	Bodo Wartke - Achillesverse	2007	music	-
15	U2 360	2010	music	8.2
16	David Garrett live	2010	music	-
17	Amy Winehouse - I Told You I Was Trouble	2007	music	7.9
18	Ice age 2	2006	animation	6.9
19	The Simpsons	2002	animation	8.9
20	Ratatouille	2007	animation	8.0
21	Family Guy	2010	animation	8.3
22	Dschungelwelten	2006	documentary	-
23	Eiswelten	2006	documentary	6.9
24	Food Inc.	2008	documentary	7.9
25	Zukunft Ohne Menschen	2008	documentary	7.7

Table 3.1: Study *WTP 1*: List of available content

to related work (e.g. see [140]), four different video quality levels were offered, i.e., the video streams were impaired by network packet loss rates between 1% (worst quality) and 0% (best quality) in order to span a reasonable quality range. The packet loss rate was determined by the network emulator *netem*[5], which generated a random packet loss distribution for the UDP connection from the server to the client with the given drop rate in percent. No forward error correction or other error concealment was used.

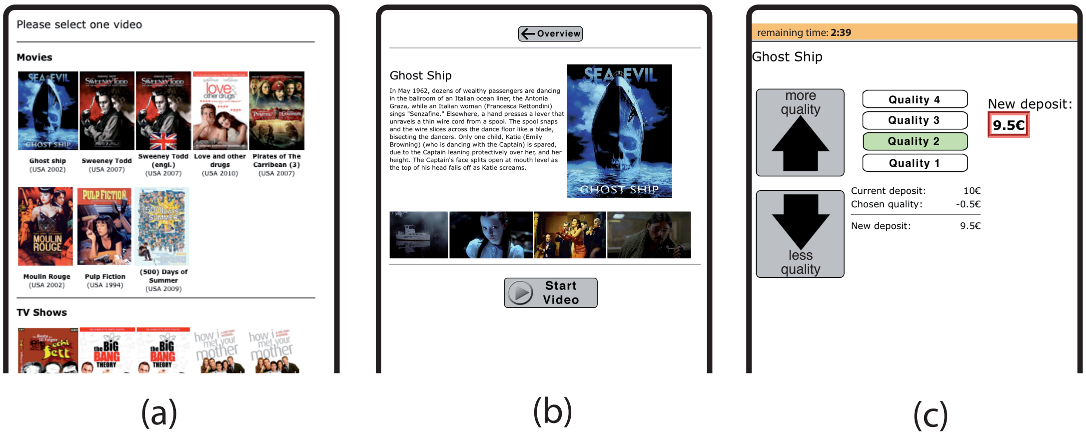


Figure 3.4: Study WTP 1: Tablet interfaces for (a) video library overview, (b) detailed video information and (c) video quality selection during playback

Video content: To guarantee appropriate and appealing video material, H.264 encoded video files from original DVDs were extracted with a constant video bitrate of 5200 Kbit/s and a resolution of 720 x 576 pixels (SD video resolution). In order to provide content of the test participant's interest a selection of 25 videos, separated in five content classes, were offered: movies, TV series, documentary reports, animation and music concerts, see Table 3.1. Videos with a length of 20 min were used, as this duration is well aligned to typical TV series and also allows the consumption of longer coherent narrative sequences from longer videos, thereby fostering the users' immersion in the content. Although it is not common in video quality trials to have videos with this durations, such long clips are needed to guarantee an appropriate and quite realistic VoD scenario, see also [141].

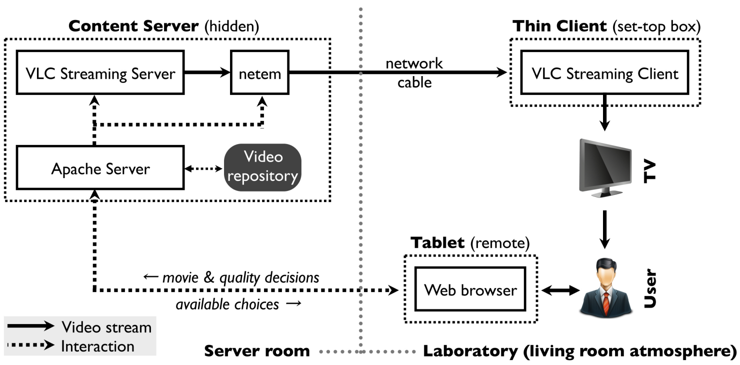
	quality class 1	**quality class 2**	**quality class 3**	**quality class 4**
packet loss [%]	1	0.25	0.085	0
delay [ms]	75	75	75	75
costs per minute [€]	0	0.025	0.05	0.075
costs per movie [€]	0	0.5	1	1.5

Table 3.2: Study *WTP 1*: Available quality classes

Complementary video QoE study: As described above, the test participants had to evaluate the perceived video quality via questionnaires after each consumed video. Most likely, these ratings were biased by the economic decision regarding the spent money

[5] http://www.linuxfoundation.org/collaborate/workgroups/networking/netem, last access: 14th August 2015

for enhanced video quality. Hence, an additional video quality study was conducted with *different* test participants but with identical technical parameters and without any user interaction regarding content, quality or spent money. Due to the applied between-subjects design approach, different users participated in the study, i.e., the users who evaluated the video quality impaired by individual decisions were *not* identical with the users who participated in the complementary study without any user decisions. In the complementary video QoE study, each user evaluated two videos and each video was presented with four different packet loss impairments. Therefore, each participant evaluated eight videos regarding perceived video quality and acceptance. Due to the lack of individual purchases, the duration of each presented video was reduced to 2 minutes (instead of 20 minutes). The authors of [142] demonstrated that differences regarding duration in common video quality assessment studies have a negligible impact on quality perception, especially for longer videos with a duration of several minutes, which legitimizes the setup of the complementary video QoE study.



(a) Technical Setup

(b) FTW's iLab

Figure 3.5: Study *WTP 1*: Technical setup and iLab

3.1.2.2 Results

The study *WTP 1* was conducted in Vienna in the FTW's iLab in October 2011, see Figure 3.5(b)[6]. After cleaning the data basis, 129 videos selected by 43 users (22 male and 21 female, mean age: 36.8 years) were analyzed. Only 5 users have had experiences with VoD platforms (2x iTunes, 2x A1 Video- store, 1x UPC on demand). Each month, these users spent between 2.5 € and 9.9 € on such services resulting in a mean of 5.48 €.

According to Figure 3.6(a), approximately 20% of the participants spent the maximum amount of money to increase the quality, i.e., the remaining deposit was 5.50 €. In contrast to this, only 10% of the users never enhanced the quality and received therefore the maximum amount of 10 € at the end. The majority, however, took an intermediary position between these two extremes. Figure 3.6(b) depicts the cumulative distribution function of the money spent per video. On average 1.01 € (standard deviation=0.49, median=1.03) were spent per movie to decrease the packet loss of the transmission. Obviously, the spent money was subject to the presented quality. Accordingly, the mean chosen quality level per movie was 3.04 (standard deviation=0.99, median=3.25s), i.e., on average the second best quality level was chosen. Overall, quality class 1 was selected 18 times, quality class 2 was selected 20 times, quality class 3 was selected 52 times and quality class 4 was selected 36 times.

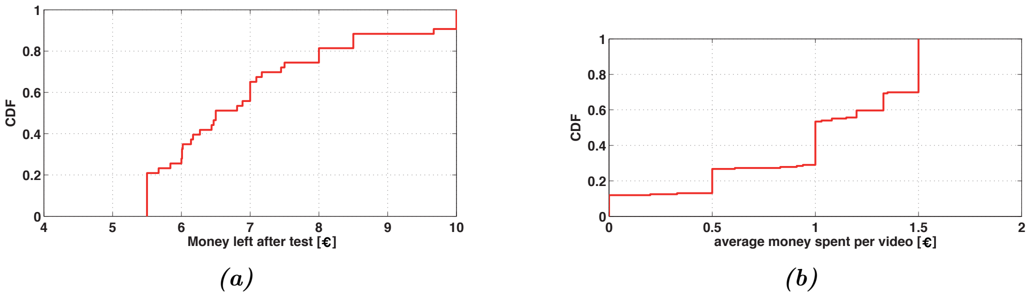


Figure 3.6: Cumulative distribution function plots of paid out money after the test (a) and of money spent per movie (b).

Most of the users were satisfied with the selected videos and the range of available videos: After each video, the users were asked via the statement: "The selected video was interesting for me" with answering options ranging from 1 (="completely agree") to 5 (="completely not agree"), and at the end of the experiment the users were asked: "The range of the available videos was large enough" with the answering options ranging from 1 (="completely agree") to 5 (="completely not agree"). The average answering value of the first question was 1.3 (standard deviation=0.46) and the average answering value for the second question was 2.0 (standard deviation=0.9). Hence, providing 25 videos separated in 5 categories led to satisfied experiment participants regarding video content.

[6] http://www.ftw.at/portfolio/i-lab, last access: 14th August 2015

According to Table 3.3 most of the money was spent at the beginning of the movies — QSP 1 as depicted in Figure 3.3 — whereas during the rest of the video duration only limited quality modifications were applied. Not even continuously decreasing prices — as calculated on the basis of the remaining share of the movie — triggered further quality upgrades. Generally, most users chose the best or the second best quality in the first QSP without extensively testing the available options. The mean quality try out rate was 4.18 in the first QSP, i.e., at the beginning of the video every user changed the quality only four times on average. Table 3.3 also depicts the decrease of interactions regarding the following QSPs 2-4. The average amount of money spent after QSP 1 was very low, due to the negligible quality enhancements after the first quality decision.

	QSP 1	**QSP 2**	**QSP 3**	**QSP 4**
avg. quality try-outs	4.18	1.00	0.71	0.63
avg. spent money	0.926	0.066	0.016	0.003
avg. chosen quality class	2.85	3.05	3.12	3.16

Table 3.3: Study *WTP 1*: User behavior during the four QSPs

The whole test duration was approx. one hour, which included browsing the video content, selecting and watching three videos labeled as iteration 1, 2 and 3. The experimental setup clearly proved that there was no behavioral difference between the beginning and the end of the test: the average amount of money spent on increasing the quality remained constant over the whole test duration, respectively over the three consumed videos, which was also true for the average try-out rate (average try-out rate overall was 6.5) and the average resulting video quality ratings. Therefore, there was no influence of the experiment duration or the amount of selected videos on the results, see Figure 3.7.

The observed behavior of the participants provided a revealing insight in the chosen purchasing and consumption strategies. Similar to the categorization of [111] discussed in Section 3.1.1.2, four types of buyers were identified:

- Strategic buyers: Some users repeatedly chose the best quality during the QSPs, which was free of charge. However before the end of each QSP, they returned to the poorest quality level again, i.e., no quality purchases were made in order to receive the maximum deposit payout.
- Generous buyers: Users were observed who always selected the best quality at the beginning of the movie without testing lower qualities. Thereafter, they stated their insensitivity to payments from their deposit and their intention to watch all movies in the best quality available.
- Budget-minded buyers: Some users declared after the experiment that they did not care about the quality as long as they received the full 10 € as payout. In some cases, even better qualities were not tested for free.
- Quality & price-aware buyers: Most of the users tested all quality classes at the beginning and finally chose quality class 3 or 4.

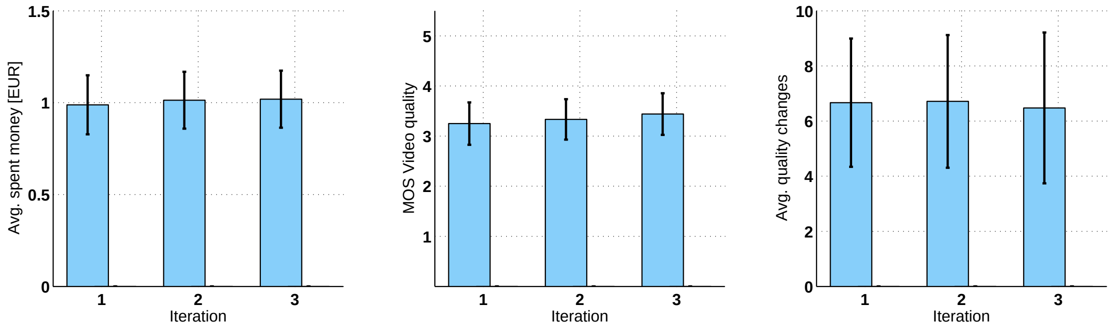


Figure 3.7: Study WTP 1: Average money spent, video quality ratings and average quality changes separated by iteration (95% confidence intervals)

Figure 3.8 depicts which movie was selected how often (corresponding movies for the displayed IDs can be found in Table 3.1). For example, the video from the TV-Show "Echt Fett" (ID=9) was selected most often, whereas the video "Bodo Wartke - Achillesverse" was never selected. The red numbers in Figure 3.8 represent the percentage amount of selections per category, i.e., TV shows were selected most often (29%) followed by documentaries (24%) and movies (23%). In addition, Figure 3.10 shows that the video category neither influences the average money spent, nor the resulting video quality ratings or the average quality changes.

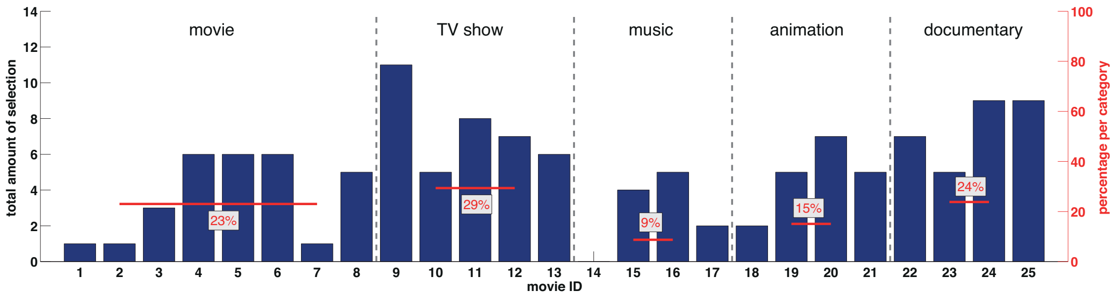


Figure 3.8: Study *WTP 1*: Selected movies and percentage of category

36 users (19 male and 17 female, mean age: 39.2 years) participated in the complementary Video QoE study — which took place in January 2012 at FTW's iLab — which did not include any user decisions regarding content, quality or spent money. Figure 3.9 depicts the video quality and acceptance rating results for both studies: For all four quality levels there is a significant difference for the four MOS pairs, also regarding the Wilcoxon rank sum test[7]. One could assume that the acceptance rate for ratings including user decisions should be close to 100% for all quality classes, because the participants selected the quality themselves. However, the question regarding acceptance focused on the usage *at home*, i.e, a certain quality might be accepted in the context of a laboratory QoE study but would be rejected if consumed at home in the living room.

[7] Quality level 1: p=0.0001164; quality level 2: p=0.0021, quality level 3: p=0.00000376, quality level 4: p=0.0072; Comparisons of ratings with user decisions vs. ratings without any user decisions.

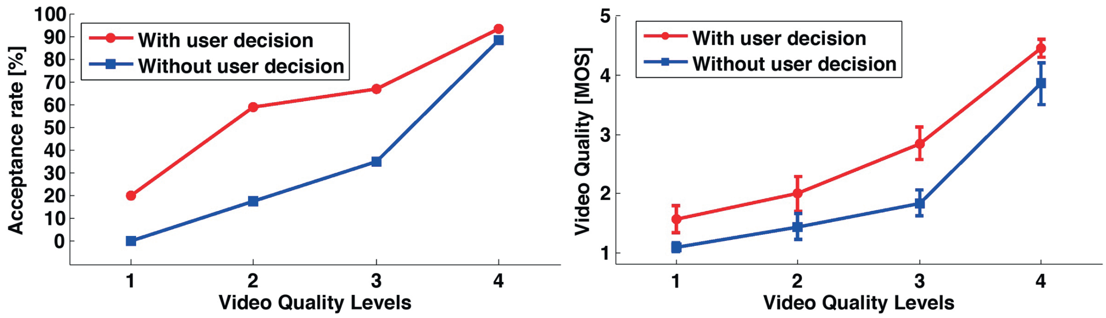


Figure 3.9: *WTP 1*: MOS video quality and acceptance ratings (90% confidence intervals)

Compared to the MOS values in Figure 3.9, the differences regarding the acceptance rate diverge even more clearly, raising the question which factors may have caused such "irrationalities" regarding quality assessment. These unexpected differences regarding subjective quality evaluation might be explained by the socio-psychological theory of cognitive dissonance, which will be discussed in the following section.

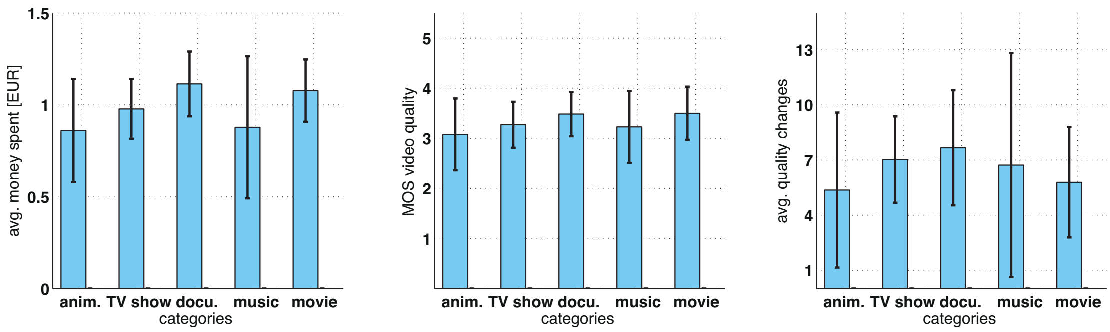


Figure 3.10: *WTP 1*: Average money spent, MOS video quality ratings and average quality changes separated by category (95% confidence intervals)

3.1.2.3 Parenthesis about Cognitive Dissonance

The term cognitive dissonance was introduced by Festinger [143] as a social-psychological approach which addresses the human aim to achieve a harmonious and consistent state of mind. Cognitive dissonance occurs if two opposed cognitive elements exist simultaneously, which causes stress and discomfort and which generates a strong demand to minimize this dissonance [144]. For example, a person who has decided to smoke is normally conscious of the risks of smoking. These two elements (the habit of smoking and the knowledge of the dangerous risks) are obviously incompatible: a cognitive dissonance occurs. Hence, there are two options: First, the person changes contrary behavior, i.e., she stops smoking, or secondly, additional information is added to reduce the dissonance. For example, a person who smokes could seek purposefully information which underlines the benefits of smoking or trivializes the risks, see [145]. This effect of *whitewashing* is also visible during/after high involvement decisions. For example, the authors of [146] show that citizens who voted for a political party, which was at the end not successful relativize their vote decision with the reason ”Maybe it is better that the other party is in charge now, this strengthens our political system”.

Since cognitive dissonance was described as a theoretical framework, many mundane behaviors can be explained in a profound way, e.g., marketing effects of advertising and sales techniques. Most relevant for the QoE-related findings is the so-called *post purchase cognitive dissonance*, which can be observed after a purchasing decision was made. For example, buying a car normally leads to high financial expenses. Hence, considering only the monetary aspect the made choice would be unfavorable. This leads to a cognitive dissonance (”I bought an expensive car”, but ”generally I want to save money”) which forces the concerned person to change something. Selective information seeking could be used to reduce this dissonance: before the purchase the potential customer possibly tries to get various information about different cars for a reasonable decision. After the decision and the purchasing, however, the customer tries to find only information which *justifies* his decision (this is even not really necessary because no information is needed as the decisions has already been made). For example, advertising about the own car (and not rejected alternatives) is extensively consumed or test reports are read which emphasize the advantages and neglect disadvantages of the own car. By using this approach the dissonance can be reduced (”Now I have a safer car with a higher resale value”; ”I drive a more environment-friendly car”, etc.). Hence, marketing plans need to consider this post-purchase behavior to prepare adequate communication strategies, see also [144].

To bring it back to the empirical QoE-results, let us focus on the questionnaire which was presented after each consumed video. Answering the binary acceptance question ”Would you consume this video in the presented quality at home?” *could* lead to cognitive dissonance: Choosing a bad video quality and answering honestly with ”no” would lead to the contradictory state ”I have chosen bad quality” and ”I would not choose this quality”. The only way to avoid this post purchase cognitive dissonance is to select ”yes” and to justify the previous decision (”It was reasonable to choose the bad quality to avoid costs, I would accept it at home too”). This also applies to the video quality MOS ratings, i.e., there is a clear tendency for less critical ratings if active user decisions are forced. One could interpose that it was primarily the methodological setup that caused the cognitive dissonance rather than the presented quality itself. For example, the question regarding acceptance might be misleading for quality estimation, so the this question would cause invalid results. This would mean that active user decisions have no influence on quality perception and only the setup generates cognitive dissonance. On the other hand, however, the assessment of the video quality does not force justifications in the same way as the acceptance question and also leads to diverging results. This points out that the observed differences are in all probability caused by user behavior and are not raised by the questionary itself. Explicitly, the observed economic cognitive dissonance effect may be compared with the observations discussed in [147], where employees working with dangerous chemicals were asked about their working conditions.

