MULTI-EPISODIC PERCEIVED QUALITY OF TELECOMMUNICATION SERVICES

vorgelegt von

Dennis Guse, M. Sc.
geb. in Berlin, Deutschland

an der Fakultät IV - Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuss

Vorsitzender	Prof. Dr. Axel Küpper
Gutachter	Prof. Dr.-Ing. Sebastian Möller
Gutachter	Prof. Dr. Judith Redi
Gutachter	Prof. Dr.-Ing. Ulrich Reiter

Tag der wissenschaftlichen Aussprache: 1. September 2016.

Berlin, 2016

Dennis Guse: *Multi-episodic Perceived Quality of Telecommunication Services*, © September 2016, Version: 4215706

Brain: Pinky, are you pondering what I'm pondering?
Pinky: I think so, Brain, but...
— Pinky & Brain

ABSTRACT

Telecommunication services have to cope with degradations resulting from the necessary transmission of data. A telecommunication service might thus not always be able to provide the same performance to a user. The resulting variation in perceived quality might affect the user's satisfaction, attitude, behavior, and also future-use intention towards a telecommunication service.

This thesis investigates the formation process of perceived quality across multiple, distinct interactions with one telecommunication service. The formation process of the so-called multi-episodic perceived quality is examined for two different time spans. Here, repeated-use in one session consisting of multiple usage episodes is investigated with an overall duration of up to 45 min. This is complemented by studying the formation process spanning several days.

This investigation was conducted by performing empirical experiments under controlled laboratory settings as well as field experiments. These experiments are based upon the *Mean Opinion Score (MOS)*, i. e., the assessment of the perceived quality of an (almost) identical stimulus/condition by multiple observers to derive the judgment of an *average observer*. The impact of individual user behavior was limited here by defining the task, content, and also time for each usage episode as well as the provided performance (defined-use method). The empirical data shows that applying the defined-use method is feasible and yields consistent results.

The results of the experiments show that more recent episodes have a higher impact on the multi-episodic perceived quality (recency effect). A saturation is observed for consecutive degraded episodes, i. e., the multi-episodic judgments remain on the same level *above* the episodic judgments of degraded episodes. In addition, a duration neglect is observed, i. e., a longer degraded episode does not have a higher negative impact on judgments of multi-episodic perceived quality.

With the empirical data, models for the prediction of multi-episodic judgments are evaluated. These models are based on the weighted average of the episodic judgments. The evaluation showed that a linear function outperforms a window function in regard to prediction accuracy and robustness.

ZUSAMMENFASSUNG

Da Datenübertragungen anfällig für Störungen sind, kann die wahrgenommene Qualität eines Telekommunikationsdienstleisters permanent Schwankungen unterliegen. Diese können die Zufriedenheit, die Meinung sowie das gegenwärtige und zukünftige Nutzungsverhalten ihres Anwenders beeinflussen.

In dieser Dissertation wird untersucht, wie sich die wahrgenommene Qualität bei wiederholter Nutzung eines Telekommunikationsdienstleisters zusammensetzt. Die sogenannte multi-episodisch wahrgenommene Qualität wurde für jeweils zwei Zeitspannen evaluiert. In einem kontrollierten Laborversuch wurden Probanden in einer Session von bis zu 45 Minuten in mehrere Nutzungsepisoden involviert. Auf Basis der gewonnenen Ergebnisse wurde eine Untersuchung der Nutzung über mehrere Tage konzipiert und als Feldversuch umgesetzt. Alle Experimente basieren auf dem *Mean Opinion Score (MOS)*, d. h. die wahrgenommene Qualität jedes einzelnen Nutzers auf einen annähernd gleichen Stimulus wird zu einem Durchschnittswert aller Nutzer zusammengefasst. Um den individuellen Einfluss des Nutzungsverhaltens auf die Bewertung möglichst gering zu halten, wurden Art, Dauer, Inhalt und Performanz jeder Nutzungsepisode streng definiert. Die empirischen Daten zeigen, dass die Methode der definierten Nutzung durchführbar ist und konsistente Ergebnisse liefert.