In both cases *negative cognitions* (poor video quality vs. dangerous working conditions, respectively) were sugarcoated by individuals (consumers vs. workers respectively) in order to defend their choice of quality purchases or employment. Although the effect of dangerous working conditions may be more severe than disutility in entertainment services, the underlying effect of cognitive dissonance seems to be on the same economic dimension.

3.1.2.4 Conclusions

This first WTP/QoE user study, which combines economical user decisions and subjective quality evaluation, demonstrates that individual economic decisions impact the quality perception process. According to the presented findings the ability to individually select certain quality classes — which are paid with the user's own money — *positively* impacts the subjective quality assessment, i.e., for all quality levels the video quality ratings were significantly more positive. One possible explanation for this outcome might be the well-known cognitive dissonance effect. Also, it has been shown that the described laboratory setting led to reasonable results, i.e., 90% of the test users spent *real* money. Moreover, the range of the available movies, the available categories and the length of the movies were reasonable.

Nevertheless, during the execution of the study *WTP 1* and during the analysis phase, several lessons were learned and also shortcomings were identified, which were considered in the following study *WTP 2*:

- Instead of using a between-subjects design a within-subjects design should be applied to get more valid results.
- According to the observed user behavior it is not necessary to provide several quality selection phases. Hence, it is better to provide only a single quality selection phase at the beginning with the possibility to record the user interaction behavior in a more detailed way, e.g., direction of quality changes with accurate time stamps.
- The network protocol UDP was used for video streaming impaired by packet loss to generate different video quality classes. The deployed network emulator generated randomly distributed packet loss, i.e., it is not guaranteed that all experiment participants perceived the same degraded videos in identical technical video quality classes. Hence, for following experiments it is recommended to use video impairments which are repeatable.
- To be future-proof, instead of SD videos HD videos should be evaluated in subsequent studies.
- It is assumed that cognitive dissonance might be the reason between the difference of the quality ratings, see Figure 3.9. This assumption should be checked via dedicated cognitive dissonance questionnaires.

- According to the stated findings, a test duration of one hour including the selection of 3 videos, each with a duration of 20 minutes, does not affect the user in a negative way, i.e., the average quality selection rate, the average amount of spent money and the MOS video quality values remain constant for all three iterations.
- 25 videos in 5 categories were provided (Animation, Action, Music, Documentary, TV Show). There are no significant differences between the spent money, the amount of quality changes and the video quality ratings regarding the categories, see Figure 3.10. Also, the selectable amount of videos were approved by the experiment participants.

3.1.3 WTP 2 User Study

The results of the study *WTP 1* provide some interesting insights into the relationship between pricing and QoE, but also additional questions occurred and some shortcomings of the 2011 setup needed to be fixed.

Regarding the technical setup, several adaptions were made to improve the experiment. First of all, a within-subjects test design was implemented, i.e., each user had to make video quality assessments which were impaired by economic decisions and assessments without any economic decision interferences. Additionally, charged quality changes were only possible at the beginning of the video consumption (in contrast to the *WTP 1* user study in which several quality selection phases were implemented). In Section 3.1.2.3 the hypothesis about cognitive dissonance was proposed to explain the rating differences. To verify this assumption, in the study *WTP 2* a specific questionnaire was used.

Additionally, the range of selectable quality classes was expanded from 4 to 20 quality classes to observe a more fine-grained selection & rating behavior. Furthermore, three different price plans A, B and C were used, i.e., in contrast to the previous study *WTP 1*, the quality classes were differently priced at specific stages during the experiment. As depicted in Table 3.9, each available quality class *Q0 - Q19* was charged differently depending on the set price plan. The lowest quality class Q0 was always for free, whereas the higher classes got linearly more expensive up to the highest quality class Q19, which is charged with 2 - 4€.

To extend the possibilities of applying different price plans during the experiment, the influence of price plan patterns was also evaluated: overall, each user had to make three economic quality decisions and for each decision one of the three previously mentioned price plans was applied, e.g., the user with the ID 18 had to make her first economic quality decision with price plan C, her second quality decision with price plan B and her last quality selection with price plan A. Therefore, several *Price Plan Patterns* were set for the users. Table 3.4 depicts the three patterns which were used in the experiment. Hence, it is possible to investigate the influence of *increasing prices* (pattern I), the influence of *decreasing prices* (pattern II) and the influence of *constant prices* (pattern III) on subjective quality assessment.

To get more insights into quality assessment behavior, the quality classes Q16 - Q19 were implemented differently compared to the other quality classes: whereas the video bitrate remained *constant* at the highest available value, the prices *increased* up to the maximum fee, see Table 3.4. Therefore, economically spoken it was not rational to choose a higher quality class than Q16.

3.1.3.1 Setup

In general, the study setup of the study *WTP 2* was similar to the study *WTP 1*. Of course, many adaptions and improvements were applied.

quality class	**Q0**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**
VBR [kBit/s]	128	181	256	362	512	724	1024	1448	2048	2896
priceplan A [€]	0	0.105	0.211	0.316	0.421	0.526	0.632	0.737	0.842	0.947
priceplan B [€]	0	0.158	0.316	0.474	0.632	0.789	0.947	1.105	1.263	1.421
priceplan C [€]	0	0.211	0.421	0.632	0.842	1.053	1.263	1.474	1.684	1.895
quality class	**Q10**	**Q11**	**Q12**	**Q13**	**Q14**	**Q15**	**Q16**	**Q17**	**Q18**	**Q19**
VBR [kBit/s]	4096	5793	8192	11585	16384	23170	32768	32768	32768	32768
priceplan A [€]	1.053	1.158	1.263	1.368	1.474	1.579	1.684	1.789	1.895	2
priceplan B [€]	1.579	1.737	1.895	2.053	2.211	2.368	2.526	2.684	2.842	3
priceplan C [€]	2.105	2.316	2.526	2.737	2.947	3.158	3.368	3.579	3.789	4

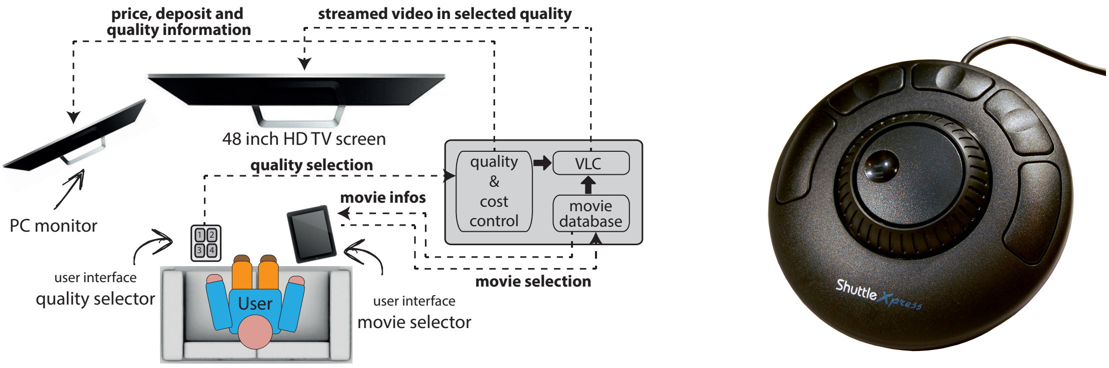
Table 3.4: Study *WTP 2*: Quality classes with video bitrate and fees

Procedure: In contrast to the previous user study *WTP 1*, in the study *WTP 2* a "pure" video quality QoE evaluation phase was executed before the user was able to select the video content and subsequently was able to adapt the quality. That means that, each user had to evaluate 17 short video clips (1080p, h.264 codec, duration of 10 seconds) encoded with 17 video bitrates ranging from 128 kBit/s to 32768 kBit/s, see quality classes Q0 to Q16 in Table 3.4. Each video bitrate was evaluated once per user. Based on [30], the participants used a standard, continuous 5-point ACR-scale ranging from "excellent" to "bad" to evaluate the presented videos and with the binary yes/no question "Would you consume this video in the presented quality at home?" the so called acceptance was determined. This video assessment phase was used to investigate the subjective quality perception without any user decisions, i.e., the participants did not make any decision regarding the consumed content, quality or expenditure. These measurements are referred to "rating 1", see Table 3.5. After that, the WTP/QoE part of the study was conducted: The users received 10 € as a deposit which could be used to increase the video quality of three individually selected videos. The users were able to try out all quality classes for free during the first 5 minutes of each 20 minute video. Thereafter, the last selection was irreversibly set and the related fee was withdrawn from the deposit. After the video had ended, the participant evaluated the perceived video quality in the same way as described for rating 1. This measurement is described as rating 2, see Table 3.5. At the end of the complete test, i.e., after three movies had been selected, consumed and evaluated, the user had to evaluate the perceived video quality of the previously selected three videos with the selected quality again, but of course with no additional payment and with a video duration of 10 seconds. This evaluation is labeled as rating 3, see Table 3.5.

	position	content selected?	monetary / quality decision?	video length
Rating 1	begin of study	no	no	10 sec.
Rating 2	after video consumption	yes	yes	20 min.
Rating 3	end of study	yes	no	10 sec.

Table 3.5: Study *WTP 2*: QoE measurements

Economic aspects: Each user received 10 € in advance as deposit to improve the video quality during video consumption of the selected videos. At the end of the test the remaining deposit was paid out in cash to the user. The lowest available quality was always for free and for the highest quality the user had to pay between 2 € and 4 € per selected movie, depending on the applied price plan, see Table 3.4. Overall, each user selected three videos (iteration 1,2 and 3) and each participant was randomly assigned to a certain price plan pattern, see Table 3.6. For example, pattern II means that price plan C was applied to the first selected movie (iteration 1), price plan B was applied to the second movie (iteration 2) and price plan C was applied to the third selected movie (iteration 3). Overall, 14 users were assigned to pattern I, 15 users to pattern II and 14 users to pattern III.



(a) Technical Setup

(b) Jogwheel as quality selection interface

Figure 3.11: Study *WTP 2*: Technical setup and jogwheel

Technical setup: A modified, adaptive streaming system was used based on Apple's HLS (HTTP Live Streaming) to ensure the possibility of adapting the video quality during consumption. The open source video player VLC 2.1 was modified in order to reduce the switching time between video streams from 7 seconds to 1 second. Via an iPad-Interface, the users were able to choose their preferred videos from a set of 20 action movies, each with a duration of 20 minutes. The same iPad-Interface was used for video content browsing and video selections as in the study *WTP 1*, see Figure 3.4(a,b). After a movie had been chosen, the test user was able to try out all video qualities with a so-called jogwheel [8], i.e., turning the wheel clockwise increased the video quality and vice versa, see Figure 3.11(b). As can be seen in Table 3.4, the user was able to switch between 20 quality classes. The current deposit and the current charge, e.g. 0.211 € for quality class Q2 in price plan A, were displayed via a small screen located near the TV set. The name of the effective quality class was *not* displayed, i.e., only monetary aspects were directly visible. Therefore, the user was forced to make a decision based on current *charge* and *perceived quality* instead of considering *objective, QoS-related* aspects like the displayed name of the currently applied quality class. Missing quality labels also led to the effect that the participants were not able to remember and automatically apply the previously set quality and therefore the actual decision was based on current quality perception and charges.

pattern	description	iteration 1	iteration 2	iteration 3
I	*increasing prices*	A	B	C
II	*decreasing princes*	C	B	A
III	*constant prices*	B	B	(C or A)

Table 3.6: Study *WTP 2*: Price plan patterns

Video content: In contrast to the previous study *WTP 1* — in which SD video content was used — high definition material (1080p, 25fps) was applied for the study WTP 2. Whereas there are lots of genres available on DVD (SD content) the range of categories available on Blu-Ray (HD content) was rather limited at the time when the study was prepared. For example, there were no appropriate TV shows available on Blu-Ray, only DVDs were obtainable. Therefore, instead of providing 5 different genres like in 2011 the study *WTP 2* focused on action movies. Additionally, the usage of only a single content class avoided the problem of the impact of content classes on QoE, e.g., a soccer sequence normally gets worse subjective ratings compared to a technically equally encoded animation sequence, see for example [148]. Several Blu-Rays were ripped with the open source tool handbreak[9] and HLS-compatible videos were generated with the HTTP Live Streaming Tools from Apple[10]. Table 3.7 provides an overview about the used movies.

Cognitive Dissonance Questionnaire:

There are several existing cognitive dissonance questionnaires. For example, the one presented in [149] was used in a study discussed in [150] for evaluating the post purchasing behavior of people who had bought premium pieces of furniture. Obviously, this approach is not appropriate for low-priced decisions like technical video quality. The authors of [151] developed the so-called "Preference For Consistency (PFC) Scale" which focuses strongly on disappointments, e.g., how a wrong individual decision affects others. Finally, an adapted cognitive dissonance questionary based on the germanophone

[8] http://ergo.contour-design.com/ergonomic-mouse/shuttlexpress, last accessed: July 10th, 2015
[9] https://handbrake.fr/, last accessed: July 10th, 2015
[10] https://developer.apple.com/library/ios/documentation/NetworkingInternet/Conceptual/StreamingMediaGuide/UsingHTTPLiveStreaming/UsingHTTPLiveStreaming.html, last accessed: July 10th, 2015

ID	title	year	IMDb rating
1	Tower Heist	2011	6.2
2	Black Gold	2011	6.7
3	Black Swan	2010	8.0
4	Cowboys & Aliens	2011	6.0
5	James Bond: Quantum of Solace	2008	6.7
6	Harry Potter and the Half-Blood Prince	2009	7.5
7	Harry Potter and the Deathly Hallows: Part 2	2011	8.1
8	In Time	2011	6.7
9	Knight and Day	2010	6.3
10	The Sorcerer and the White Snake	2011	5.9
11	Pirates of the Caribbean: On Stranger Tides	2011	6.7
12	SALT	2010	6.4
13	Sherlock Holmes	2009	7.6
14	The Fall	2006	7.9
15	Transformers: Dark of the Moon	2011	6.3
16	Wickie auf großer Fahrt	2011	5.5
17	Harry Potter and the Deathly Hallows: Part 2 (English)	2011	8.1
18	James Bond: Quantum of Solace (English)	2008	6.7
19	Pirates of the Caribbean: On Stranger Tides (English)	2011	6.7
20	In Time (English)	2011	6.7

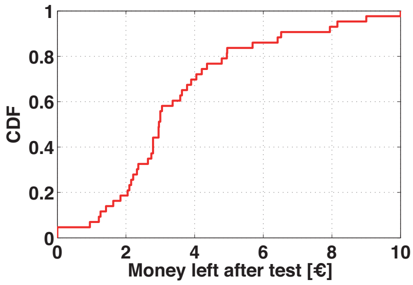
Table 3.7: Study *WTP 2*: List of available content

Münchener Dissonanzskala (MDS-K) was used in the study *WTP 2*. The creator of the scale Christoph Piesbergen [152] kindly provided the permission to use and to adapt the questionnaire. The questionnaire was handed out once per participant after the video quality evaluation of the first selected video was, see also Appendix A.

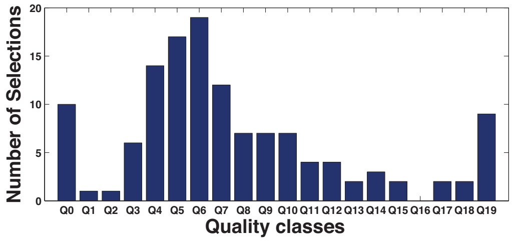
3.1.3.2 Results

Overall, 43 test users from Austria participated in the study (31 female, 12 male users). 26 of them were between 18 and 30 years old, 10 were between 31 and 45 years old and 7 were older than 45 years. Most of them (16) were employed or students (16), only 8 participants were in a relationship or married and only 3 users were experienced in charged VoD services.

Only two users did not spend any amount of their deposit to increase the presented video quality, i.e., a large amount of the participants actually spent money and therefore interacted with the system and made decisions regarding enhancement, see Figure 3.12(a). This result shows that the adaptions of the setup — change from SD to HD, shrinking the amount of content classes from five to one — were reasonable regarding the spent amount of money. Hence, in the study WTP 2 the choice regarding deposit, charging, available content and video duration was suitable for the test purpose.



(a) Paid out money at the end of the test. Only 4.6% of our users did not spend any money on quality enhancement and therefore received the maximum paid out deposit of 10 €

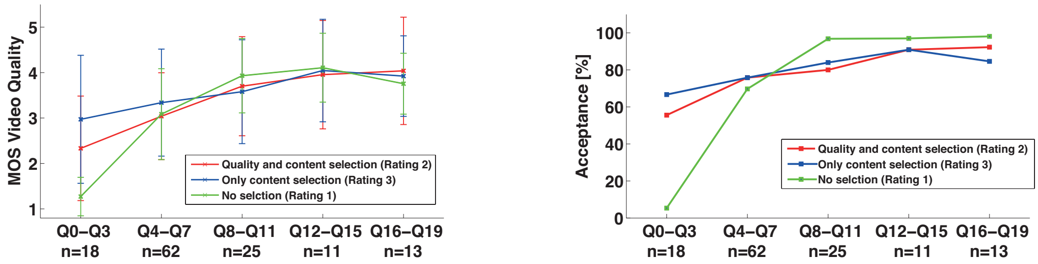


(b) Number of selections for each quality class for all users and all iterations

Figure 3.12: Study *WTP 2*: Money left and quality class selections.

Figure 3.12(b) shows that some quality levels were more popular than others. There is some kind of normal distribution around quality class Q6 (which is obviously a good trade-off between technical quality and price). Interestingly, the most expensive quality class Q19 and the free of charge quality class Q0 were remarkably often selected. In general, this unbalanced selection leads to some methodological problems. For example, quality class Q5 was selected 19 times, whereas quality class Q17 was chosen only twice. Thus, there may not be meaningful results for each individual quality class as some only have a sample size of two. Therefore, for methodological reasons the 20 quality classes were aggregated. The grouping of 4 quality classes into one class — resulting in 5 classes (Q0-Q3, Q4-Q7...etc.) — led to useful results by increasing the sample sizes. On the one hand some granularity is lost regarding individual quality classes, but on the other hand a more robust analysis can be conducted.

Figure 3.13(a) depicts the results regarding the video quality MOS values for all three quality ratings (rating 1,2 and 3, see Table 3.5). The extremely large overlapping confidence intervals and the similar mean values make it difficult to draw any conclusions. It seems that for the lower quality classes Q0 to Q4 the MOS values are low if no content/quality decision was made and a little bit higher if a user decision was involved (rating 2 and 3). Figure 3.13(b) shows that the acceptance rates for quality classes greater than or equal to Q4 are rather similar for all three types of ratings. Only for low quality classes the acceptance rate for ratings excluding any user decision is unambiguously lower compared to ratings which included a user decision. Except for the acceptance ratings for Q0-Q3, these findings are in contrast to the findings of the study *WTP 1*: No significant positive impact of individual economic decisions occurs. Of course, the limited number of ratings per category and the consequent aggregation may distort the findings.



(a) Original 20 quality classes were aggregated to 5 classes; MOS video quality ratings: 5=best and 1=worst, confidence intervals are given for a level of 95%, ratings 1,2 and 3 refer to Table 3.5

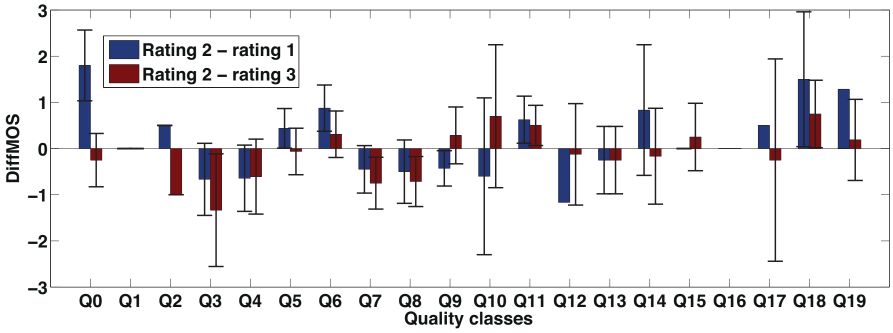
(b) Original 20 quality classes were aggregated to 5 classes, acceptance is the percentage amount of users who positively answered the question "Would consume this video in this quality at home?", ratings 1,2 and 3 refer to Table 3.5

Figure 3.13: Study *WTP 2*: MOS and acceptance results for aggregated ratings.

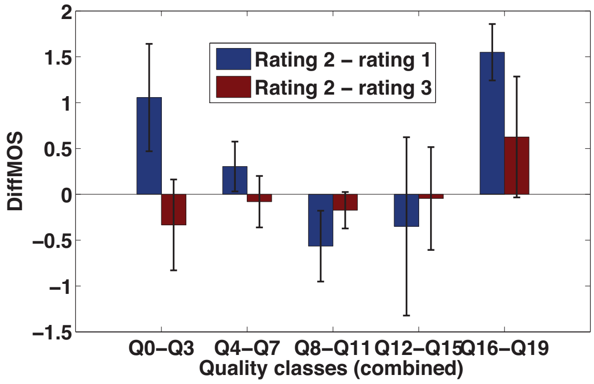
		$DiffMOS$	median	standard deviation	confidence interval	$p_{ranksum}$
Rating 2 - Rating 1	Q0-Q3	1.0556	0.5000	1.4337	**[0.4696, 1.6416]**	**0.001**
	Q4-Q7	0.3036	0.5000	1.2161	**[0.0318, 0.5754]**	0.0913
	Q8-Q11	-0.5652	0	1.0798	**[-0.9511, -0.1793]**	0.2540
	Q12-Q15	-0.3500	-0.5000	1.7958	[-1.3224, 0.6224]	0.1919
	Q16-Q19	1.5500	1.5000	0.5986	**[1.2420, 1.8580]**	**0.0024**
Rating 2 - Rating 3	Q0-Q3	-0.3333	-0.2500	1.2127	[-0.8289, 0.1623]	0.4804
	Q4-Q7	-0.0804	-0.5000	1.2569	[-0.3613, 0.2005]	0.5091
	Q8-Q11	-0.1739	0	0.5561	[-0.3726, 0.0248]	0.7631
	Q12-Q15	-0.0455	0	1.0357	[-0.6063, 0.5153]	1
	Q16-Q19	0.6250	0.7500		[-0.0343, 1.2843]	0.0768

Table 3.8: Study *WTP 2*: Ratings Overview

In Figure 3.14(a) the $DiffMOS$ values separated by all 20 quality classes are depicted. The low amount of ratings per quality classes (see Figure 3.12(b)) and the following large confidence intervals make it hard to draw any conclusions. Hence, also for the $DiffMOS$ results quality classes were combined to increase the sample number for each $DiffMOS$ value and to reduce the corresponding confidence interval. Figure 3.14(b) depicts the $DiffMOS$ results for the combined quality classes. Except for the $DiffMOS$ ratings based on rating 2 and rating 1 (blue bar) for the lowest and the highest quality classes, no significant differences regarding the influence of individual, economic user decisions on quality assessment are visible. So far, the results depicted in Figures 3.13 and 3.14 do not validate the findings from the *WTP 1* user study, which suggested that an individual, economic user decision generally leads to a better quality assessment. Although the sample sizes for each quality are low, which makes drawing conclusions rather difficult, the results for the $DiffMOS$ values for the higher quality classes Q16 to Q19 in Figure 3.14(a) indicate that constant quality is evaluated better when prices get higher. This is in line with empirical user studies dealing with the impact of prices on product quality perception, see [112], which is discussed in Section 3.1.1.



(a) $DiffMOS$ for *WTP 2* user study



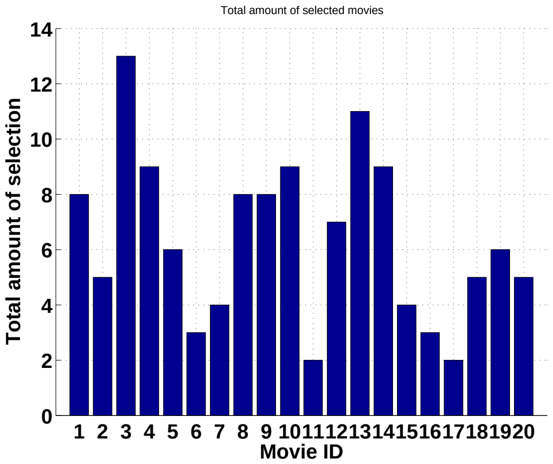
(b) $DiffMOS$ for *WTP 2* user study (combined quality classes)

Figure 3.14: Study *WTP 2*: $DiffMOS$ for all quality classes and for combined quality classes (90% confidence intervals)

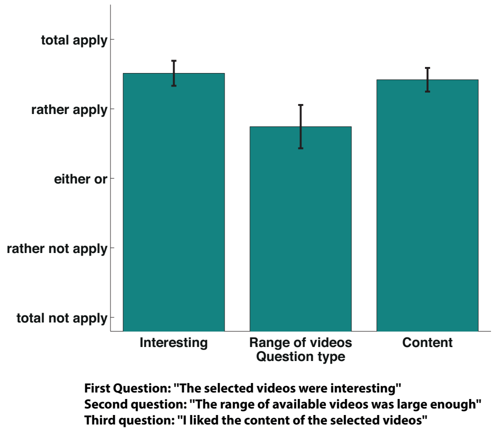
Nevertheless, the $p_{ranksum}$ values for combined quality classes depicted in Table 3.8 show that individual, economic decisions concerning free/cheap quality classes (Q0-Q3) influence quality assessment significantly, as well as the highest/most expensive quality classes (Q16-Q19). Figure 3.16(a) additionally depicts the CDF-plot for $DiffRatings$ corresponding to the difference between rating 2 and rating 1 (see Table 3.5). Especially for the highest quality classes Q16 and Q19 the findings are all positive, which indicates a strong positive effect of economic decision on quality assessment.

Figure 3.15(b) depicts how satisfied the users were with the available content. In general, most of the users were satisfied with the range of the available videos and with the individually selected content.

As depicted in Table 3.4 various price plans were applied, e.g., the user with the ID 17 was assigned to price plan A for his first movie selection (Iteration 1) and to price plan B for his second movie selection (Iteration 2). Figure 3.17 (left) shows small changes in both average selected quality (red line) and average spent money (blue line) for different charging: Even if prices increased on average the participants did not significantly reduce

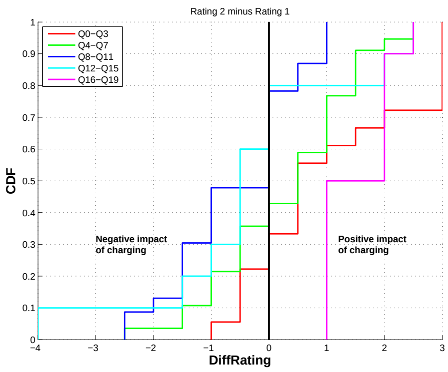


(a) Selected movies

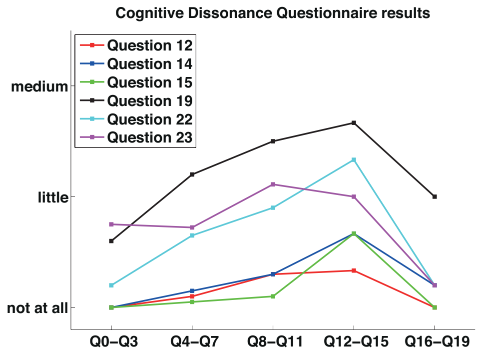


(b) Content satisfaction (90% confidence interval)

Figure 3.15: Study *WTP 2*: Movie selection and content satisfaction



(a) CDf plot for *DiffRatings* (rating 2 - rating 1; combined quality classes)



(b) Cognitive dissonance questionnaire results (related questions can be found in Annex A).

Figure 3.16: Study *WTP 2*: CDF-plots and cognitive dissonance results

the selected quality (red line), which led to an increase on average money spent (blue line) when more expensive price plans were applied. At least under the circumstances of the laboratory experiment there was a tendency that users prefered constant quality to constant payments on average. The MOS ratings in Figure 3.17 (right, red line) are in line with the average selected quality ratings in Figure 3.17 (left, red line), i.e., different price plans did not influence the subjective perceived quality.

As depicted in Table 3.6, various combinations of the price plans A, B and C were applied. In Figure 3.18 the influence of the patterns on the average selected quality, the average money spent, the video quality MOS ratings and acceptance ratings are

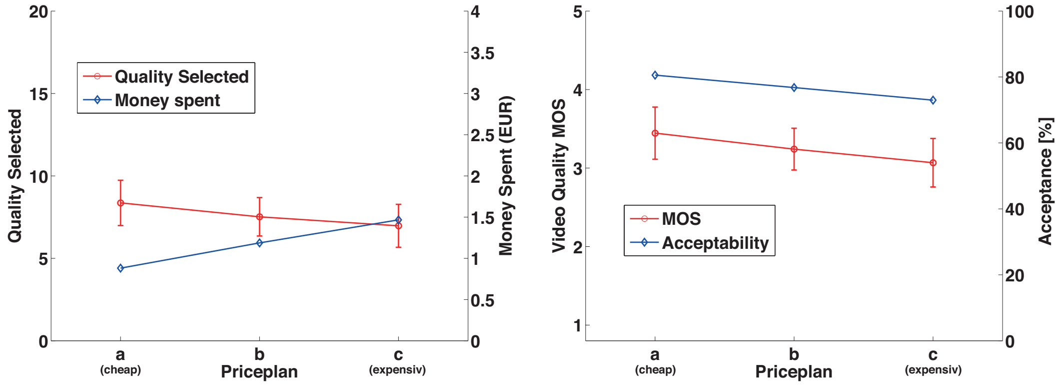


Figure 3.17: Influence of price plans on average selected quality and money spent (left) and average subjective perceived quality and acceptance (right)

depicted. Whereas the selected quality for the increasing price plan (A-B-C) did not change during the first two iterations, the subjective ratings regarding video quality MOS and the acceptance rate decreased (red lines). For decreasing prices (C-B-A) there was hardly any change in selected quality and MOS ratings. It seems that the subjective quality perception is influenced by previous pricing factors: if a comparison between *current* and *previous* iteration leads to a *negative* evaluation, i.e., the identical technical quality is more expensive, the resulting subjective perceived quality is lowered. The depicted amount of spent money per price plan pattern is in line with the findings stated in the previous paragraph: whereas quality is held constant, personal payment strategies are adapted to increasing or decreasing prices.

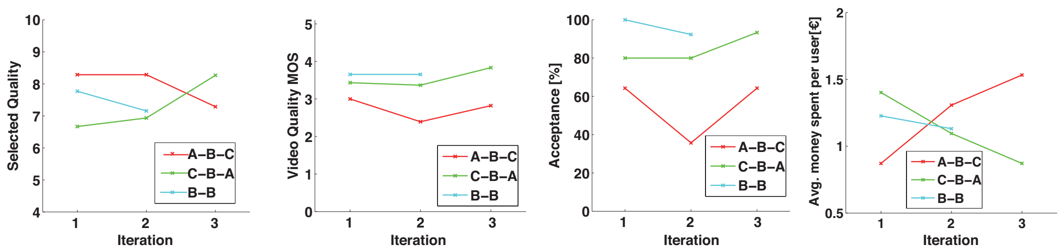


Figure 3.18: Study *WTP 2*: Influence of price plan patterns on average selected quality, average subjective perceived quality, acceptance rate and average amount of spent money.

As shown in Table 3.4 the quality classes Q16, Q17, Q18 and Q19 were identically coded with 32,768 kBit/s, but differently charged. Hence, from a rational point of view it does not make any sense to invest in a quality class above Q16. Interestingly, nobody chose quality class Q16 whereas higher quality classes were selected 13 times, see Figure 3.12(b). Regarding the three price plans the distribution of choosing these three quality classes is equal: A=4 (11.1%), B=6 (10.7%) and C=3(8.1%). Therefore, the price plans have no influence on squandering money. Additionally, a t-test was conducted to compare the $DiffMOS$ values between quality selections above and below quality class Q16 but no differences have been found, i.e., generosity does not effect the quality perception in a negative or positive way. In particular, especially the segment of quality-seeking customers may be subdivided and targeted by different maximum prices according to their individual willingness-to-pay and price sensitivity.

After the first iteration, the participants filled out a questionary dealing with questions related to cognitive dissonance (the questionnaire and the related question items can be found in the Annex A). As depicted in Figure 3.16(b) there is a small tendency that with increasing quality classes (Q0 to Q15) and charging fees the approval of this question increases. However, for the quality classes higher than Q15 this approval drops. It seems that spendthrift users have no problems with paying the highest price even if it is unnecessary from a rational point of view. Hence, spending more money for better video quality leads to slightly higher cognitive dissonance, but spending the highest amount of money reduces this effect significantly.

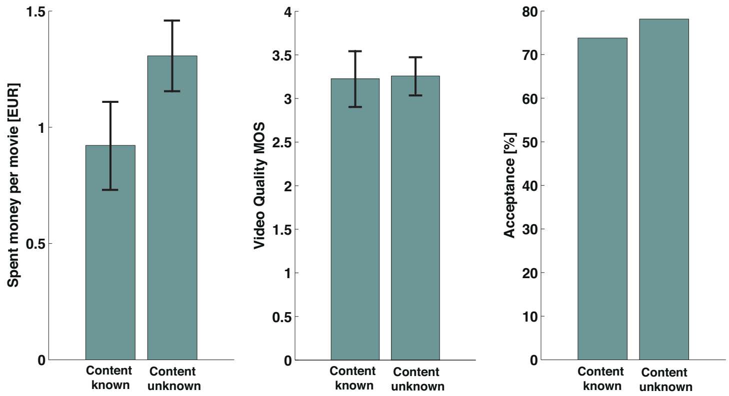


Figure 3.19: Study *WTP 2*: Content knowledge and user behavior (90% confidence intervals)

After each selected movie the users were asked if they have already known the movie. 42 movies were indicated as known (32%) and 87 were indicated as not known (68%). Hence, roughly a third of the movies were already known by the users who selected these movies. According to Figure 3.19 users spend more money if the movie is unknown, i.e., the movie is consumed in higher quality. Nevertheless, there is no significance regarding the acceptance rate or the resulting video quality MOS values.

In the study *WTP 1* a simple customer segmentation was conducted which resulted in four user segments. According to the previously mentioned methodological shortcomings, a valid and reliable user classification for study *WTP 2* is rather unfeasible and is not discussed here, for more details please see [153].

3.1.3.3 Conclusions

According to the findings plotted in Figure 3.14 and 3.13 it is not possible to verify the hypothesis that individual economic decisions positively impact the subjective video quality assessment in a VoD context. Additionally, the low amount of quality ratings per quality class makes it difficult to get profound results. Nevertheless, the aggregated $DiffMOS$ findings in Figure 3.14(b) show that there *might* be a positive impact of economic user decisions for the lowest and the highest available quality. Hence, to get a definite evaluation of the "positive impact"-hypothesis, an additional WTP user study is necessary which focuses on fewer quality classes to achieve reasonable numbers of selection per class. Nevertheless, especially the aggregated $DiffMOS$ findings for quality classes Q16-Q19 show that the irrationality of purchasing decisions — which has been discussed in the related work Section 3.1.1.1 — also impacts the subjective quality assessment and the quality selection: the overpriced quality classes Q17 to Q19 were selected disproportionally often. As depicted in Figure 3.16(b) increasing prices lead to higher cognitive dissonance, but this effect is abruptly reduced if the best and most expensive quality classes are selected. Hence, cognitive dissonance plays a not to be neglected role in subjective quality assessment. However, the small amount of ratings per quality class also makes it difficult in this case to draw final conclusions.

At least under the circumstances of the laboratory experiment there is a tendency that users prefer *constant quality* over *constant payments*, see Figure 3.17: Even if higher prices are charged the selected technical quality remains constant. It has also been shown that there is a negative impact of increasing price plans on quality assessment as depicted in Figure 3.18: Whereas the average selected quality remains constant the MOS value and the acceptance rate drop (red lines in Figure 3.18). Nevertheless, if these results are applied to real world scenarios, one should strongly consider the artificial laboratory setting and — once again — the low amount of ratings per quality class.

Roughly one third of the selected movies were already known by the users but more money was spent for movies which were unknown, see Figure 3.19. Therefore, to stimulate purchasing behavior a large amount of latest movies should be provided. Nevertheless, resulting MOS and acceptance rates indicate no impact of the users' acquaintance with selected movies.

3.1.4 WTP 3 User Study

To overcome the shortcomings of the study *WTP 2* — especially the low number of ratings per quality class — severe modifications in terms of the selectable content classes and the pricing aspects were urgently needed to get valid results regarding the hypotheses that individual, economic decisions positively affect quality assessments.

3.1.4.1 Setup

For the study *WTP 3* a highly similar setup — like the one in the study *WTP 2* — was used, see Section 3.1.3.1. In order to avoid small sample sizes for the selectable quality classes the amount of selectable quality classes was reduced from 20 to 4. Furthermore, the video duration was reduced from 20 minutes to 15 minutes to get one additional iteration, i.e., overall each user was able to select and evaluate 4 videos instead of 3 videos. In contrast to the study *WTP 2*, the PC monitor (see Figure 3.11) displayed the actual deposit, the resulting deposit if the current selected quality was applied for the complete movie, and the currently selected quality indicated by green bars (these green bars were not used in the study *WTP 2*). Similar to the previous study in 2012, there were reference video ratings before and after the movie selection phase. The same videos were used as in the study *WTP 2*, see Table 3.7. Table 3.9 depicts the video bitrates and the costs per quality class for the study *WTP 3*. Each user received 10 € as a deposit and selected overall four movies. Hence, the paid out deposit at the end of the experiment was between 0 € and 2 €.

quality class	**price per movie [EUR]**	**video Bit rate [kBit/s]**	**overall amount of ratings**
Q1	0	181	15 (ca. 10%)
Q2	0.7	724	54 (ca. 38%)
Q3	1.4	1448	39 (ca. 27 %)
Q4	2	32768	36 (ca. 25 %)

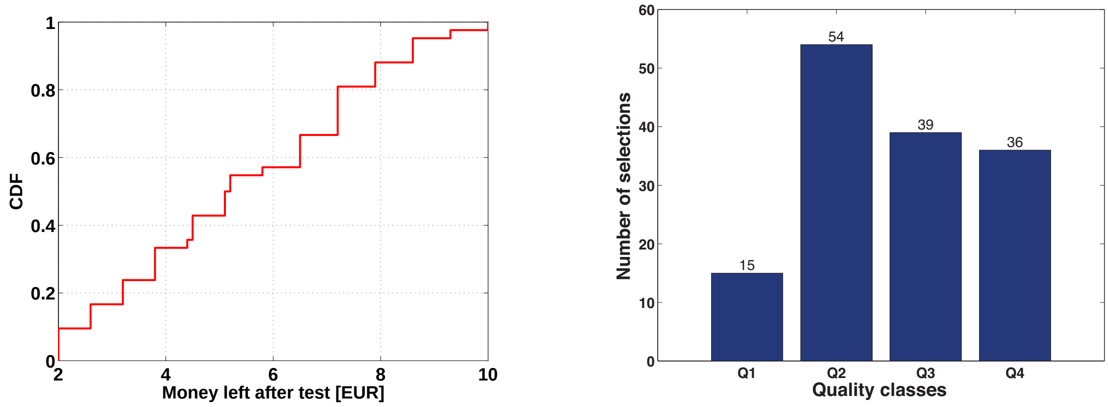
Table 3.9: Study *WTP 3*: Overview about selectable quality classes in study *WTP 3* with video bitrates, fees and the overall amount of ratings at the end of the trial.

3.1.4.2 Results

Overall, 36 users participated in the study *WTP 3* (18 female and 18 male users; mean age was 38.6 years with a median of 34 years; age groups: 13 users were below 30 years old, 12 users between 30 and 44 years and 11 users were above 44 years old).

Figure 3.20(a) depicts the CDF for the money left at the end of the experiment, i.e., the paid out cash to the test participants. Similar to the previous studies, the combination of initial deposit, movie length, video content and the price for enhanced video quality led to suitable results: only one user did not spend any money on quality enhancement and therefore received the maximum paid out deposit of 10 €.

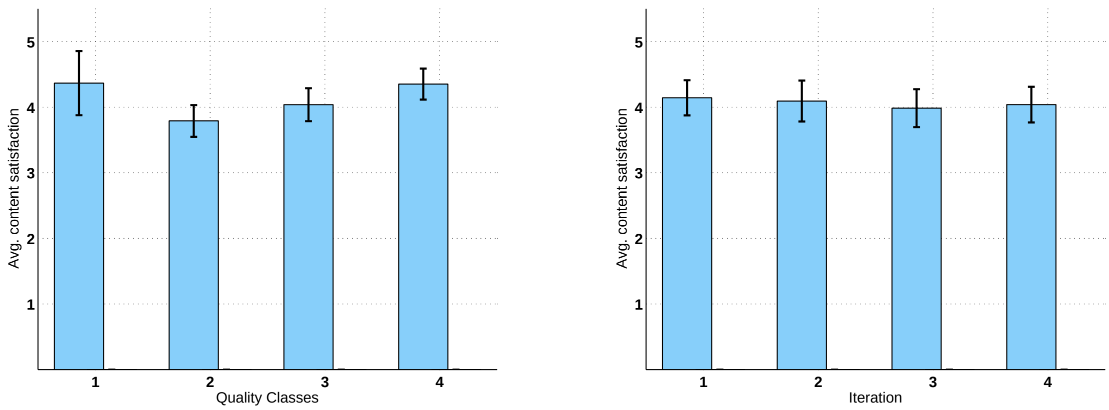
The same video content was used as in the study *WTP 2* and the users' satisfaction regarding the selected content was evaluated per movie, i.e., after each selected movie the participants were asked: "How satisfied are you with the content of the selected movie?" with answering options ranging from "very satisfied" (5) to "not satisfied" (1). The mean value for all users was 4.07 (median=4.13), i.e., on average the content was evaluated as "satisfying". Figure 3.21(a) shows the results for content-satisfaction separated by selected quality class. Obviously, there are no significant differences regarding the selected quality and the resulting content satisfaction. Figure 3.21(b) additionally shows that the iteration had no impact on the content satisfaction, i.e., the users were constantly satisfied with the available content. Hence, it was avoided that after the first movie selections users were not able to find and select interesting content.