Die Ergebnisse der durchgeführten Experimente zeigen, dass die wahrgenommene Qualität zeitlich später erlebter Nutzungsepisoden einen größeren Einfluss auf die multi-episodisch wahrgenommene Qualität des gesamten Erlebnisses hat. Weiterhin wurde eine Sättigung für aufeinander folgende gestörte Nutzungsepisoden beobachtet: Obwohl mehrere gestörte Nutzungsepisoden hintereinander präsentiert wurden, sinkt die Bewertung der multi-episodisch wahrgenommenen Qualität nicht weiter. Auffällig ist, dass die multi-episodische Bewertung deutlich positiver ausfällt als die Bewertung der einzelnen gestörten Episoden. Unterschiede in der Dauer einer gestörten Nutzungsepisode zeigen hingegen keinen Einfluss auf die multi-episodisch wahrgenommene Qualität.

Auf Basis der empirischen Untersuchungen wurde ein Modell zur Vorhersage der multi-episodischen Bewertung konzipiert. Dieses beruht auf dem gewichteten Mittelwert der einzelnen Bewertungen aller bisher erlebten Episoden. Hierbei erweist sich eine lineare Gewichtung als genauer und stabiler als eine Fensterfunktion.

PUBLICATIONS

Parts of this PhD thesis have already appeared in the following publications:

Guse, Dennis and Möller, Sebastian (2013). "Macro-temporal Development of QoE: Impact of Varying Performance on QoE over Multiple Interactions." In: *Proceedings of AIA-DAGA Conference on Acoustics*. Vol. 46. Merano, Italy: Deutsche Gesellschaft für Akustik, pp. 452–455.

Guse, Dennis, Weiss, Benjamin, and Möller, Sebastian (2014). "Modelling multi-episodic quality perception for different telecommunication services: first insights." In: *Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*. Singapore. IEEE, pp. 105–110.

Weiss, Benjamin, Guse, Dennis, Möller, Sebastian, Raake, Alexander, Borowiak, Adam, and Reiter, Ulrich (2014). "Temporal development of quality of experience." In: *Quality of Experience*. Ed. by Sebastian Möller and Alexander Raake. Springer International Publishing, pp. 133–147. ISBN: 978-3-319-02681-7.

In Guse and Möller (2013) and Guse et al. (2014), the main work was done by myself. This includes designing the experiments, setting up required technical systems, conducting the experiments, analyzing the data, and writing the publications. Please note that in Guse et al. (2014) also experimental data from Möller et al. (2011), an experiment in which I was not involved, was used to develop and evaluate prediction models for multi-episodic judgments. My colleague Dr. Benjamin Weiss and my supervisor Prof. Dr.-Ing. Sebastian Möller were involved in discussing the experimental designs, conducting data analysis, interpreting the results, and designing potential prediction models.

In Weiss et al. (2014), I have written the section *10.5 Assessing Multi-Episodic QoE* (p. 142ff.). This work was possible due to the intensive discussions with Dr. Benjamin Weiss, Prof. Dr.-Ing. Sebastian Möller, and Prof. Dr.-Ing. Alexander Raake, who all provided valuable feedback.

ACKNOWLEDGMENTS

During the time working on this dissertation, I had the pleasure to meet, work, and getting to know a large number of awesome people.

I would like to thank my supervisor Prof. Dr-Ing. Sebastian Möller for providing me the opportunity to pursue my doctoral degree. Also, I would like to thank Prof. Dr. Judith Redi and Prof. Dr.-Ing. Ulrich Reiter for serving on my doctoral committee.

Also, I am grateful that I had the opportunity to work closely with Dr. Benjamin Weiss, who provided important feedback, near-endless discussions, and his steady support. The same holds for Prof. Dr-Ing. Alexander Raake.

My deepest thanks goes to my colleagues and friends Anna Wunderlich and Frank Haase. Without their impressive work, their clever ideas, and their precise preparation, the conducted experiments and thus this dissertation would have been impossible. It felt like pure luxury working with them. Thanks also to Henrique Orefice for the opportunity to build an awesome software project.

I would also like to thank my colleagues and friends Friedemann Köster, Falk Schiffner, Steffen Zander, Janto Skowronek, Dr. Hagen Wierstorf, Dr.-Ing. Sebastian Arndt, Johannes Rummel, Prof. Dr.-Ing. Jens Ahrens, and Dr. Oliver Hohlfeld for introducing me to interesting topics, starting fancy research projects, and, most important, cheering me up. Especially, the near-infinite number of table soccer games and the impressive niveau have been very important to me.

Last but not least, I am grateful to my girlfriend Regine Hähnel. Without her energetic, continuous support, I would not have been able to finish this dissertation. I am so glad, she endured all the stressful times and helped enthusiastically.