(a) Paid out money at the end of the test

(b) Number of selections for each quality class for all users and all iterations

Figure 3.20: Study *WTP 3*: Money left and quality selections.

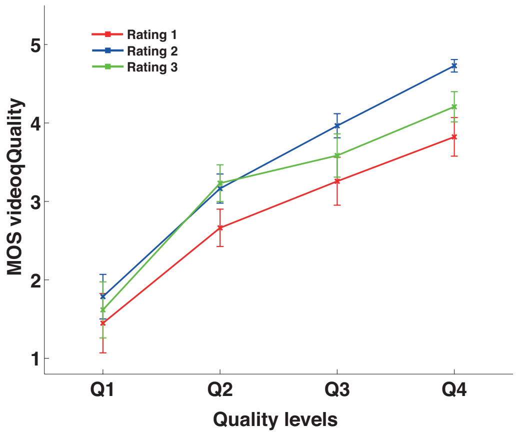


(a) Content satisfaction vs. selected quality classes

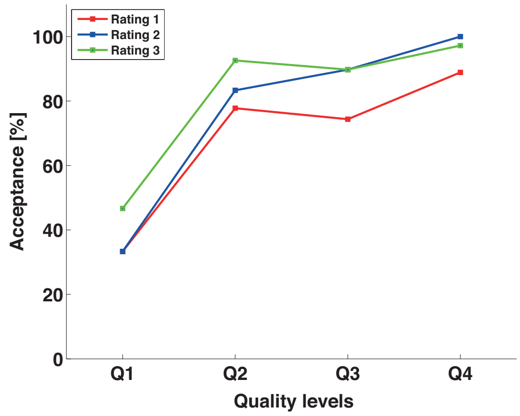
(b) Content satisfaction vs. iteration during study

Figure 3.21: Study *WTP 3*: Content satisfaction.

In the previous study *WTP 2* the relatively large amount of available quality classes led to unevenly distributed quality ratings and also to too few ratings in some quality classes, see Figure 3.12(b). Figure 3.20(b) depicts how often each of the 4 available quality classes was chosen during the WTP 3 user study. Each user — 36 in total — selected four movies and selected one of four quality class for each movie, i.e., overall 144 quality evaluation ratings were gathered. According to Figure 3.20(b) the selections were not equally distributed. Whereas quality class Q2 was selected 54 times, quality class Q1 was only selected 15 times. Hence, this data makes statistical analysis reasonable but a more frequent selection of quality class Q1 would lead to more significant results. However, the reduction from 20 to 4 quality classes led to a useful basis for the analysis.



(a) Video quality ratings for study WTP 3 (90% confidence intervals)



(b) Video quality accceptance ratings for study WTP 3

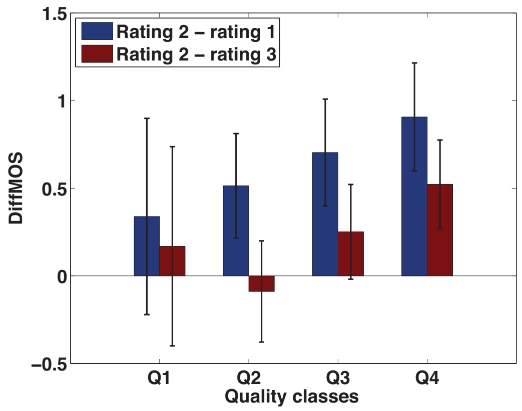
Figure 3.22: Study *WTP 3*: MOS and acceptance results.

Figure 3.22 depicts the results of the subjective video quality ratings regarding MOS values and acceptance rates. Ratings 1, 2 and 3 are explained in Table 3.5. Due to the higher amount of ratings per quality class the confidence intervals are smaller compared to the study *WTP 2*. Nevertheless, for the lowest quality class Q1 all confidence intervals are overlapping, which means there is no significant influence of priced quality selection on quality assessment. For higher quality classes (Q2-Q4) the MOS values for rating 1 (red line - reference ratings without user decision regarding quality) and rating 2 (blue line - reference ratings with user decision regarding quality) are significantly different, i.e., regarding these quality classes there is a positive effect of individually selected, charged quality on subjective quality assessment. The acceptance ratings in Figure 3.22(b) confirm this finding: Except for the lowest quality class Q1 the acceptance is higher — up to +16% for quality class Q3 — if the user individually selects the movie and the charged quality.

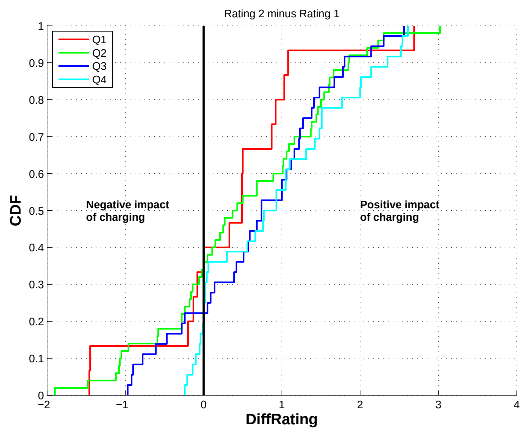
The findings in Figure 3.23(a) regarding $DiffMOS$ confirm the impact of charged quality on quality perception: Especially for the difference between rating 2 and rating 1 the resulting value is clearly positive for quality classes Q2-Q4 (blue bars in Figure 3.23(a). Additionally, the CDF-plots in Figure 3.23(b) clearly show that most $DiffRatings$ are positive. Also the $p_{ranksum}$ values depicted in Table 3.10 confirm the significant positive impact of individual economic decisions on subjective quality assessment.

Figure 3.24(b) depicts which movie was selected how often (Table 3.7 provides an overview about the movies and the corresponding IDs). In contrast to the study *WTP 2*, the movies "Harry Potter and the Half-Blood Prince" and "Transformers: Dark of the Moon" were not chosen[11]. Nevertheless, the distribution of the movie selection seems to be reasonable.

[11] The English version of "Harry Potter and the Deathly Hallows: Part 2" was selected five times, which relativizes the non-selection of the German version of the movie (ID=7).



(a) $DiffMOS$ ratings (95% confidence intervals)



(b) CDF-plots for "rating 2 - rating 1"

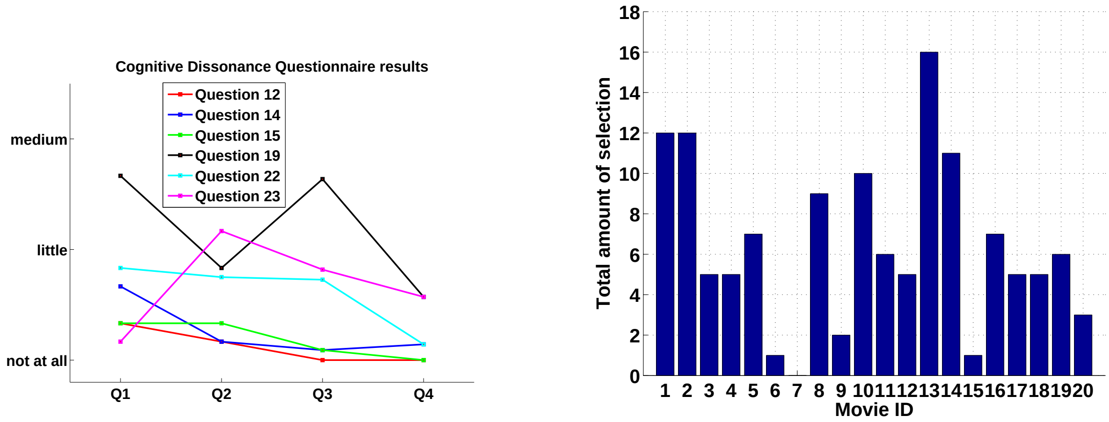
Figure 3.23: Study WTP 3: $DiffMOS$ and CDF-plots of $DiffRatings$

		$DiffMOS$	median	standard deviation	confidence interval	$p_{ranksum}$
rating 2 - rating 1	Q1	0.3387	0.4900	1.0171	-0.2210 - 0.8984	**0.0119**
	Q2	0.5132	0.4000	1.0933	**0.2149 - 0.8115**	**0.0074**
	Q3	0.7036	0.7400	0.9412	**0.3988 - 1.0084**	**0.0025**
	Q4	0.9064	0.8500	0.9117	**0.5982 - 1.2146**	**0.000001**
rating 2 - rating 3	Q1	0.1687	0	1.0327	-0.3996 - 0.7370	0.2608
	Q2	-0.0892	-0.0100	1.0580	-0.3779 - 0.1995	0.6771
	Q3	0.2511	0.1800	0.8338	-0.0190 - 0.5212	0.0917
	Q4	0.5225	0.3750	0.7472	**0.2699 - 0.7751**	**0.00016**

Table 3.10: Study *WTP 3*: Overview about ratings.

In the previous user study *WTP 2* the results from the cognitive dissonance questionnaire (see Figure 3.16(b)) indicated that spending more money on enhanced quality leads to increased dissonance which is finally reduced for the highest possible and therefore most expensive quality classes. The results of the identical cognitive dissonance questionnaire for the study *WTP 3* are less unambiguous: it seems that dissonance indicators (question items 12, 14, 15 and 22) are lower for higher quality classes, i.e., no initial increase could be observed. Also question items 19 and 23 seem to be a little bit random. Hence, the findings in the study *WTP 2* regarding cognitive dissonance as a possible explanation of observed quality perception could not be verified in the study *WTP 3*.

As described in Section 2.2.2.1, a questionnaire about expectations was included in the study *WTP 3*. The participants were asked about their general *desired* expectations regarding VoD services. At the beginning of the experiment the participants were invited to rank five items according to their individual importance in the context of Video-on-Demand services: (1) large amount of available movies, (2) excellent support, (3) low



(a) Cognitive dissonance questionnaire results   *(b)* Movie selections

Figure 3.24: Study *WTP 3*: Cognitive dissonance and movie selections

costs, (4) excellent video quality and (5) short contract duration. In the context of the user study *WTP 3* the relevant items were no. 3 (ordinal variable $expect_{costs}$, ranging from 1 to 5) and no. 4 (ordinal variable $expect_{quality}$, ranging from 1 to 5), i.e., according to their positions in the ranking the individual importance and relevance of video quality and costs in VoD-contexts was derived for each user.

This information was used to model the user rating behavior, as shown in the the following model extension steps. The common IPTV video quality model ITU-T P.1201 [154] was used, which is described in detail in [155]. In [155] the predicted video quality QV is calculated via the quality impact of video compression ($QcodV$) and the impact of packet loss ($QtraV$).

$$QV = 100 - QcodV - QtraV \tag{3.1}$$

For the purpose of the study *WTP 3* the packet loss impairment factor $QtraV$ is neglected. $QcodV$ is calculated as follows:

$$\begin{aligned} QcodV = (a_{1V} + ResC) * e^{a_{2V} * BitPerPixel * Res} \\ + a_{3V} * ContentComplexity + a_{4v} \end{aligned} \tag{3.2}$$

In the study *WTP 3* all videos were in HD resolution and had identical frame rates. Therefore, $ResC$ and Res can be treated as constant factors. $BitPerPixel$ is calculated as follows:

$$BitPerPixel = \frac{VideoBitrate * 10^6}{NumPixelsFrame * FrameRate} \tag{3.3}$$

NumPixelsFrame and *FrameRate* are the same for all videos, only the *VideoBitrate* depends on the selected quality class.

Similar action content was used, i.e., *ContentComplexity* is applied as a constant factor to reduce the complexity of the model. a_{1V}, a_{2V}, a_{3V} and a_{4V} are the coefficients, their values depend on the final fitting. By excluding all constant factors, the following simplified and special-purpose model can be used to describe the ratings impaired by economic quality decision. Since the model describes the quality via the R-scale (ranging from 0 (=worst quality) to 100(=best quality), see [81] for more details about the R-scale), the subjective ratings from the participants in the *WTP 3* study need to be converted to this scale. Hence, the conversion described in the ITU-T Recommendation G.107 [81] was applied, which was also used in the context of the video model P.1201.2, for an example see [156].

The resulting video quality model (Equation 3.4) was applied to the converted ratings[12] regarding situations in which users consumed a specific video via a specific, charged quality class. According to the first line of Table 3.11 significant coefficients and an adjusted R^2 of 0.68 and an RMSE of 13.63 were achieved[13].

$$R = 100 - a_1 * e^{a_2 * VideoBitRate} \tag{3.4}$$

By expanding the model shown in Equation 3.4 with two linear terms — representing individual, desired expectations regarding video quality and costs — the adapted model presented in the second line of Table 3.11 is generated. It finally includes $VideoBitRate$ and individual, general expectations regarding desired video quality (variable $expect_{quality}$) and desired service costs (variable $expect_{cost}$):

$$\begin{aligned} R = 100 - a_1 * e^{a_2 * VideoBitRate} + \\ a_3 * expect_{quality} + a_4 * expect_{costs} + a_5 \end{aligned} \tag{3.5}$$

By applying this model to the data an adjusted R^2 of 0.73 and an RMSE of 12.58 are calculated. The sum of the squared, linear correlations between the residuals of the model stated in Equation 3.4 and the variables $expect_{quality}$ and $expect_{costs}$ is 0.0941, i.e., these variables add an additional explanatory value of approx. 9.4% to the model.

[12] MOS values to R values.

[13] Significant coefficients: p-value for the t statistic that the coefficient=0 is smaller than 0.05; adjusted R^2: Proportion of total sum of squares explained by the model, number of coefficients are considered; RMSE: Root mean squared error

Hence, including individual expectations in models for charged situations leads to an enhanced MOS prediction accuracy. Table 3.11 depicts the modeling results and specific values for the coefficients.

independent variables	model	RMSE	adj. R^2	add. expl. Value
VBR	$R = 100 - 80.54 * e^{-0.00104*VideoBitRate}$	13.63	0.68	-
VBR, $Expect_{quality,costs}$	$R = 100 - 81.67 * e^{0.0012*VideoBitRate} - 4.49 * exp_{quality} - 4.72 * exp_{costs} + 18.68$	12.58	0.73	ca. 9.4%

Table 3.11: Study *WTP 3*: Model overview

3.1.4.3 Conclusions

The results of the final study *WTP 3* demonstrate that there is a *significant, positive effect of individual, economic decisions on subjective quality assessment* of increasing prices. Hence, if a customer does not have to pay for delivered quality — but the decision for this type of consumption is made, i.e., ”worst quality for free” — the subjective quality assessment is not impaired. This finding is in line with the price-quality relationship discussed in Section 3.1.1.2, i.e., it seems that higher prices are used as indicator for higher quality. Hence, the resulting quality assessment ratings are higher.

As shown in Figure 3.16(b) only a small impact of cognitive dissonance is visible. In contrast to the results of study *WTP 2*, no increase of dissonance could be observed and there was no drop for the highest quality class happened. Hence, a strong impact of cognitive dissonance could not be verified due to the fact that only a small amount of quality classes — Q1, Q2, Q3 and Q4 — was selectable.

An approach to measuring individual expectations in order to enhance the accuracy of MOS prediction models has been discussed in Section 2.2.3. This method has also been used to enhance the accuracy of a video quality MOS model by 9.4% for the results of the user study *WTP 3*.

3.1.5 Recommendations for User Studies Involving Individual, Economic Decisions

In this section recommendations — which have been developed in recent years via the user studies *WTP 1*, *WTP 2* and *WTP 3* — are presented as to how to successfully conduct QoE-related user studies with a focus on individual, economic user decisions.

20 selectable action videos guarantee satisfied users: In ideal video quality evaluation studies all users evaluate a limited, standardized set of short videos encoded via different qualities [157]. However, a large library of various videos should be offered in order to enable users to select desired and interesting videos. Yet, this is at odds with the aforementioned design principle because . Therefore, a balance is necessary to conform scientific standards and also to legitimate individual decisions regarding video content. Reducing the amount of available content classes from initially five in the study *WTP 1* (movies, TV series, documentary reports, animation and music concerts; 25 videos overall) to one in the study *WTP 3* (action movies, 20 videos overall) does not significantly affect the users' satisfaction with the presented video library size, i.e., the test participants agreed that the content library was large enough to find interesting videos. Hence, one content class (*action movies*) containing 20 videos is sufficient.

study	**n**	**nr. quality classes**	**init. depo-sit**	**max. fee per movie**	**min. payout**	**QoS**	**focus of study**
WTP 1	43	4	10 €	1.5 €	5.5 €	packet loss	Combine active users with quality evaluation
WTP 2	43	20	10 €	4 €	0 €	VBR	HD content & varying price plans
WTP 3	35	4	10 €	2 €	2 €	VBR	Validate results & setup, include expectations

Table 3.12: WTP studies overview

Video clip durations of 15 minutes are long enough: For a pure video quality evaluation short clip durations (10-20 seconds) are sufficient, see [142]. Additionally, short videos accommodate an efficient testing procedure, i.e., more clips can be evaluated in the available amount of time. However, in the WTP user studies the participants were supposed to pay for enhanced video quality of the selected content and obviously nobody would pay extra fees for a clip that only lasts for a few seconds. From a user's perspective the entire movie duration would be the most appropriate video length (e.g. 90 - 120 minutes), which is of course not feasible in an experimental, laboratory setup as compared with a field study. Therefore, choosing the appropriate video length is crucial. In the study *WTP 1* the used videos had a duration of 20 minutes, which was finally reduced to a duration of 15 minutes in the study *WTP 3*. In all three WTP user studies the majority of the users spent a considerable amount of money and most of them were satisfied with the video duration. Therefore, a video length of 15 minutes is long enough to evoke valid economic decisions in terms of a trade-off between money and quality.

Avoid large numbers of selectable quality classes: Since users are allowed to choose quality classes by themselves the resulting ratings are *not* uniformly distributed across the available options and the final distribution is only available at the end of the experiment, i.e., it is difficult to intervene during the execution of the study. As shown in Table 3.9, there are 15 ratings in quality class Q1, but 54 ratings in Q2 in the study *WTP 3*. However, a minimum amount of ratings for each class is required to obtain valid statistical results. Therefore, according to the gained experiences the following instruction helps to calculate how many users are needed to get valid results (note: number of selectable quality classes and ratings per test participant are predefined).

Assuming that the ratings were uniformly distributed (=ideal case) over all selectable quality classes the amount of ratings per quality class would be:

$$Ratings/class_{uniform} = \frac{Nr.\ of\ ratings}{Nr.\ of\ quality\ classes} \tag{3.6}$$

According to the empirical findings during the WTP user studies, it is assumed that in reality the most unpopular class only reaches approximately 50% of this amount:

$$Ratings/class_{worst\ case} = \frac{Nr.\ of\ ratings}{Nr.\ of\ quality\ classes} * \frac{1}{2} \tag{3.7}$$

To get valid results the authors of [30] recommend a minimum of 15 ratings per quality class:

$$15 = \frac{Nr.\ of\ ratings}{Nr.\ of\ Quality\ Classes} * \frac{1}{2} \tag{3.8}$$

Therefore, the number of ratings should be:

$$Nr.\ of\ ratings = 30 * Nr.\ of\ quality\ classes \tag{3.9}$$

Hence, to calculate the number of users to participate in the study the following equation can be used:

$$Users_{planned} = \frac{30 * Nr.\ of\ quality\ classes}{ratings\ per\ user} \tag{3.10}$$

For example, if each user evaluates 4 videos and 4 quality classes are available the minimum amount of required users is 30. If the amount of selectable quality classes is duplicated, i.e., 8 classes are available, 60 users are needed to get a minimum of 15 ratings for the least often selected quality class.

Provide a free option and always pay out: It is important to find the right balance between initial deposit and the price for enhanced video quality. In Table 3.12 an overview of all three conducted experiments is given including information about initial deposit, minimum pay outs and maximum fees per movie. The study *WTP 3* provides the most uniform and therefore desired distribution of spent money, as the CDF in Figure 3.20(a) shows. In the user study *WTP 3* the minimum amount of pay out was €2 with an initial deposit of €10. Therefore, it is proposed that subjects received at least 20% of the initial deposit at the end of the test session in any case, even if the most expensive option was always chosen by the test participant. Additionally, the lowest quality should always be for free to avoid an enforced spending of money.

Conduct reference ratings before and after the economic part: In the studies *WTP 2* and *WTP 3*, before and after the actual WTP/QoE part was conducted, reference ratings were collected and labeled as rating 1 and rating 3. To decide which of these ratings were more consistent, the average standard deviation over all quality classes was calculated, see Table 3.13. In the study *WTP 2* the average standard deviations of ratings 1 were smaller compared to ratings 3, i.e., one can assume that the ratings were more consistent. In contrast to this, in the study *WTP 3* the ratings 3 seemed to be more consistent because of smaller average standard deviations. Ratings 3 referred to quality class and content class regarding the previous selections, but ratings 1 were made without any user influence, i.e., they represent more adequately the common way of video quality assessment. Hence, at the moment it is recommended to conduct both types of reference ratings for QoE studies containing individual, economic user decisions.

Video switching time of two seconds is good enough: As described in Section 3.1.3.1, in the studies *WTP 2* and *WTP 3* Apple's HLS was used to enable video quality switching during playback, which finally resulted in a switching time of approx. 2 seconds, i.e., after the user had selected a quality class it took 2 seconds before the quality of the currently played video adapted. The switching time resulted from 1 second HSL video file chunks and 1 second VLC internal adaption time. Faster switching times might be technically realizable for audio-video content, in the WTP user studies a switching time of 2 seconds was appropriate for the users.

		study *WTP 2*	study *WTP 3*
rating 1	avg. standard deviation (combined)	0.7337	-
	avg. standard deviation (all)	0.6448	0.9720
rating 3	avg. standard deviation (combined)	1.1485	-
	avg. standard deviation (all)	0.7632	0.8815

Table 3.13: Average standard deviation for ratings 1 and 3 during study *WTP 2* and *WTP 3*.

Displaying selected quality class is optional: To examine real world scenarios the currently selected quality class was presented to the test participant in the study *WTP 3*, i.e., like in everyday life the user was informed about her current deposit, fees and which quality was currently selected. In contrast to this, if the unadulterated effect of spending money on quality perception is to be investigated no quality class related information should be displayed. Therefore, displaying quality information to the user is optional and depends on the specific research question.

Use within-subjects design: In the study *WTP 1* a *between-subjects* design was applied: users of group A selected the content and the quality classes and subsequently evaluated the video qualities, whereas users of group B evaluated the video qualities without any content/quality decisions. Then, the ratings were compared between the groups, see Figure 3.9. In the studies *WTP 2* and *WTP 3* a *within-subjects* design was applied, e.g., each user evaluated the individually selected videos *and* the reference videos. Although within-subject designs entail disadvantages (e.g. learning effect, fatigue, etc. see [158]), the positive effect of reduced errors caused by individual differences outweighs, especially if the objective is to examine individual factors affecting subjective quality perception.

Use DiffMOS, confidence intervals and CDFs to interpret individual difference ratings: If the recommended within-subjects design is applied, a $DiffRating$ can be calculated for each user and each condition, which can be further used to calculate the $DiffMOS$ values, see Section 1.3. Additionally, for each $DiffMOS$ value the corresponding confidence interval can be calculated and plotted as error bars, see for example Figure 3.23(a). Hence, if the confidence interval does not include zero, one can assume that a significant effect of the economic user decision on quality assessment occurred. Additionally, it is recommended to plot CDFs of the $DiffRatings$, see for example Figure 3.23(b) and to include Wilcoxon rank sum tests.

Include user expectations data to enhance rating prediction: As discussed in the previous section, including individual expectations in modeling leads to enhanced MOS prediction accuracy. The approach described in Section 2.2.2.1 was applied.

3.2 Non-Economic Decisions

In the previous section experiments regarding the impact of individual, *economic* decisions on subjective quality assessment has been presented. In this section, the impact of non-economic user decisions on quality assessment is discussed.

Current technological evolution in the context of network-based services — mobile gaming, 4k video streaming, cloud applications, etc. — demands not only highly sophisticated network infrastructure, but also novel and holistic approaches to measuring and evaluating subjective quality perception to guarantee high Quality of Experience, which requires comprehensive QoE models and QoE management solutions. So far, there are several QoE frameworks and models which include various impact factors to predict perceived quality of various technological systems. Some of them focus on technical features [159] or user features [36], etc. In the context of user features, [160] emphasizes the role of individual goals and [161] stresses the importance of motivation on QoE. Also, the impact of moods and emotions on QoE has been addressed in empirical user studies [162], as well as the role of individual expectations in quality perception, see Chapter 2.

However, so far, the role of *non-economic, individual user decisions* has not been considered in the field of QoE evaluation, modeling and management. Therefore, a paradigm shift from passive users towards *active, decision-making users* is necessary. As a consequence, the common way to conduct empirical QoE studies needs to be adapted: in common QoE experiments most factors like content to consume, devices to use, etc. are strictly *predefined* by the test design. As a consequence, the test user has no chance to make any individual decision and has to act in a prescribed way. In contrast to this, in real world scenarios most of these aspects are determined by the user himself, e.g., the user decides if she wants to access a certain webpage via a smartphone or a laptop.

Therefore, the commonly gained laboratory results might be biased by these predefined aspects, which has not been examined so far. Hence, it is necessary to examine the the impact of non-economic, individual user decisions on subjective quality assessment results.

This section sets out to answer this question on on the basis of the results of three empirical, laboratory Web and video QoE experiments that examine the influence of user decisions on quality evaluation ratings.

3.2.1 Background: Non-Economic Decisions

The following sections present an overview about relevant (socio-) psychological studies which empirically examine the role of user decisions on human behavior and related fields. Many relevant aspects of individual decisions have already been discussed in Section 3.1.1, which is why only some additions are presented regarding available options and decisions.

The influence of distinct choices on an individual's wellbeing and the inherent need for control over the environment has been discussed in various research fields. According to the authors of [163] there are various types of choices, e.g., there are single, complex and extensive decisions like which university to attend and there are also myriad, small and everyday decisions which might be below the state of the user's conscious awareness. The mere existence of options influences the individual satisfaction, even if additional options do not provide any additional value, see [164]. Many experiments have been conducted to examine the influence of user decisions on attitude change, e.g., the author of [165] presents an experiment in which the participants had to rank art prints from the most liked to the least liked. Afterwards, the subjects had to choose between two pairs of the prints and finally they had to rank the prints again. The test participants typically re-ranked the chosen prints higher and the rejected options lower compared to the initial ranking. According to [163], the need for control and therefore the substantial necessity of making individual decisions is not acquired through learning, but rather innate. This means that "we are born to choose" [163].

In general, one has to distinguish between *having* and *making* a choice, e.g., in [166] it is shown that 95% of 823 study participants indicated that *having* choices in the context of healthcare is very important, whereas only 30% revealed that *making* a choice is very important. Also, the *amount* of available options is crucial in the context of individual decisions. In a study conducted by [167] the test participants were divided into two groups: subjects in group A had to choose from an array of six gourmet jams, whereas in group B the participants had to choose from an array of 24 to 30 jams. Participants from group A were more satisfied with their selections, i.e., too many available options might reduce the contentment with the individual choice.

Regarding methodological aspects of QoE assessment in laboratory environments, the authors of [168] found out that the provisioning of choice increases intrinsic motivation and enhances performance on a variety of tasks. Especially in the context of repetitive, non-exciting tasks like evaluating the perceived quality of short video snippets, the implementation of user decision may increase the validity of the gained results.

To summarize, there is widespread empirical evidence that individual user decisions in general influence the attitude of the user and satisfaction. In the specific context of QoE the influence of individual, non-economic user decisions on subjective quality perception has been hitherto neglected.

3.2.2 Experimental Setups

Three independent user studies have been conducted to examine the influence of individual, non-economic user decisions on subjective quality assessment. All studies were conducted at FTW's iLab in Vienna. Due to the initial investigation perspective and in order to get a broad insight, each study focused on a specific aspect related to user decisions: content, used end device and quality impairment. The selectable amount of options was limited, e.g., the users could choose between *three* end devices or choose between *two* Web sites. This limitation was necessary to guarantee a sufficient amount of ratings per selectable option (see also recommendations for economic user studies in Section 3.1.5). Additionally, related work shows that a smaller amount of available options leads to higher user satisfaction and decisions are made more easily, see [167]. Table 3.14 gives an overview about the three experiments.

study	**decision regarding**	**user task**	**device(s)**	**varied quality parameter (QoS)**	**n**
Decision 1	quality impairment	watch videos	large TV screen	initial delay, stallings	26
Decision 2	end device	browse web pages	smartphone, laptop, tablet	downlink network speed	36
Decisions 3	Web content	browse web pages	laptop	downlink network speed	29

Table 3.14: Non-economic experiments: Study overview

3.2.2.1 Temporal Impairment Decision (Study Decision 1)

In this study the users had to evaluate the *disturbance* of two kinds of temporal impairments during video streaming: initial delays (i.e. the video playback was paused at the beginning of the video) and stalling events during the video playback. At the beginning of the trial the users — who were sitting in front of a large HD TV screen — had to evaluate eight short videos impaired by a certain temporal impairment (see Table 3.15). After each video playback had finished, the user had to rate the subjectively perceived disturbance via a 5-point ACR Scale, ranging from ”not disturbing” to ”very disturbing”. These so-called *reference ratings* were not affected by any user decision. The videos were generated from HD Action movie sequences.

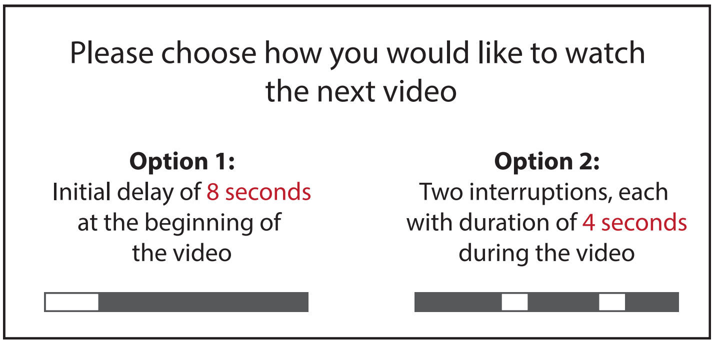


Figure 3.25: Study *Decision 1*: Example for a question regarding temporal impairment, displayed via a large TV screen.

After some other video quality evaluation tasks, the evaluation of the previously described videos started again. This time the users had to make a decision regarding the temporal impairment before the video was shown: Before the playback of each video started, a short question was displayed on the TV screen, see Fig 3.25 for an example. The user had to make her decision and the video with the selected impairment was displayed and afterwards the user had to evaluate the annoyance of the temporal impairment. Overall, each user had to make 16 decisions (all combinations of 4 initial delays and 4 stalling events).

video	init. delay [sec.]	stalling [sec.]	video duration [sec.]
1	-	2x1	32
2	-	2x2	34
3	-	2x4	38
4	-	2x8	46
5	2	-	32
6	4	-	34
7	8	-	38
8	16	-	46

Table 3.15: Study A: Evaluated videos; video duration was 30 seconds + temporal impairment duration, e.g., overall playback duration of video 4 was 46 seconds

3.2.2.2 End Device Decision (Study Decision 2)

In this study the participants had to evaluate the subjective quality regarding Web QoE, i.e., they accessed several Web pages via varying downlink bandwidth speeds. At the beginning, the subjects had to use Google Maps via an iPhone 5, a Nexus 7 tablet (Version 2013) and a 15.4 inch HP laptop in random order and with different network speeds (downlink bandwidth was 256 Kbit/s (low), 1024 Kbit/s (mid) and 4096 Kbit/s (high)). After each task, the users had to evaluate the subjective experience regarding network speed via a standard 5-point ACR scale ranging from "excellent" to "bad", see

[83]. After some other tasks, the users had to decide with which of the three devices they would like to conduct the following task, which was also browsing Google Maps. During these conditions the downlink bandwidth was shaped too and subsequently the participants had to evaluate the perceived network speed. Therefore, for each user three reference ratings (low/mid/high downlink bandwidth) and three ratings (low/mid/high downlink bandwidth) which were potentially influenced by the end user device decision were collected.

3.2.2.3 Web Content Decision (Study Decision 3)

Similar to study *Decision 2*, in this study the users had to evaluate Web QoE, i.e., they browsed a news site (www.spiegel.de) and Google Maps on a Laptop with different downlink bandwidth values and with a subsequent quality assessment, which resulted in subjective reference ratings. To generate quality ratings impaired by user decisions, after some other tasks the users were asked: "For the next three conditions, do you want to browse a news site or Google Maps?". Google Maps was accessed via a 2, 4 and 8 Mbit/s downlink bandwidth and the News Site was accessed via 1,2, and 4 Mbit/s. For the analysis, the bandwidth levels are labeled as *low*, *mid* and *high*.

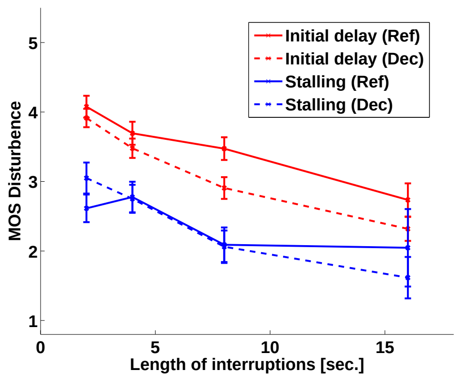
3.2.3 Results

Similar to the previously discussed experiments in this thesis, the $DiffRatings$ and the $DiffMOS$ values were calculated. Additionally, for each $DiffMOS$ the confidence intervals (90% confidence level) were calculated, which are used in the graphics in this section. For more information about $DiffRatings$ and $DiffMOS$ please see Section 1.3.

3.2.3.1 Temporal Impairment Decision during Video Streaming (Study Decision 1)

In study *Decision 1* ratings from 26 users were gathered (13 male and 13 female, with a mean age of 39.23 years and a median of 38.5 years). Figure 3.26(b) shows which impairment was chosen how often, e.g., stallings with an overall duration of 2 seconds were selected by 77% of the users instead of an 8 seconds initial delay, which was chosen by only 23% of the users. Obviously, the study participants prefer initial delay instead of stalling events during the video consumption. This is in line with the results presented in [169], where the users were more tolerant regarding (pre-selected) initial delay compared to (pre-selected) stalling events during video consumption. In Figure 3.26(a) the resulting MOS disturbance values are depicted.

Figure 3.27(a) shows the average $DiffMOS$ values for each temporal impairment. For individually selected initial delays the user decision always caused a significant *negative* impact regarding subjective quality assessment, i.e., the ratings were significantly lower. For individually selected stalling events there was only a significant *positive* effect for short stalling events (2 x 1 sec.). For longer stalling events there was no significant effect of individual user decisions on subjective quality assessment. Figure 3.27(b) shows the cumulative distribution function of all $DiffRatings$. The rating-distribution for 2 x 1 sec. stalling is clearly positive, i.e., most of the ratings (approx. 78%) are located on the right side of the vertical zero-line and are greater than zero. In contrast to this, approx. 75% of the $Diffratings$ regarding 8 sec. initial delay are negative, which indicates a negative effect of individual user decisions on subjective quality assessment ratings.



(a) MOS disturbance ratings (5=not disturbing, 1=very disturbing, 90% confidence intervals); Ref=Reference, Dec=Decision

vs.	**Stalling Events [sec.]**			
Initial Delay [sec.]	**2x1**	**2x2**	**2x4**	**2x8**
2	96%	100%	92%	96%
4	62%	77%	92%	96%
8	77%	58%	81%	92%
16	85%	62%	54%	73%

(Green indicates that stalling events were preferred, red indicates that initial delay was preferred)

(b) Percentual amount of selected temporal impairments.

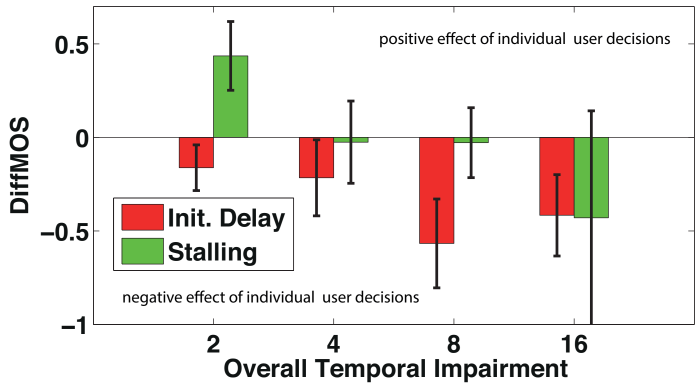
Figure 3.26: Study *Decision 1*: MOS disturbance values and selections

		$DiffMOS$	median	standard deviation	confidence interval	$p_{ranksum}$
initial delay	2	-0.1622	-0.1500	0.7349	**[-0.2842, -0.0402]**	**0.0459**
	4	-0.2160	-0.1100	1.1275	**[-0.4194, -0.0126]**	**0.0273**
	8	-0.5665	-0.3200	1.1201	**[-0.8040, -0.3290]**	**0.0001**
	16	-0.4164	-0.3100	0.8690	**[-0.6339, -0.1989]**	0.0681
stalling	2x1	0.4357	0.4600	0.7988	**[0.2520, 0.6194]**	**0.0192**
	2x2	-0.0254	0.1000	0.7928	[-0.2453, 0.1945]	0.8288
	2x4	-0.0283	0.0200	0.5232	[-0.2153, 0.1587]	0.9474
	2x8	-0.4300	-0.2200	1.0563	[-1.0020, 0.1420]	0.4473

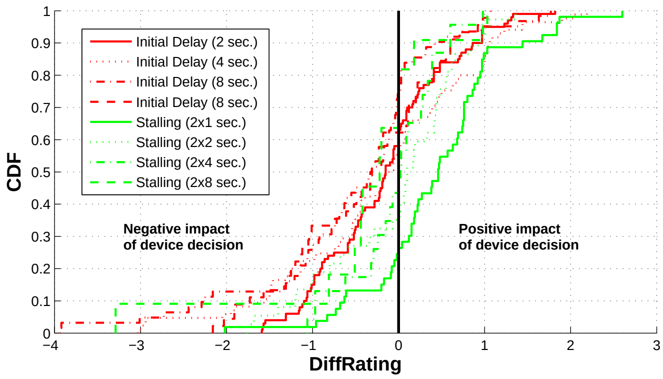
Table 3.16: Study *Decision 1*: Rating overview

3.2.3.2 End-Device Decision during Web Usage (Study Decision 2)

Overall, 36 users (19 female, 17 male with a mean age of 36.6 years and a median of 32 years) participated in this experiment. 10 participants selected the iPhone, 20 users chose the Android tablet and 6 users selected the laptop. Obviously, the low amount of users per device resulted in relatively large confidence intervals, especially for ratings concerning laptop decisions, see Figure 3.28(a). Figure 3.28(a) also reveals that individual device decisions have a *positive* impact on quality assessment for *low QoS*



(a) $DiffMOS$ values for all temporal impairments (90% confidence intervals)



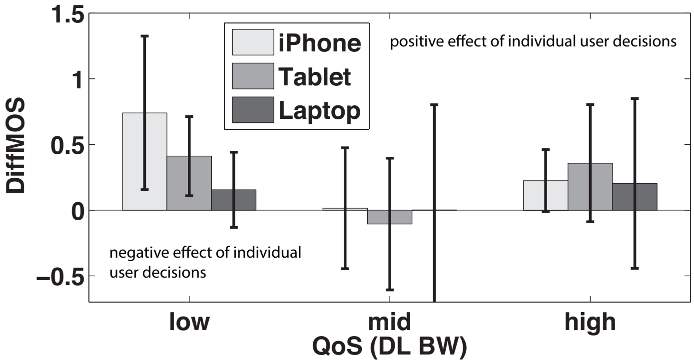
(b) CDF-plots of $DiffRatings$ for each initial delay and stalling value

Figure 3.27: Study *Decision 1*: $DiffMOS$ results and CDF-plots of $DiffRatings$.

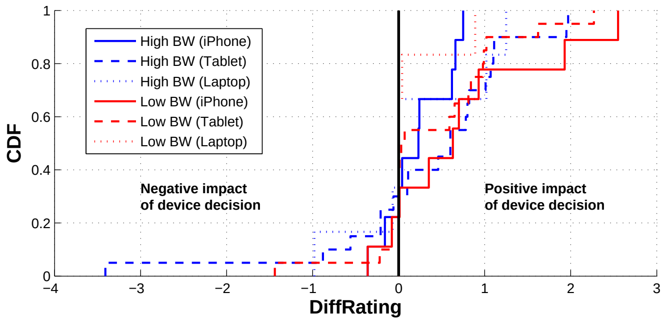
situations, e.g., browsing Google Maps with a downlink bandwidth of 256 Kbit/s. For higher downlink bandwidth levels there is no — neither positive nor negative — effect of individual device decisions. Figure 3.28(b) shows the distribution of the $DiffRatings$ of study *Decisions 2*. However, due to limited space only for low and high downlink bandwidth conditions. For low bandwidth conditions (red lines) most of the ratings are positive, approx. 70%.

3.2.3.3 Content Decision during Web surfing (Study Decision 3)

In study *Decision 3* ratings from 29 users were collected (13 male, 16 female with a mean age of 34.7 years and a median of 31 years). Overall, during the decision-phase of the experiment 16 participants selected Google Maps and 13 participants chose the news site. Figure 3.29(a) shows that there was no significant impact of individual user



(a) DiffMOS values for all selected devices (90% confidence intevals)



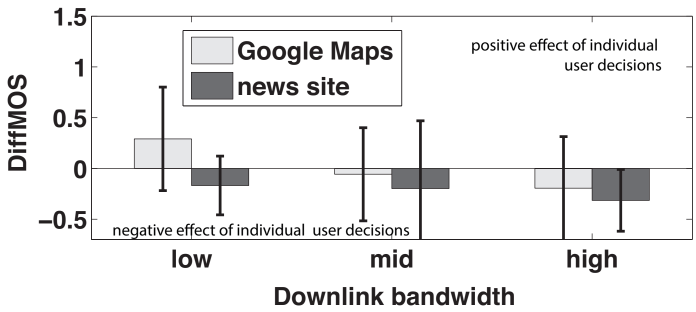
(b) CDF for each selected device

Figure 3.28: Study *Decision 2*: $DiffMOS$ results and CDF-plots of $DiffRatings$.