It was pleasure meeting you all and becoming part of my life.

Thanks.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AAC	Advanced Audio Coding
ACR	Absolute Category Rating
AoD	Audio-on-Demand
CoCR	Continuous Category Rating
ETSI	European Telecommunications Standards Institute
FEC	Forward Error Encoding
FPS	Frames per Second
HP	High Performance
ITU	International Telecommunications Union
LP	Low Performance
MOS	Mean Opinion Score
MP	Medium Performance
MP3	MPEG-1 Audio Layer III
QoE	Quality of Experience
QP	Quantification Parameter
PLC	Packet-loss Concealment
POLQA	Perceptual Objective Listening Quality Assessment
RMSD	Root Mean Square Deviation
RTP	Real-Time Transport Protocol
SCS	Short Conversation Scenario
SIP	Session Initiation Protocol
SSCQE	Single Stimulus Continuous Quality Evaluation
VoD	Video-on-Demand
VoIP	Voice-over-IP
UDP	User Datagram Protocol
USB	Universal Serial Bus
UX	User Experience

1

INTRODUCTION

1.1 MOTIVATION

In contrast to products, which can be manufactured and stored, a service is created on demand. A service is simultaneously produced by the *service provider* and consumed by the *customer*. A typical example are *telecommunication services*, such as speech telephony or Internet access. Telecommunication services provide communication between one or more parties in the form of data transmission. Here, a party can be a person or a computer. A service provider maintains the network infrastructure to provide telecommunication services and makes it available to potential users[1]. An actual service is created when a user interacts with a telecommunication service.

As a service is created when it is used, the provided *performance* might vary within a usage instance as well as between distinct usage instances. Performance is the "ability of a unit to provide the function it has been designed for" (Möller, 2005, p. 360). Performance can be determined by measurable parameters of a service, such as transmission bandwidth and one-way transmission delay (Möller, 2000, p. 12). In the case of a telecommunication service, varying performance might be the result of current network load conditions, network equipment failure, or related to end-user devices. The usage of telecommunication services can be divided into two aspects. First, a user can actively interact with a service, such as starting and engaging a telephone conversation. Second, a service can also be used passively, e. g., providing reachability of a telephone.

The active use of a telecommunication service results in a *perceived quality* of this interaction. It is assumed that the perceived quality results from a comparison between the *desired experience* and the *actual experience* (Raake and Egger, 2014, p. 13). Experience includes all perceptions resulting from this interaction (Raake and Egger, 2014, p. 13). The perceived quality of an interaction with a service is determined by the experienced performance, but also by individual factors, such as expectations, usage situation, etc. (e. g., Reiter et al., 2014, p. 55ff.).

Understanding the relationship between performance of a service and perceived quality has become an important field of research

1 Throughout this thesis, persons engaged in a service are called *users* of this service without distinction if they are the actual customers.

investigating the so-called *Quality of Experience (QoE)*. Starting from speech transmission for telephony (IEEE Audio and Electroacoustics Group, 1969) and its inherent impairments, QoE now mainly considers digital transmission of multimedia data. This includes the production, coding, transmission, decoding, and reproduction of text, voice, speech, audio, image, and video information. Specific characteristics of the human perceptual system have been used to develop enhanced coding mechanisms, such as *MPEG-1 Audio Layer III (MP3)* and the video codec H.264. These apply compression to provide a reduction in data rate while achieving only little to no reduction in perceived quality. Knowledge about QoE is also applied to planning and monitoring of the network transmission infrastructure for telecommunication services (Schatz et al., 2014). This enables telecommunication service providers to tailor their network infrastructure to the estimated need and on-the-fly control their infrastructure to avoid reductions in perceived quality for their users. For example, in the case of a video streaming service, stalling provoked by a temporary limitation in network bandwidth might be avoidable by reducing the video encoding bandwidth. Although perceived quality is likely to be affected, the service is still provided and remains useful for the user.

In general, a user can choose from a variety of service providers which provide similar services and select the one provider that suits his needs best. For a reasonable selection, a user must know his needs, requirements, financial constraints, and estimate the perceived quality, i. e., *assumed quality* (Raake and Egger, 2014, p. 13). Based on this knowledge, he can estimate if the usage of a service is likely to be satisfactory to him. After experiencing an interaction with a service, the user can then evaluate if this service fulfilled his needs including his perceived quality. He can then decide to use the service again or select a different service provider when he once again needs this type of service (Geerts et al., 2010).

1.2 RESEARCH QUESTION

Of special interest is the perception of varying performance over distinct interactions with one telecommunication service and the resulting perceived quality. Although the impact of varying performance during one usage instance has been investigated, it is so far not known how perceived quality evolves over several distinct instances, i. e., individual experiences. In fact, this is one factor that contributes to the *service quality* of a service (Berry et al., 1985; Zeithaml et al., 1996). The construct service quality maintains a holistic view from a business perspective of service usage, focusing on aspects such as customer loyalty (Parasuraman et al., 1985).

This thesis investigates perceived quality over several distinct and meaningful interactions with a service. Such an interaction is denoted

as a *usage episode* (for the definition see Section 2.2). The perceived quality of a usage episode is denoted as *episodic quality*. The perceived quality over several usage episodes is denoted as *multi-episodic perceived quality*.

This particular thesis addresses the following research question:

RESEARCH QUESTION: How does the multi-episodic perceived quality for one user evolve over several usage episodes with a single service?

So far, the formation process of perceived quality of distinct usage episodes into a multi-episodic perceived quality is not yet known. It could be a continuous process that integrates current experiences immediately into a current view or a retrospective assessment evaluating all memorized and recallable information (Hogarth and Einhorn, 1992). In fact, it might also be a mixture of both.

1.3 GOALS

In this thesis, I pursue two goals towards understanding the formation process of multi-episodic perceived quality. I focus solely on telecommunication services, because these are prone to variations in performance and usage episodes can be relatively short. As the formation process of multi-episodic perceived quality cannot be observed directly (cf. Chapter 2), this investigation relies on the judgments of multi-episodic perceived quality.

This leads to two goals:

GOAL 1: To investigate how the performance of a sequence of usage episodes determines judgments of multi-episodic perceived quality.

GOAL 2: To investigate how multi-episodic judgments can be predicted based on episodic judgments.

I conduct this investigation for two *usage periods*. Usage periods denote the time frame in which a user interacts repeatedly with a service. First, I investigate multi-episodic perceived quality in one session consisting of several usage episodes. The duration of one session is chosen here to be up to 45 min long, as this is a common duration for subjective experiments on perceived quality. Second, I investigate the formation process of multi-episodic perceived quality over a usage period of several days. These two usage periods provide an initial starting point for the investigation of the formation process of multi-episodic perceived quality. Considering two different usage periods is necessary because it is not yet known if the time between

usage episodes affects the formation process of multi-episodic perceived quality.

1.4 STRUCTURE

This thesis is structured as follows: In Chapter 2, I introduce concepts and fundamentals that form the basis for my investigation of multi-episodic perceived quality. It starts with an introduction to psychophysics and QoE, followed by an introduction into human memory and known biases for retrospective judgments. Following this, the state of the art on perceived quality under performance fluctuations is presented. Finally, the prior work on multi-episodic perceived quality is presented and discussed.

Subsequently, I present my work towards multi-episodic perceived quality. In Chapter 3, I describe the considered aspects of the defined-use method as well as the hypotheses that were investigated. These hypotheses form the basis for my investigations on multi-episodic perceived quality in one session (Chapter 4) as well as over several days (Chapter 5). Based on the conducted experiments, I evaluate potential models for the prediction of multi-episodic judgments based on prior episodic judgments (Chapter 6). Finally, I summarize my thesis, discuss the results, and present directions for future work in Chapter 7.

2

RELATED WORK

2.1 PERCEPTUAL QUALITY

2.1.1 *Perception and Psychophysics*

Humans use their senses, i. e., perceptual organs, to perceive events in their environment. Based on those events, an internal model is created and updated which incorporates knowledge about the environment. A *perceptual event* occurs within a human observer when a *physical event* stimulates a sensory organ (Blauert, 1996, p. 5). A physical event is an observable occurrence in time, location, and character (Le Callet et al., 2013). A perceptual event cannot be observed directly, as it occurs inside the observer due to perceptual and mental processing. It can, however, be described by the observer. The description process requires that a perceptual event can be observed consciously. In addition, a perceptual event can also be observed indirectly with behavioral measurements and physiological measurements. In fact, these measurements might be considered a special form of description. A psychometric function can be derived by measuring the properties of physical events and relating those to the descriptions.