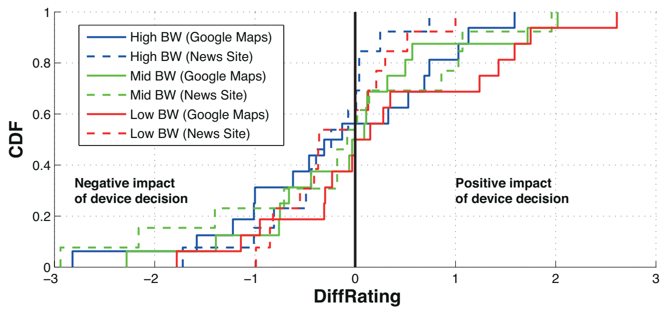
		$DiffMOS$	median	standard deviation	confidence interval	$p_{ranksum}$
low	iPhone	0.7400	0.6300	0.9559	**[0.1559, 1.3241]**	0.0864
	tablet	0.4110	0.0500	0.7827	**[0.1091, 0.7129]**	**0.0178**
	laptop	0.1550	0	0.3604	[-0.1309, 0.4409]	0.6970
mid	iPhone	0.0144	-0.0700	0.7532	[-0.4459, 0.4747]	0.7962
	tablet	-0.1055	0.0300	1.2988	[-0.6064, 0.3954]	1
	laptop	0.0017	0.2100	1.0090	[-0.7988, 0.8022]	1
high	iPhone	0.2244	0.2300	0.3867	[-0.0119, 0.4607]	0.1950
	tablet	0.3575	0.6000	1.1566	[-0.0886, 0.8036]	0.2733
	laptop	0.2033	0		[-0.4431, 0.8497]	1

Table 3.17: Study *Decision 2*: Rating overview

decisions on subjective quality assessment. Also the CDF-plots presented in Figure 3.29(b) indicate that neither a positive nor a negative effect prevails.



(a) DiffMOS values for web application selection (90% confidence intervals)



(b) CDF for both web applications

Figure 3.29: Study *Decision 3*: $DiffMOS$ results and CDF-plots of $DiffRatings$.

3.2.3.4 Influence of User Variables on DiffRatings

The CDF-figures 3.27(b), 3.28(b) and 3.29(b) reveal that for most conditions the $DiffRatings$ are widely distributed around the 0-line, i.e., there are positive *and* negative values. Therefore, for some users there was a positive and for some users there was a negative effect of individual decisions regarding subjective quality assessment. To obtain more detailed knowledge regarding the experiment participants, some additional user information like age, Internet usage habits, education, etc. was collected. Tables 3.18 and 3.19 show significant correlation results between this user information and related $DiffRatings$ for study *Decision 1* and *Decision 2*[14]. In study *Decision 2* the increased usage of laptops, tablets, Youtube and email services resulted in more positive $DiffRatings$, i.e., in this case the individual user decision had a more positive impact regarding subjective quality assessment ratings, see Table 3.19. In the video streaming study *Decision 1*, the Youtube usage frequency and the general TV satisfaction correlates with the use ratings for some initial delay and stalling conditions, see Table 3.18. However, for both studies *Decision 1* and *Decision 2* the correlations do not exist for all conditions, which makes it difficult to get definite results. To derive final conclusions, more research is needed.

[14] Regarding study *Decision 3*, there were no significant correlations

		YouTube Frequency	TV satisfaction
init. delay [sec.]	2		0.28** (n.s.)
	4	0.23* (n.s.)	
	8		
	16		
stalling 2x [sec.]	1		n.s. (0.26*)
	2	-0.34* (n.s.)	0.35* (n.s.)
	4		
	8		

Table 3.18: Study Decision 1: Pearson's linear correlation coefficients between $DiffRatings$ and user variables. Values in brackets indicate Kendall's rank correlation coefficient, * = 0.05 significance level, ** = 0.01 significance level, n.s. = no significant correlation.

		laptop usage	tablet usage	Youtube usage on tablet	email usage on smartphone
DL BW	low				
	mid	0.34* (n.s.)		n.s. (0.55*)	0.41* (n.s.)
	high		0.64* (n.s.)		
	avg.		0.62* (n.s.)		

Table 3.19: Study Decision 2: Pearson's linear correlation coefficients between $DiffRatings$ and user variables. Values in brackets indicate Kendall's rank correlation coefficient, * = 0.05 significance level, ** = 0.01 significance level, n.s. = no significant correlation.

3.2.4 Conclusions

For low QoS Web situations (i.e. accessing Google Maps with a downlink bandwidth of 256 Kbit/s) the choice regarding the used end device had a *small, positive and significant impact* on the subjective quality ratings. It seems that users were more tolerant regarding bad network conditions if at least the used device could be individually chosen. These findings are in line with the related work discussed in Section 3.2.1, i.e., the presence of choice generally has a positive impact. A not yet verified explanation might be that the annoyance of low downlink bandwidth values can be compensated by making an individual decision.

In contrast to this positive impact, if the users were able to individually select a certain quality parameter — in study Decision 1 the temporal impairment regarding video streaming — the individual user decision for initial delays had a *significant, negative impact* on subjective quality ratings. This is opposed to related work which indicates a general positive impact of user decisions. As shown in [169] long, pre-selected initial delays in video stream scenarios only have a minor impact on disturbance, whereas even small stalling events during video playback lead to significant lower quality ratings. The presence of user choice only affects the more tolerated initial delays, whereas stalling events are not affected.

However, the individual selection of web content (study *Decision 3*) does not seem to significantly affect the subjective quality assessment.

From a practical study execution perspective, the possibility of user decisions, and hence the lack of predictability of the final choices, requires test designs that take this challenge into account (see also Section 3.1.5). For example, in study *Decision 2* only 6 of 36 users selected the laptop which impeded the generation of statistically valid results for this device category. Hence, one cannot expect a uniform distribution of cases over all options at the end of the experiment. As a consequence, it is strictly recommended to limit the amount of available options in order to increase the number of ratings per category (which is also recommended by the authors of [167]).

3.3 Conclusions Regarding Individual, Economic/Non-Economic Decisions

The six empirical studies about economic/non-economic user decisions presented in this chapter of the thesis provided insights into how user behavior impacts quality assessment. So far, there is a positive impact of purchasing decisions on subjective quality perception if the user decides to spend money: if a free-of-charge option is selected, there is no difference between decision-impaired quality ratings and a situation without any user decisions. The findings also demonstrate that there is a correlation of price and quality, i.e., the price is used as a basis for quality, which results in larger $DiffMOS$ values for increasing prices.

It seems that the socio-psychological effect of cognitive dissonance could play a role in these settings, but so far no final conclusion could be drawn. It seems that a steadily increasing dissonance is reduced if the most expensive options are chosen (Requirement: lots of options are available, i.e., in the study *WTP 2* twenty options were offered). Regarding the relevance of pricing and QoE it is anticipated that more empirical work including individual user decisions (about price, quality, content, etc.) will be conduced in the future. Hence, in Section 3.1.5 recommendations were presented to provide guidance for setting up laboratory experiments including quality assessment and decisions. A positive impact of individual, non-economic user decisions also occurs if the participant is able to select the end device for Web browsing and a low QoS condition is applied (e.g. low downlink bandwidth). One could assume that the individual decision compensates for the annoying network behavior to some extent. Totally unexpected, also a negative impact of individual, non-economic user decisions was observed in study *Decision 1*.

One has to consider that the discussed results about individual user decisions were gained in laboratory settings which might have led to biased results. For example, the decision regarding a certain video quality might be a clear low involvement activity. However, in the context of a user trial this decision gains a lot of attention which could lead to a high involvement situation (see also Section 3.1.1.1). Additionally, in the described experiments a decision was mandatory, i.e., the users had to selected a certain video quality, they had to select a certain Web content for browsing, etc. In real life scenarios, it is also possible to avoid a decision or to interrupt the current task. Although it is common to provide real money as deposit in WTP user studies (see Section 3.1.1.3 some kind of artificiality cannot be denied. Also, most purchasing decision models (see Section 3.1.1.1) presuppose a certain amount of competitors enabling economic competition, which was not implemented in the described user studies.

To overcome the stated shortcomings to a certain extent, a field study could be a reasonable approach in order to conquer the artificial limitations of laboratory experiments. For example, the usage and purchasing behavior of VoD-customers could be tracked and additionally some parameters could be modified over time, e.g., the price for enhanced video quality. In this scenario, the participants would be forced to use their own money and their behavior would be less limited, e.g., it would be possible for them to consume a video whenever it suits them.

4 Conclusions & Outlook

In this final chapter of the thesis an overall conclusion regarding the previously discussed user studies is provided. For this reason, the research questions and hypotheses — introduced in Section 1.2 — are evaluated. Furthermore, an outlook is provided, which bases on the results of this thesis and could lead to more insights regarding expectations, decisions and QoE.

4.1 Conclusions

In this thesis a first approach was presented to including user expectations systematically in QoE-related research. Hence, the first research question is:

RQ1: How relevant are expectations in QoE?

To investigate the role of expectations in the context of subjective QoE assessment and QoE modeling, a comprehensive literature study was conducted which revealed relevant aspects of expectations gained in various research fields, see Section 2.1.1. Furthermore, this information was used to generate an appropriate conceptual model which includes expectations in the process of quality assessment, see Section 2.1.2. This was achieved by incorporating desired and adequate expectations in the quality formation process introduced and discussed in [12].

Then, it was shown that explicit triggering of expectations — in the context of laboratory QoE assessment studies — is possible and reasonable if the applied setup is elaborate: expectation triggering via manual activation of a switch to change Internet connection types (wireline vs. wireless) in the experiments *Trigger 1* and *Trigger 2* (see Section 2.2.1.1) led to reasonable results, whereas the simple presentation of textual hints regarding the resolution of a video before the video was consumed (see Section 2.2.1.3) or asking the participants just to imagine applying a certain VoD contract while watching a video (see Section 2.2.1.3) did not lead to satisfying results. Hence, the first hypotheses H1.1 could only be validated for cases which include appropriate triggering mechanism.

H1.1: Explicitly communicated information about quality triggers user expectations, which leads to different quality ratings compared to experiments without explicit triggering.

In Section 2.2.3 it was discussed how quantified information about individual adequate and desired user expectations can be utilized to enhance the accuracy of MOS prediction models. Even with the rather simple approach of adding linear factors to a simple exponential fitting model it was shown that the accuracy can be enhanced up to 12%. Hence, it is reasonable to include the two short questionnaires regarding desired and adequate expectations (see Annex A) in empirical user studies to improve QoE models with little effort. Therefore, hypothesis H1.2 could be validated.

H1.2: Information about individual quality expectations can be utilized to significantly improve the prediction accuracy of QoE models.

Regarding the role of user expectations in the context of subjective QoE assessment and QoE modeling, the main output of this thesis is twofold:

First, it is possible to trigger expectations regarding technical quality with a subsequent impact on subjective quality assessment. Apparently, to be effective these trigger mechanisms have to be appropriate and need to be well included in the experimental study setup. This research finding is relevant for further QoE user studies in two different ways: (1) If the context of the usage of a certain service is relevant, these context factors need to be explicitly communicated to the test participants to trigger the correct expectations which influence the quality assessment. (2) One should avoid unintended triggering of expectations in experimental setups. Of course, as described before triggering of expectations in a laboratory setup is a complex task, so the chance of unintended triggering might be low. Nevertheless, if expectations are triggered, they might impact the assessment output.

Second, it was demonstrated that the gathered information about adequate and desired expectations via questionnaires can be utilized to enhance the prediction of quantitative QoE models. In contrast to established expectation related questionnaires like SERVQUAL (see Section 2.1.1), the questionnaires used in this thesis are short enough to include in QoE studies without spending too much time on this task.

The second part of the thesis covered the second research question dealing with the influence of individual economic/non-economic user decisions on subjective quality assessment:

RQ2: What influence do individual user decisions have on QoE assessment?

Section 3.1 showed that answering this research question requires fairly sophisticated experimental setups, i.e., three Willingness-to-pay user studies *WTP 1*, *WTP 2* and *WTP 3* were discussed (see Sections 3.1.2, 3.1.3 and 3.1.4) which demonstrated how methodological shortcomings were systematically overcome to finally achieve a first insight into the complex connection between economic decisions and subjective quality perception. Individual purchasing decisions in the context of Video-on-Demand led to more positive quality ratings with increasing prices if the user herself decided to spend money. Hence, hypothesis H2.1 could only be validated for conditions featuring individually selected and *charged* technical quality, and not for conditions featuring individually selected *free-of-charge* quality.

H2.1: Individual economic decisions regarding technical quality influence the subjective assessment of quality in a positive way.

The results of this thesis demonstrate that individual economic decisions regarding technical quality impact the QoE assessment. Hence, for QoE user studies in which the context of usage is relevant, this influencing factor is crucial and has to be considered. Consequently, VoD-related QoE experiments which neglect economic user decisions regarding the technical quality do not provide valid and generalizable results.

In Section 3.2 the influence of individual decisions regarding quality impairment, end device and content on subjective QoE assessment was discussed. A positive effect of individual user decisions was only confirmed for low QoS conditions and if the user selected the end device by herself. If the type of the temporal impairment of the video playback was chosen, the ratings were more negative compared to a situation in which no impairments were selected, which is not in line with related work. The individual selection of Web browsing content has no influence on subjective quality assessment. Hence, hypothesis H2.2 could not be validated.

H2.2: Including individual choices in subjective quality experiments leads to a significant positive difference in quality ratings compared to experiments without any user decisions.

So far, no final conclusions can be made regarding the impact of individual non-economic decisions in the context of subjective QoE assessment. The three conducted experiments *Decision 1*,*Decision 2* and *Decision 3* (see Section 3.2) demonstrated a negative, a positive and no impact of a particular user decision. So far, the individual selection of Web content seems to have no impact on subjective quality assessment, i.e., according to the discussed findings it is not necessary to provide different contents in Web QoE studies. On the other hand, the individual selection of the end device (smartphone, tablet and laptop) impacts the resulting quality rating, i.e., this influence needs to be considered in QoE experiments in which the participants are allowed to select the device. As shown in study Decision 1, individual user decisions can also lead to a negative impact on subjective quality assessment, i.e., one should be careful if individual user decisions are included in QoE experiments. So far, it cannot be ruled out that any individual user decision could have a negative, unintended impact on the quality rating.

4.2 Outlook

In this last section open research questions are addressed which emerged during the research process — but were out of scope of the thesis — and which also derive from the gained results.

4.2.1 Further Development of Expectation Questionnaires

The used questionnaires about adequate expectations were designed for certain situations, e.g., the evaluation of expected downlink bandwidth in a mobile Internet usage scenario. This approach requires some kind of abstraction by the involved users and for some use cases this approach might not be appropriate. For example, it is undefined which is the most appropriate quality feature regarding video quality ought to be included in the adequate expectations questionnaire. In the studies presented in this thesis the participants had to state expectations regarding video bit rate and video resolution. Nevertheless, less abstract features are required, e.g., regarding Web QoE it is reasonable to ask about indirect quality indicators like page load times or file download times. One promising approach could be to provide examples of different video qualities and the user has to select an adequate quality.

In Section 2.2.2 the assessment of desired and adequate expectations via quantitative questionnaires is described. According to the presented findings indirect questioning via ranking questions regarding desired expectations leads to reasonable results. However, also other methods like directly asking about desired expectations should be considered in future work.

Additionally, it might be interesting to extend the adequate expectation questionnaire with questions regarding the specific (e.g. downlink bandwidth, video resolution, etc.) *barely accepted* quality and the *wished* quality instead of just asking about a single value, which might be useful for advancing quantitative QoE models.

4.2.2 Advancement of Quantitative QoE Models

In the quantitative QoE models described in this thesis, data from the adequate and the desired expectations questionnaire was used separately, i.e., independent additional factors representing these two kinds of expectations were added to the model. Another approach could be to use the ranking information of the desired expectation questionnaire to weight the information of the adequate expectation questionnaire. For example, if a high technical quality is generally rather unimportant for a user, the specific adequate expectation regarding the technical quality of a service should also be less relevant in the quantitative QoE model. Hence, combinations of these kinds of information should be evaluated.

Even the plain approach of utilizing linear fitting models to demonstrate how expectation information can be used to enhance MOS prediction accuracy of QoE-models led to promising results. Nevertheless, also other methods like machine learning, decision trees, neural networks, etc. could be applied and evaluated.

4.2.3 Overfulfillment of Expectations

Table 4.1 shows that most of the users (50 to 65%) in the study *Trigger 2* stated expectations regarding quality that were higher than the finally accepted quality levels in the related empirical user trial. So far, it is unclear why expectations are generally stated higher than observed expectation fulfillment. One reason might be that users are generally not able to give precise information about their expectations, or the design of the used questionnaires was not optimal. However, more research is needed regarding the relationship between stated expectations and their observed fulfillment.

expectations stated via questionnaires are ... than actual quality ratings	fixed		mobile	
	speed [Mbit/s]	DL duration [sec.]	speed [Mbit/s]	DL duration [sec.]
lower	11%	11%	11%	11%
equal	36%	23%	38%	26%
higher	52%	65%	50%	61%

Table 4.1: Study *Trigger 2*: Percentual amount of users who describe their expectations regarding downloading a 50 MB file via a questionnaire in relation to real acceptance ratings (Binary yes/no question:"Would you accept this quality at home?"); DL=Download.

4.2.4 Emergence and Development of Expectations

The results of study *Trigger 1* and *Trigger 2* show that expectations adapt over time (cf. Figure 2.9). So far, only little research has been done to investigate which factors influence the modification of expectations over time. One could assume that adapted usage behavior (e.g. changing from a 3G to a LTE Internet connection on the smartphone) and media consumption (e.g. advertisements, which has been demonstrated by [41]) shape expectations, but so far strong empirical evidence is missing.

According to the related work discussed in Section 2.1, desired expectations are rather stable and more or less independent from context. Nevertheless, technology progress and changes in the general way of life might also influence this kind of expectations to some extent. For example, a couple of years ago SD video quality (576p) was accepted as the standard video quality in the context of TV consumption. Nowadays, HD video quality (720p or 1080p) is seen as the standard quality in this context and SD quality is perceived as degradation [170]. One can assume that this impacts the viewer's expectations. Nevertheless, this long-term impact has not been investigated in QoE-related research so far.

4.2.5 Indirect Measurement of Expectations

Collecting information about individual expectations via questionnaires might be a suitable approach for laboratory experiments with a small number of participants. For large scale applications — e.g. all customers of a telecommunication provider — this method might be inappropriate. Hence, other approaches of collecting information about user expectations should be evaluated. For example, information about expectations could be derived to some extent from existing socio-economical data, i.e., the combination of data of sex, age, usage habits, etc.

4.2.6 The Reason Behind the Economic Impact in QoE Studies

The main reason behind the positive $DiffMOS$ values occurring in the study *WTP 3* (see Section 3.1.4) is still not fully understood: the more a user pays, the bigger the video quality rating difference is[1]. According to the presented findings, cognitive dissonance might not be the main reason behind it, see Figure 3.24(a). Hence, additional studies, e.g., large field trials, might be helpful to get more insights.

The results of the study *WTP 2* (see Section 3.1.3) demonstrated that if a large amount of charged quality classes is available, the lowest (=free of charge) and the highest (=most expensive) quality class are selected unexpectedly often, see Figure 3.12(b). Additionally, the resulting $DiffMOS$ values seem to be significantly positive, i.e., for these two special cases there is a strong positive impact of economic user decisions on subjective quality assessment. Nevertheless, the low amount of ratings per quality class makes it almost impossible to draw a final conclusion here. Therefore, future work is needed to get more valid results regarding the influence of a large amount of selectable, charged quality classes in the context of subjective quality assessment.

4.2.7 Methodological Challenges of how to do WTP Studies

Laboratory experiments including user decisions are costly and sophisticated, as has been shown in this thesis. Hence, less elaborate study setups including economic decisions should be evaluated. For example, in the related-work Section 3.1.1.3 the constant-sum-technique [105] was discussed which could be a promising approach in the context of QoE laboratory studies. Instead of paying real money for a certain technical quality, the test participant has to find a balance between several options, e.g., better quality vs. more content to choose from vs. shorter test duration. For example, the amount of available video content is reduced in compensation for higher technical video quality. Nevertheless, also this approach leads to methodological challenges. Hence, the most appropriate way of getting highly valid information about economic-decision impaired QoE might be field tests. Here, the test participants could use their habitual VoD service at home while spending their own money — i.e. no handed out deposit — for consuming video content in the preferred charged quality.

4.2.8 Negative Impact of Individual Decision on QoE Assessment

The unexpected significant negative $DiffMOS$ values for study *Decision 1* (user decision about temporal impairment of videos, see Section 3.27) regarding initial delay raises

[1] Difference of quality assessment without economic decision vs. quality assessment impaired by economic decisions

many questions. So far, it is unclear why the selection of initial delay led to this negative quality assessment and why, on the other hand, the selection of stalling events had no negative impact. Also, this negative impact of individual decisions is not in line with related studies, which postulate a positive impact of decisions on test subjects (note: amount of choices was not too large).

4.2.9 Interaction Between User Expectations, Decisions and QoE Assessment

In this thesis, the one-sided impact of decisions on subjective quality assessment was empirically evaluated via the user studies *WTP 1-3* and *Decision 1-3*. Yet, as depicted in the model in Figure 1.3, this connection is not one-sided, i.e., there is an interplay between user decisions, expectations and quality assessment. Hence, these complex mechanisms need to be analyzed empirically in a more comprehensive way. In this thesis, the quality rating is the dependent variable and the user decision is the independent variable, i.e., the reversed case needs to be examined. This is aggravated by the fact that subjective quality assessments are not always explicit, i.e., it is not mandatory that the output of "comparison & judgment" in the quality perception process in Figure 1.3 automatically triggers an explicit "quality rating" via "encoding". Obviously, it is rather difficult to quantify quality assessments which are implicit, but this is necessary to fully understand the interplay between decisions, expectations and quality assessment.

4.2.10 Improvement of QoE Management by Including User Decisions

The results of study *Decision 2* (individual decisions about used end devices, see Section 3.2.3.2) could be used to evaluate novel QoE modeling and management approaches. For example, if the user is able to select her end device, the provided downlink throughput can be reduced and a certain level of user satisfaction might still be reached. Hence, the impact of network congestions and other network incidents on subjective quality assessment could be reduced. Additionally, as discussed in the work of [166], the mere *availability* of options increases the satisfaction level in general. Hence, regarding QoE management it might be sufficient to focus on the availability of options, i.e. most of the users might not be interested in actually *making* a decision, but the mere availability of options might evoke a positive effect on subjective quality assessment. To prove this assumption, more research is needed.

A Appendix: Questionnaires

The paper questionnaires were originally generated in German. Regarding further use and in the context of this thesis, translated versions are included in the appendix. If desired, the original questionnaires can be requested from the author of this thesis (mail address: andreas.sackl@gmx.at). Only the relevant parts of the questionnaires are depicted, i.e., test numbers, pagination, test user ID, etc. are omitted.

To what extent do you agree or disagree with the following statements:

	full applies	rather applies	neither nor	rather not applies	not applies
Mobile Internet access is standard.	☐	☐	☐	☐	☐
I expect 100% stability from my mobile Internet access.	☐	☐	☐	☐	☐
Mobile high-speed Internet access is very important	☐	☐	☐	☐	☐

	full applies	rather applies	neither nor	rather not applies	not applies
Fixed Internet access is standard.	☐	☐	☐	☐	☐
I expect 100% stability from my fixed Internet access.	☐	☐	☐	☐	☐
Fixed high-speed Internet access is very important	☐	☐	☐	☐	☐

Figure A.1: Study *Trigger 1* and *Trigger 2* - Questionnaire about the assessment of general expectations w.r.t. fixed and mobile internet access

.

What do you expect from your Internet connection? Please order the following features of an Internet connection regarding your personal importance. (1=most important, 2=second important, etc.)

___ High Speed
___ Low monthly fee
___ Short Contract duration
___ Good Support via E-Mail and telephone
___ Unlimited download volume

Figure A.2: Study *Measuring 1* - Questionnaire about desired expectations w.r.t. Internet access features

.

What do you expect from a Video on Demand Provider? Explanation: Via a Video on Demand Service it is possible to consume movies and TV Series over the Internet on your TV for a certain fee. The videos are rent and streamed, i.e., the videos are not bought.

What do you expect from a Video on Demand provider? Please order the following regarding your personal importance. (1=most important, 2=second important, etc.)

___ Large amount of available movies and TV series
___ Good support via E-Mail and telephone
___ Low monthly fee
___ High video quality
___ Short contract duration

Figure A.3: Study *Measuring 2*/Study *WTP 3* - Questionnaire about desired expectations w.r.t. Video on Demand providers

.

How fast should your **mobile Internet connection** be (data stick connected to your Laptop) while you browse a typical **Internet page** like maps.google.at or youtube.com?

☐ 1 MBit/s ☐ 4 MBit/s ☐ 14 MBit/s ☐ 30 MBit/s ☐ 45 MBit/s

The reply to this question was:

Very easy ☐ ☐ ☐ ☐ ☐ very difficult (=I could not answer this question)

How long should it maximally take to **download a 50 MB** file via your **mobile Internet connection** (data stick connected to your Laptop)?

☐ 10 seconds ☐ 30 seconds ☐ 1 minute ☐ 1 minute 30 seconds ☐ 2 minutes

The reply to this question was:

Very easy ☐ ☐ ☐ ☐ ☐ very difficult (=I could not answer this question)

How long should it take **maximally** to **download** an **Internet page completely** (at home via ADSL/Cable,etc.)?

☐ 2 seconds ☐ 3-7 seconds ☐ 8-12 seconds ☐ 13-17 seconds ☐ more than 17 Seconds

The reply to this question was:

Very easy ☐ ☐ ☐ ☐ ☐ very difficult (=I could not answer this question)

How fast should your **Internet connection** be at **home** (ADSL, Cable, etc.) while you **download a 50 MB** file?

☐ 1 MBit/s ☐ 4 MBit/s ☐ 14 MBit/s ☐ 30 MBit/s ☐ 45 MBit/s

The reply to this question was:

Very easy ☐ ☐ ☐ ☐ ☐ very difficult (=I could not answer this question)

Figure A.4: *Study Measuring 3/Study WTP 3* - Questionnaire about adequate expectations w.r.t. mobile/fixed Internet connections

.

How fast should your **Internet connection** at home (ADSL, Cable, etc.) be while you browse a typical **Internet page** like maps.google.at or youtube.com?

☐ 1 MBit/s ☐ 4 MBit/s ☐ 14 MBit/s ☐ 30 MBit/s ☐ 45 MBit/s

The reply to this question was:

Very easy ☐ ☐ ☐ ☐ ☐ very difficult (=I could not answer this question)

How long should it maximally take to **download a 50 MB** file via your **Internet connection at home** (ADSL, Cable, etc.)?

☐ 10 seconds ☐30 seconds ☐ 1 minute ☐ 1 minute 30 seconds ☐ 2 minutes

The reply to this question was:

Very easy ☐ ☐ ☐ ☐ ☐ very difficult (=I could not answer this question)

How long should it take **maximally** to **download** an **Internet page completely** with your mobile Internet connection (data stick connected to your Laptop)?

☐ 2 seconds ☐ 3-7 seconds ☐ 8-12 seconds ☐ 13-17 seconds ☐ more than 18 seconds

The reply to this question was:

Very easy ☐ ☐ ☐ ☐ ☐ very difficult (=I could not answer this question)

How fast should your **mobile Internet connection** (data stick connected to your Laptop) be while you **download a 50 MB** file?

☐ 1 MBit/s ☐ 4 MBit/s ☐ 14 MBit/s ☐ 30 MBit/s ☐ 45 MBit/s

The reply to this question was:

Very easy ☐ ☐ ☐ ☐ ☐ very difficult (=I could not answer this question)

Figure A.5: *Study Measuring 3/Study WTP 3* - Questionnaire about adequate expectations w.r.t. mobile/fixed Internet connections

.

How did you feel after you selected the video quality?

Please answer the following questions by checking **one** item per question.

1 totally not applies	**2** rather not applies	**3** neither nor	**4** rather applies	**5** totally applies

After I have selected the video quality with the corresponding costs ...

01.	I felt desperate	1	2	3	4	5
02.	I felt irritated	1	2	3	4	5
03.	I felt disappointed by myself	1	2	3	4	5
04.	I felt anxious	1	2	3	4	5
05.	I felt empty	1	2	3	4	5
06.	I felt angry	1	2	3	4	5
07.	I felt uncomfortable	1	2	3	4	5
08.	I felt disillusioned in myself	1	2	3	4	5
09.	I felt upset	1	2	3	4	5
10.	I felt frustrated	1	2	3	4	5
11.	I felt pain	1	2	3	4	5
12.	I felt depressed	1	2	3	4	5
13.	I felt furious about myself	1	2	3	4	5
14.	I felt unwell	1	2	3	4	5
15.	I felt agonized	1	2	3	4	5
16.	I asked myself if I really need this video quality	1	2	3	4	5
17.	I asked myself if it was necessary to spend money	1	2	3	4	5
18.	I asked myself if I selected the correct video quality	1	2	3	4	5
19.	I asked myself if it was properly to buy the selected video quality	1	2	3	4	5
20.	I asked myself if I was deluded after buying the selected video quality	1	2	3	4	5
21.	I asked myself if I was persuaded into buying the video quality	1	2	3	4	5
22.	I asked myself if the purchasing of the video quality was some kind of foul	1	2	3	4	5

How painful was the **financial decisions** to by the selected video quality?

[1] very painful [2] rather painful [3] neither nor [4] rather not painful [5] not painful

Figure A.6: Studies *WTP 2* and *WTP 3* - Questionnaire about cognitive dissonance.
.

Bibliography

[1] Andreas Sackl, Sebastian Egger, Patrick Zwickl, and Peter Reichl. The QoE Alchemy: Turning Quality into Money. Experiences with a Refined Methodology for the Evaluation of Willingness-to-pay for Service Quality. In *Proc. QoMEX (Quality of the Multimedia Experience) 2012, Yarra Valley, Australia*. IEEE, July 2012.

[2] Andreas Sackl, Kathrin Masuch, Sebastian Egger, and Raimund Schatz. Wireless vs. wireline shootout: How user expectations influence quality of experience. In *Proc. QoMEX (Quality of the Multimedia Experience) 2012, Yarra Valley, Australia*. IEEE, July 2012.

[3] Andreas Sackl, Patrick Zwickl, Sebastian Egger, and Peter Reichl. The role of cognitive dissonance for QoE evaluation of multimedia services. In *Proceedings of QoEMC2012, Anaheim, California*, 2012.

[4] Andreas Sackl, Patrick Zwickl, and Peter Reichl. QoE alchemy 2.0: An improved test setup for the pecuniary bias of QoE. In *Proc. The fifth international workshop on Quality of Multimedia Experience (QoMEX)*, 2013.

[5] Andreas Sackl and Raimund Schatz. Evaluating the impact of expectations on end-user quality perception. In *Proc. The Fourth International Workshop on Perceptual Quality of Systems (PQS 2013), Vienna, Austria*, Sep 2013.

[6] Andreas Sackl, Patrick Zwickl, and Peter Reichl. The trouble with choice: An empirical study to investigate the influence of charging strategies and content selection on QoE. In *Proc. ICQT'13, Zurich, Switzerland*, 2013.

[7] Andreas Sackl and Raimund Schatz. Got what you want? Modeling expectations to enhance web qoe prediction. In *Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEX) 2014, Singapore*, Sep 2014.

[8] Andreas Sackl and Raimund Schatz. Evaluating the influence of expectations, price and content selection on video quality perception. In *Proceedings of the International Workshop on Quality of Multimedia Experience (QoMEX) 2014, Singapore*, Sep 2014.

[9] Andreas Sackl and Raimund Schatz. The influence of user decisions on subjective quality assessment ratings. In *Proceedings of the IEEE Workshop on Quality of Experience-based Management for Future Internet Applications and Services (QoE-FI) - in conjunction with IEEE ICC 2015*, London, UK, Sep 2015.

[10] Andreas Sackl, Raimund Schatz, and Alexander Raake. Do you know what you want? Expectations and subjective quality perception. *International Journal of Human-Computer Studies*, [submitted].

[11] Valarie A. Zeithaml, Ananthanarayanan Parasuraman, and Leonard L. Berry. *Delivering service quality: balancing consumer perceptions and expectations*. The Free Press, New York, 1990.

[12] Alexander Raake and Sebastian Egger. *Quality of experience. Advanced concepts, applications and methods*, chapter Quality and Quality of Experience. Springer, Berlin–Heidelberg–New York NY, 2014.

[13] Leonard L. Berry, Valarie A. Zeithaml, and Ananthanarayanan Parasuraman. The nature and determinants of customer expectations of service. *Journal of the Academy of Marketing Science*, 21(1):1–12, 1993.

[14] Peter Reichl, Bruno Tuffin, and Raimund Schatz. Logarithmic laws in service quality perception: where microeconomics meets psychophysics and quality of experience. *Telecommunication Systems*, 52(2):587–600, 2013.

[15] Gerard Briscoe and Philippe De Wilde. Digital ecosystems: Evolving service-orientated architectures. In *Proceedings of the 1st International Conference on Bio Inspired Models of Network, Information and Computing Systems*, BIONETICS '06, New York, NY, USA, 2006. ACM.

[16] Frank Jackson. Epiphenomenal qualia. *The Philosophical Quarterly*, 32(127):127–136, 1982.

[17] Ute Jekosch. *Voice and Speech Quality Perception: Assessment and Evaluation*. Springer Berlin Heidelberg, 2006.

[18] International Telecommunication Union (ITU). Quality of telecommunication services: concepts, models, objectives and dependability planning – terms and definitions related to the quality of telecommunication services. *ITU-T Recommendation E.800*, 2008.

[19] International Organization for Standardization. Quality management systems - fundamentals and vocabulary. Technical report, ISO 9000, 2005.

[20] Harald Martens and Magni Martens. *Multivariate analysis of quality : an introduction*. Wiley, 2001.

[21] Sudip Misra, Snigdha Das, Manas Khatua, and Mohammad S. Obaidat. QoS-guaranteed bandwidth shifting and redistribution in mobile cloud environment. *EEE Transactions on Cloud Computing*, 2(2):181–193, April 2014.

[22] Sun Mifen and Zhang Xingming. Design of a QoS control scheme for real-time mobile video based on h.264. In *International Conference on Computer Science Service System (CSSS)*, pages 934–937, Aug 2012.

[23] Michael S. Borella, Debbie Swider, Suleyman Uludag, and Gregory B. Brewster. Internet packet loss: measurement and implications for end-to-end QoS. In *Proceedings of the 1998 ICPP Workshops on Architectural and OS Support for Multimedia Applications/Flexible Communication Systems/Wireless Networks and Mobile Computing*, pages 3–12, Aug 1998.

[24] Patrick Le Callet, Sebastian Möller, and Andrew Perkis eds. Qualinet white paper on definitions of quality of experience (2012). Technical report, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), 2012, Lausanne, Switzerland.

[25] Ulrich Reiter, Kjell Brunnström, Katrien De Moor, Mohamed-Chaker Larabi, Manuela Pereira, Antonio Pinheiro, Junyong You, and Andrej Zgank. *Quality of experience. Advanced concepts, applications and methods*, chapter Factors Influencing Quality of Experience. Springer, Berlin–Heidelberg–New York NY, 2014.

[26] Sebastian Egger, Raimund Schatz, Tobias Hoßfeld, and Waltraud Müllner. Perceptual events in a page view cycle. Technical report, Contributed to ITU-T , work group SG12, 2012.

[27] Sebastian Möller, Wai-Yip Chan, Nicolas Côté, Tiago H. Falk, Alexander Raake, and Marcel Wältermann. Speech quality estimation: Models and trends. In *IEEE Signal Process*, volume 28, pages 18–28, 2011.

[28] Shelley Buchinger, Simone Kriglstein, and Helmut Hlavacs. A comprehensive view on user studies: survey and open issues for mobile tv. In *EuroITV'09, Leuven, Belgium*, 2009.

[29] Tiia Sutinen and Timon Ojala. Case study in assessing subjective QoS of a mobile multimedia web service in a real multi-access network. In Herman de Meer and Nina Bhatti, editors, *Lecture Notes in Computer Science, Quality of Service – IWQoS 2005*, volume 3552, pages 298–312, 2005.

[30] International Telecommunication Union (ITU). Methodology for the subjective assessment of the quality of television pictures. *ITU-T Recommendation BT.500-7*, 2002.

[31] International Telecommunication Union (ITU). Methods for subjective determination of transmission quality. *ITU-T Recommendation P.800*, 2008.

[32] Sebastian Möller. *Assessment and prediction of speech quality in telecommunications.* Springer US, 2000.

[33] International Telecommunication Union (ITU). Subjective video quality assessment methods for multimedia applications. *ITU-T Recommendation P.910*, 2002.

[34] Frank Wilcoxon. Individual comparisons by ranking methods. In *Biometrics Bulletin*, volume 1, pages 80–83. International Biometric Society, 1945.

[35] Andy Field. *Discovering statistics using SPSS.* Sage, 2013.

[36] Sebastian Möller and Alexander Raake, editors. *Quality of experience: advanced concepts, applications and methods.* Springer, Berlin–Heidelberg–New York NY, 2014.

[37] Raimund Schatz, Tobias Hoßfeld, Lucjan Janowski, and Sebastian Egger. From packets to people: Quality of experience as a new measurement challenge. In Ernst Biersack, Christian Callegari, and Maja Matijasevic, editors, *Data Traffic Monitoring and Analysis*, volume 7754 of *Lecture Notes in Computer Science*, pages 219–263. Springer Berlin Heidelberg, 2013.

[38] Robert J. Sternberg. *Cognitive psychology.* Wadsworth Publishing, 6. edition, 2011.

[39] Richard J. Gerrig. *Psychology and life.* Pearson, 20. edition, 2012.

[40] Joseph A. Bonito, Judee K. Burgoon, and Bjorn Bengtsson. The role of expectations in human-computer interaction. In *Proceedings of the international ACM SIGGROUP conference on Supporting group work*, GROUP '99, pages 229–238, New York, NY, USA, 1999. ACM.

[41] Bronwyn Higgs, Michael Jay Polonsky, and Mary Hollick. Measuring expectations: forecast vs. ideal expectations. Does it really matter? *Journal of Retailing and Consumer Services*, 12(1):49–64, 2005.

[42] William Boulding, Ajay Kalra, Richard Staelin, and Valarie A. Zeithaml. A dynamic process model of service quality: from expectations to behavioral intentions. *Journal of Marketing Research*, 30(1):7–27, 1993.

[43] Francis Buttle. Servqual: review, critique, research agenda. *European Journal of Marketing*, 30(1):8–32, 1996.

[44] Richard L. Oliver. *Satisfaction: a behavioral perspective on the consumer.* Irwin-Mc Graw-Hill, 1996.

[45] Ananthanarayanan Parasuraman, Valarie A. Zeithaml, and Leonard L. Berry. SERVQUAL: a multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1):12–40, 1988.

[46] Richard T. Watson, Leyland F. Pitt, and C. Bruce Kavan. Measuring information systems service quality: lessons from two longitudinal case studies. *MIS Quarterly*, 22(1):61–79, 1998.

[47] Dan J. Kim, Donald L. Ferrin, and H. Raghav Rao. A study of the effect of consumer trust on consumer expectations and satisfaction: the korean experience. In *Proceedings of the 5th international conference on Electronic commerce*, ICEC '03, pages 310–315, New York, USA, 2003. ACM.

[48] Kai Kristensen, Anne Martensen, and Lars Gronholdt. Measuring the impact of buying behaviour on customer satisfaction. *Total Quality Management*, 10(4-5):602–614, 1999.

[49] Jukka Ojasalo. Managing customer expectations in professional services. *Managing Service Quality*, 11(3):200, 2001.

[50] Ernest R. Cadotte, Robert B. Woodruff, and Roger L. Jenkins. Expectations and norms in models of consumer satisfaction. *Journal of Marketing Research*, 24(3):305–314, 1987.

[51] Yen-Hao Hsieh and Soe-Tsyr Yuan. A quantitative approach to measure customer expectation for service innovation within service experience delivery. In *Proceedings of the 11th International Conference on Electronic Commerce*, ICEC '09, pages 38–46, New York, NY, USA, 2009. ACM.

[52] Albert Williams and Dixon E. Is this what you expected? The use of expectation measuers in usability testing. In *Proceedings of Usability Professionals Association Conference*, Scottsdale, AZ, 2003.

[53] Ann L. McGill and Dawn Iacobucci. The role of post-experience comparison standards in the evaluation of unfamiliar services. *Advances in Consumer Research*, 19:570–578, 1992.

[54] Miyuri Shirai and Robert Meyer. Learning and the cognitive algebra of price expectations. *Journal of Consumer Psychology*, 4:365–388, 1997.

[55] Martin O'Neil and Adrian Palmer. An exploratory study of the effects of experience on consumer perceptions of the service quality construct. *Managing Service Quality*, 13(3):187–196, 2003.

[56] Alexander Raake. *Speech quality of VoIP – assessment and prediction.* John Wiley & Sons Ltd, Chichester, West Sussex, UK, 2006.

[57] Dominik Strohmeier, Satu Jumisko-Pyykkö, and Kristina Kunze. Open profiling of quality: a mixed method approach to understanding multimodal quality perception. *Advances in Multimedia*, 2010(3):28, 2010.

[58] Marcel Wältermann, Alexander Raake, and Sebastian Möller. Quality dimensions of narrowband and wideband speech transmission. *Acta Acustica united with Acustica*, 96(6):1090–1103, 2010.

[59] Vlado Menkovski, Georgios Exarchakos, and Antonio Liotta. Online QoE prediction. In *Second International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 118–123, Trondheim, June 2010.

[60] Thomas Monath, Mario Kind, Tobias Heger, Maximilian Schlesinger, and José Aznar. Economical analysis of experience-optimized service delivery. In *9th Conference on Telecommunications Internet and Media Techno Economics (CTTE)*, Ghent, 2010.

[61] Davide Micheli, Alessandro Curto, Marina Mazzurco, Laura Pierucci, Romano Fantacci, and Renato Simoni. Handset and network quality performance benchmarking for QoE improvement. In *Globecom Workshops (GC Wkshps)*, Atlanta, GA, USA, 2013.

[62] Karan Mitra, Arkady Zaslavsky, and Christer Åhlund. A probabilistic context-aware approach for quality of experience measurement in pervasive systems. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, TaiChung, Taiwan, 2011.

[63] William L. Wilkie. *Consumer behaviour*. John Wiley and Sons, Ltd., 1994.

[64] Staelens Nicolas, Wendy Van den Broeck, Yohann Pitrey, Brecht Vermeulen, and Piet Demeester. Lessons learned during real-life QoE assessment. In *10th European conference on interactive TV and video (Euro ITV-2012)*, Ghent University, Department of information technology, 2012.

[65] Stéphane Péchard, Mathieu Carnec, Patrick Le Callet, and Dominique Barba. From SD to HD television: effects of h.264 distortions versus display size on quality of experience. In *IEEE International Conference on Image Processing*, pages 409–412, Atlanta, GA, USA, 2006. IEEE.

[66] Alhad Kuwadekar and Khalid Al-Begain. User centric quality of experience testing for video on demand over IMS. In *Proceedings of Computational Intelligence, Communication Systems and Networks*, Indore, 2009.

[67] Oliver Hohlfeld, Enric Pujol, Florin Ciucu, Anja Feldmann, and Paul Barford. A QoE perspective on sizing network buffers. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, Vancouver, BC, Canada, 2014.

[68] Karel Vandenbroucke, Denzil Ferreira, Jorge Goncalves, Vassilis Kostakos, and Katrien De Moor. Mobile cloud storage: a contextual experience. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices and services*, Toronto, ON, Canada, 2014.

[69] Pham-Thi Huong, Hoang-Van Hiep, Takumi Miyoshi, and Tatsuya Yamazaki. QoE-driven bandwidth allocation method based on user characteristics. In *Proceedings of the 16th Network Operations and Management Symposium (APNOMS)*, Taiwan, 2014.

[70] Bernadette Szajna and Richard W. Scamell. The effects of information system user expectations on their performance and perceptions. *MIS Quarterly*, 17(4):493–516, 1993.

[71] Ralf Plücker. Untersuchung des Einflusses der Qualitätserwartung und der Telefonier–Situation auf die Beurteilung von Telefonsprache. Study project report (unpublished). Institut für Kommunikationsakustik, Ruhr Universität Bochum, 1998.

[72] Joachim Riedel. Einfluß der Erwartungshaltung des Benutzers auf die Qualitätsbeurteilung von Telefonübertragungen an Internet–Terminals. Study project report (unpublished). Institut für Kommunikationsakustik, Ruhr–Universität, Deutschland Bochum, 1998.

[73] Carol A. Reeves and David A. Bednar. Defining quality: Alternatives and implications. *The Academy of Management Review*, 19(3):419–445, 1994.

[74] Mikko Hartikainen, Eas-Pekka Salonen, and Markuu Turunen. Subjective evaluation of spoken dialogue systems using servqual method. In *Interspeech 2004 - ICSLP, 8th International Conference on Spoken Language Processing, Jeju Island, Korea*, 2004.

[75] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. Anatomy of a crowdsourcing platform—using the example of microworkers.com. In *Workshop on future internet and next generation networks (FINGNet)*, Seoul, Korea, 2011.

[76] Huang Yajun, Zhou Wenan, and Du Yu. Research on the user behavior-based QoE evaluation method for http mobile streaming. In *Proceedings of the 10th International Conference on Broadband and Wireless Computing, Communication and Applications*, Krakow, Poland, 2014.

[77] Eduardo Cerqueira, Luis Veloso, Marillia Curado, Edmundo Monteiro, and Paulo Milheiro Mendes. Quality level control for multi-user sessions in future generation networks. In *IEEE GLOBECOM Global Telecommunications Conference*, New Orleans, LO, USA, 2008.

[78] Slawomir Zielinski, Francis Rumsey, and Søren Bech. On some biases encountered in modern audio quality listening tests – a review. *J. Audio Eng. Soc*, 56(6):427–451, 2008.

[79] Maarten Wijnants, Stijn Agten, Peter Quax, and Wim Lamotte. Investigating the relationship between QoS and QoE in a mixed desktop/handheld gaming setting. In *Co-Next Student Workshop '09: Proceedings of the 5th international student workshop on Emerging networking experiments and technologies*, Rome, Italy, 2009.

[80] Tobias Hoßfeld and Christian Keimel. *Quality of experience. Advanced concepts, applications and methods*, chapter Crowdsourcing in QoE Evaluation. Springer, Berlin–Heidelberg–New York NY, 2014.

[81] International Telecommunication Union. The E-model, a computational model for use in transmission planning. *ITU-T Recommendation G.107*, 2005.

[82] Andrea Bacioccola, Claudio Cicconetti, and Giovanni Stea. User-level performance evaluation of VoIP using ns-2. In *ValueTools '07 Proceedings of the 2nd international conference on Performance evaluation methodologies and tools*, Brussels, Belgium, 2007.

[83] International Telecommunication Union (ITU). Estimating end-to-end performance in IP networks for data applications. *ITU-T Recommendation G.1030*, 2005.

[84] Richard L. Oliver. Measurement and evaluation of satisfaction processes in retail settings. *Journal of Retailing*, 57(3):25, 1981.

[85] Silviu Apostu, Anas Al-Nuaimi, Eckehard Steinbach, Michael Fahrmair, Xiaohang Song, and Andreas Möller. Towards the design of an intuitive multi-view video navigation interface based on spatial information. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, Munich, Germany, 2013.

[86] Michael J. Munson and Shelby H. McIntyre. Developing practical procedures for the measurement of personal values in cross-cultural marketing. *Journal of Marketing Research*, 16(1):48–52, 1979.

[87] Milton Rokeach. *Beliefs, attitudes, and values: a theory of organization and change*. Jossey-Bass, San Francisco, 1968.

[88] Carroll JD. *Multidimensional scaling: theory and applications in the behavioral sciences*, volume 1, chapter Individual preferences and multidimensional scaling, pages 105–155. R. N. Shepard, A. K. Romney and S.B. Nerlove, 1972.

[89] Ville-Veikko Mattila. Perceptual analysis of speech quality in mobile communications (doctoral thesis), 2001.

[90] Ute Jekosch. *Voice and speech quality perception – assessment and evaluation signals and communication technology*. Springer, Berlin, 2005.

[91] Vlado Menkovski, Adetola Oredope, Antonio Liotta, and Antonio Cuadra Sánchez. Predicting quality of experience in multimedia streaming. In *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia*, MoMM '09, pages 52–59, New York, NY, USA, 2009. ACM.

[92] Elisangela Aguiar, André Riker, Antônio Abelém, Eduardo Cerqueira, and Mu Mu. Video quality estimator for wireless mesh networks. In *Proceedings of the 2012 IEEE 20th International Workshop on Quality of Service*, IWQoS '12, pages 1–9, Piscataway, NJ, USA, 2012. IEEE Press.

[93] Andreas Sackl, Sebastian Egger, and Raimund Schatz. Where's the music? Comparing the QoE impact of temporal impairments between music and video streaming. In *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 64–69, July 2013.

[94] Sebastian Egger, Peter Reichl, Tobias Hoßfeld, and Raimund Schatz. Time is bandwidth? Narrowing the gap between subjective time perception and quality of experience. In *2012 IEEE International Conference on Communications (ICC)*, pages 1325–1330, June 2012.

[95] Erich M. Kirchler. *Wirtschaftspsychologie*. Hogeref, 1999.

[96] Erik Angner. *A course in behavioral economics*. Macmillan Education, 2012.

[97] Alan Lewis. *The cambridge handbook of psychology and economic behaviour*. Cambridge University Press, 2012.

[98] Herbert A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, 1955.

[99] Daniel Kahneman. Maps of bounded rationality: psychology for behavioral economics. In *The American Economic Review*, volume 93, pages 1449–1475, 2003.

[100] Peer Soelberg. Unprogrammed decision making. *Academy of Management Proceedings*, 1966(1):3–16, 1966.

[101] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[102] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–91, 1979.

[103] Daniel Kahneman. *Thinking, fast and slow.* Penguin, 2012.

[104] Judith Lynne Zaichkowsky. Measuring the involvement construct. *Journal of Consumer Research*, 12(3):341–52, 1985.

[105] Werner Kroeber-Riel, Peter Weinberg, and Andrea Groeppel-Klein. *Konsumentverhalten.* Franz Vahlen Verlag, München, 9 edition, 2009.

[106] Darrel D. Muehling, Russell N. Laczniak, and J. Craig Andrews. Defining, operationalizing, and using involvement in advertising research: A review. *Journal of Current Issues & Research in Advertising*, 15(1):21–57, 1993.

[107] Christopher K. Hsee, Jiao Zhang, Fang Yu, and Yiheng Xi. Lay rationalism and inconsistency between predicted experience and decision. *Journal of Behavioral Decision Making*, 16(4):257–272, 2003.

[108] Christopher K. Hsee. Less is better: when low-value options are valued more highly than high-value options. *Journal of Behavioral Decision Making*, 11(2):107–121, 1998.

[109] James R. Bettman. Information processing models of consumer behavior. *Journal of Marketing Research*, 7:370–376, 1970.

[110] Klaus Peter Kaas. *Empirische Preisabsatzfunktionen bei Konsumgütern.* Springer Verlag Berlin, 1977.

[111] Thomas T. Nagle and Reed K. Holden. *The strategy and tactics of pricing: a guide to growing more profitably.* Pearson Prentice Hall, third edition, 2002.

[112] J. Douglas McConnell. Effect of pricing on perception of product quality. In *Journal of Applied Psychology*, volume 52, pages 331–334, 1968.

[113] Roben Peterson. The price-perceived quality relationship: experimental evidence. In *Journal of Marketing Research*, pages 525–528, 1970.

[114] Rose L. Johnson and James J. Kellaris. An exploratory study of price/perceived-quality relationships among consumer services. In *Advances in Consumer Research*, volume 15, page 316, 1988.

[115] V.K. Venkataraman. The price-quality relationship in an experimental setting. *Journal of Advertising Research*, 21(4):49–52, 1981.

[116] Frank R. Kardes, James J. Kellaris, Maria L. Cronley, and Steven S. Posavac. The role of selective information processing in price-quality inference. *Journal of Consumer Researchs*, 31, 2004.

[117] Pierre Chandon, Vicki G. Morwitz, and Werner J. Reinartz. Do intentions really predict behavior? Self-generated validity effects in survey research. *Journal of Marketing*, 69(2):1–14, 2005.

[118] Christoph Breidert, Michael Hahsler, and Thomas Reutterer. A review of methods for measuring willingness-to-pay. In *Innovative Marketing*, volume 2, pages 8–32, 2006.

[119] Michal Ries, Olivia Nemethova, and Markus Rupp. On the willingness to pay in relation to delivered quality of mobile video streaming. In *Proceedings of the International Conference on Consumer Electronics*. IEEE, 2008.

[120] Nadine Friedrich. Nutzung? Forschung! Ergebnisse einer repräsentativen Umfrage. In *TVienna*, 2008.

[121] Steve Duck. Diaries and logs. In Barbara M. Montgomery and Steve Duck, editors, *Studying interpersonal interaction*. New York: Guilford, 1991.

[122] Stefan E. Hormuth. The sampling of experiences in situ. *Journal of Personality*, 54(1):262–293, 1986.

[123] Christoph Breidert. *Estimation of willingness-to-pay. Theory, measurement, and application*. Doctoral thesis, WU Vienna University of Economics and Business, 2005.

[124] Alvin J. Silk and Glen L. Urban. Pre-test-market evaluation of new packaged goods: A model and measurement methodology. In *Journal of Marketing Research*, volume 15, pages 171–191, 1978.

[125] Michael Jarschel, Daniel Schlosser, Sven Scheuring, and Tobias Hoßfeld. An evaluation of QoE in cloud gaming based on subjective tests. In *Fifth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, 2011.

[126] David Hands. *FP5 Project M3I, IST–1999–11429, Deliverable 15/2 – M3I user experiment results*. 2002.

[127] Peter Reichl and Florian Hammer. Charging for quality-of-experience - a new paradigm for pricing IP-based services. In *Proceedings of 2nd ISCA Tutorial and Research Workshop on Perceptual Quality of Systems*, 2006.

[128] Ron Cocchi, Scott Shenker, Deborah Estrin, and Lixia Zhang. Pricing in computer networks: motivation, formulation and example. *IEEE/ACM Transactions on Networking*, 1(6):614–627, 1993.

[129] Jeffrey K. MacKie-Mason and Hal R. Varian. Pricing congestible network resources. *IEEE Journal on selected areas in communications*, 13(7), 1995.

[130] Patrick Maillé and Bruno Tuffin. Multi-bid auctions for bandwidth allocation in communication networks. In *IEEE Infocom, Hong Kong*, 2004.

[131] Antonis M. Hadjiantonis and Burkhard Stiller. *Telecommunication economics*, chapter Economics of Quality of Experience. Springer, 2012.

[132] Peter Reichl, Patrick Maillé, and Patrick Zwickl. On the fixpoint problem of QoE-based charging (invited paper). In *In Proc. The 6th International Conference on Performance Evaluation Methodologies and Tools*, 2012.

[133] Andrew Perkis, Peter Reichl, and Sergio Beker. *Quality of experience. Advanced concepts, applications and methods*, chapter Business Perspectives on Quality of Experience. Springer, Berlin–Heidelberg–New York NY, 2014.

[134] C. Whan Park. Joint decisions in home purchasing: a muddling-through process. *Journal of Consumer Research*, 9(2):151–162, 1982.

[135] Paul't Hart. Irving l. janis' victims of groupthink. *Political Psychology*, 12(2):247–278, 1991.

[136] Yoram Wind. Preference of relevant others and individual choice models. *Journal of Consumer Research*, 3(1):50–57, 1976.

[137] Erich M. Kirchler. *Kaufentscheidungen im privaten Haushalt. Eine Sozialpsychologische Analyse des Familienalltags.* Verlag für Psychologie Hogrefe, Goettingen, 1989.

[138] J.G. March and Zur Shapira. Behavioral decision theory and organizational decision theory. In Mary Zey, editor, *Decision Making: Alternatives to rational choice models*, CA, 1992. Newbury Park Sage Publications.

[139] Andreas Sackl, Patrick Zwickl, and Peter Reichl. From quality of experience to willingness to pay for interconnection service quality. In *Proceedings of Networking 2012*. Springer, 2012.

[140] Dapeng Wu, Yiewei Thomas Hou, and Ya-Qin Zhang. Transporting real-time video over the internet: challenges and approaches. In *Proceedings of the IEEE*, volume 88, pages 1855 – 1875, 2000.

[141] Nicolas Staelens, Stefaan Moens, Wendy Van den Broeck, Ilse Mariën, Brecht Vermeulen, Peter Lambert, Rik Van de Walle, and Piet Demeester. Assessing quality of experience of iptv and video on demand services in real-life environments. *IEEE Transactions on Broadcasting*, 56(4):458–466, Dec 2010.

[142] Peter Froehlich, Sebastian Egger, Raimund Schatz, Michael Mühlegger, Kathrin Masuch, and Bruno Gardlo. QoE in 10 seconds: are short video clip lengths sufficient for quality of experience assessment? In *Proceedings of the fourth International Workshop on Quality of Multimedia Experience QoMEX*, 2012.

[143] Leon Festinger. *A theory of cognitive dissonance.* Stanford University Press, Stanford, CA, 1957.

[144] Tim Breker. *The theory of cognitive dissonance and its application in marketing.* VDM Verlag Dr. Müller, Saarbrücken, Germany, 2009.

[145] Leon Mann. *Sozialpsychologie.* Beltz, 2001.

[146] Aaron C. Kay, Maria C. Jimenez, Joim T. Jost, Aaron C. Kay, Maria C. Jimenez, and John T. Jost. Sour grapes, sweet lemons, and the anticipatory rationalization of the status quo. *Personality and Social Psychology Bulletin*, 28(9), 2002.

[147] Daniel Ben-Horin. Dying to work: Occupational cynicism plagues chemical workers. In *These Times, June 27/July 3*, volume 24, 1979.

[148] Bruno Gardlo, Michal Ries, and Markus Rupp. A QoE evaluation methodology for HD video streaming using social networking. In *IEEE International Symposium on Multimedia, ISM*, 2011.

[149] Jillian C. Sweene, Douglas Hausknecht, and Geoffrey N. Soutar. Cognitive dissonance after purchase: a multidimensional scale. *Psychology and Marketing*, 17(5):369–385, 2000.

[150] Geoffrey N. Soutar and Jillian C. Sweene. Are there cognitive dissonance segments? *Australian Journal of Management*, 28(3), 2003.

[151] Robert B. Cialdini, Melanie R. Trost, and Jason T. Newsom. Preference for consistency: The development of a valid measure and the discovery of surprising behavioral implications. *Journal of Personality and Social Psychology*, 69(2), August 1995.

[152] Christoph Piesbergen and Nina Sarubin. Erfassung von kognitiver Dissonanz nach Kaufentscheidungen (http://nbn-resolving.de/urn /resolver.pl?urn=nbn:de:bvb:19-epub-12415-9). 2011.

[153] Patrick Zwickl, Andreas Sackl, and Peter Reichl. Market entrance, user interaction and willingness-to-pay: exploring fundamentals of QoE-based charging for VoD services. In *Globecom 2013 - Communications QoS, Reliability and Modelling Symposium, Atlanta, USA*, 2013.

[154] International Telecommunication Union (ITU). Parametric non-intrusive assessment of audiovisual media streaming quality. *ITU-T Recommendation P.1201.2*, 2012.

[155] Marie-Neige Garcia, Peter List, Savvas Argyropoulos, David Lindegren, Martin M. Pettersson, Bernhard Feiten, Jorgen Gustafsson, and Alexander Raake. Parametric model for audiovisual quality assessment in IPTV: ITU-T Rec. P.1201.2. In *15th IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pages 482–487, Sept 2013.

[156] Alexander Raake, Marie-Neige Garcia, Sebastian Moller, Jens Berger, Frederik Kling, Peter List, J. Johann, and Cornelius Heidemann. TV-model: parameter-based prediction of IPTV quality. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 1149–1152, March 2008.

[157] Stefan Winkler. *Digital video quality*. John Wiley and Sons, 2005.

[158] Mark L. Mitchell and Janina M. Jolley. *Research design explained*. Wadsworth Inc Fulfillment, Belmont, CA, USA, 2009.

[159] Harold Batteram, Gerard Damm, Amit Mukhopadhyay, Laurent Philippart, Rhodo Odysseos, and Carlos Urrutia-Valdes. Delivering quality of experience in multimedia networks. *Bell Labs Technical Journal*, 15(1):175–193, June 2010.

[160] Robin M. Williams. The nature of human values. *Political Science Quarterly*, 89(2):399–401, June 1974.

[161] Richard M. Ryan and Edward L. Deci. Intrinsic and extrinsic motivations: classic definitions and new directions. *Contemporary Educational Psychology*, 25(1):54 – 67, 2000.

[162] Ina Wechsung, Matthias Schulz, Klaus-Peter Engelbrecht, Julia Niemann, and Sebastian Möller. All users are (not) equal - the influence of user characteristics on perceived quality, modality choice and performance. In *Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop*, pages 175–186. Springer New York, 2011.

[163] Lauren A. Leotti, Sheena S. Iyengar, and Kevin N. Ochsner. Born to choose: The origins and value of the need for control. *Trends in cognitive sciences*, 14(10):457–463, Oct 2010.

[164] Nicola J. Bown, Daniel Read, and Barbara Summers. The lure of choice. *Journal of Behavioral Decision Making*, 16(4):297–308, 2003.

[165] Jack W. Brehm. Postdecision changes in the desirability of alternatives. *The Journal of Abnormal and Social Psychology*, 52(3):384–389, May 1956.

[166] Jane Ogden, Emma Daniells, and Julie Barnett. The value of choice: development of a new measurement tool. *The British Journal of General Practice*, 58(554):614–618, Sep 2008.

[167] Sheena S. Iyengar and Mark R. Lepper. When choice is demotivating: can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79(6):995–1006, December 2000.

[168] Miron Zuckerman, Joseph Porac, Drew Lathin, and Edward L. Deci. On the importance of self-determination for intrinsically-motivated behavior. *Personality and Social Psychology Bulletin*, 4(3):443–446, 1978.

[169] Andreas Sackl, Sebastian Egger, and Raimund Schatz. Where's the music? Comparing the QoE impact of temporal impairments between music and video streaming. In *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 64–69, July 2013.

[170] Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. Understanding the impact of video quality on user engagement. *SIGCOMM Comput. Commun. Rev.*, 41(4):362–373, August 2011.