A psychophysical experiment is conducted by presenting one or more physical events as a stimulus to one or more observers. Each individual observer derives the description of his perceptual event. A description can be expressed in a quantitative form by selecting the best fitting answer from a predefined set or in a qualitative form. For different observers, the description of the perceptual events resulting from a physical event does not necessarily have to be identical. In fact, each observer describes his individual perceptual event with regard to his concepts (Blauert, 1996, p. 11).

The perception of a physical event may even change the internal state of an observer (Raake and Egger, 2014). A physical event reaching a sensory organ can affect the sensitivity of this organ, or the observer might react to this physical event. Thus, the perception of following physical events might be affected by prior physical events. Also, the description of successive stimuli might be affected by the presentation order, since previous stimuli can be used as reference. Moreover, the *active* observation might also affect the perceptual process (Raake and Egger, 2014, p. 30). Here, an observer might focus

his attention on the perception to derive a description of a perceptual event. This might actually influence the perceptual process, i. e., it may lead to a different perceptual event and thus also affect the description.

2.1.2 *Perceptual Quality and Formation Process*

Perceptual quality is a branch of psychophysics focusing on the experience due to perception and the resulting quality of this experience.

Definition 2.1 (Experiencing) *"is the individual stream of perceptions (of feelings, sensory percepts and concepts) that occurs in a particular situation of reference." (Raake and Egger, 2014, p. 13)*

With regard to the quality of an experience and the underlying perceptual events, Jekosch (2005) formulates the *quality formation process* as an individual comparison process between the desired or expected outcome with the experienced outcome. The comparison of expectations and the actual experience results in a *quality event* inside the observer. It is assumed that a perceptual event is evaluated by a *comparing system* within the observer. This system derives the *quality features* of this event (cf. Jekosch, 2005, p. 17).[2]

Definition 2.2 (Quality Feature) *"is a recognized and designated characteristic of an entity that is relevant to the entity's quality." (Jekosch, 2005, p. 17)*

It is assumed that the evaluation of the difference between the *perceived quality features* and the *desired quality features* results in the experienced quality (Raake and Egger, 2014, p. 23). With regard to telecommunication services and multimedia systems, the term *perceived quality* has been extended to *Quality of Experience (QoE)*. In QoE, an observer is not regarded as an instrument of measurement, but as an *actor* striving for a *satisfying* perceived quality with regard to his expectations, requirements, and needs. In difference to an observer, who only describes an event, an actor can proactively make decisions and react to his environment.

Definition 2.3 (Quality of Experience) *"is the degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfillment of his or her expectations and needs with respect to the utility and / or enjoyment in the light of the person's context, personality and current state." (Raake and Egger, 2014, p. 19)*

2 Jekosch (2005) uses the term *entity* with regard to the quality formation process. As entity does not convey a temporal component, the term *event* is in this work used instead; following the notion of Blauert (1996).

Raake and Egger (2014) extended the *quality formation process* of Jekosch (2005). Here, both an anticipation process and an update process are added. These processes incorporate current experiences into expectations and might result in an adjustment of the desired quality features. Thus, the perceived quality of subsequent experiences might be affected. In addition, the concept of *assumed quality* is derived.

Definition 2.4 (Assumed Quality) *"corresponds to the quality and quality features that users, developers, manufacturers or service providers assume regarding a system, service or product that they intend to be using, or will be producing, without however grounding these assumptions on an explicit assessment of quality based on experiencing." (Raake and Egger, 2014, p. 17)*

Here, an experience has not yet taken place, but the quality formation process is based on expectations and prior knowledge. Based on *assumed quality*, a user might decide to initiate an interaction or rather avoid it.

The actual experience and resulting perceived quality is affected by influence factors.

Definition 2.5 (Influence Factor) *"Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user." (Reiter et al., 2014, p. 56)*

Contextual factors, which include task, user behavior, and usage situation (Reiter et al., 2014, p. 56), are important as these are in general not invariant. For example, high delay might not be noticed in a two-party telephone conversation if only one speaker is talking. In such a case, the delay is not perceived as degradation and thus would not negatively affect the quality formation process.

The formation process of perceived quality is still under investigation because it is not yet fully understood. In fact, assessment methods for perceived quality and the broader concept of QoE are still under development.

2.1.3 *Assessment Methods*

For the assessment of perceived quality, experiments are conducted that allow an actor (cf. Section 2.1.2) to use and thus experience a service or product. Information about the quality of an experience can be derived by monitoring the actor, observing his behavior, or requesting him to describe his experience.

For a quantitative description, scales with predefined labels are often used. An actor describes his experience by selecting the label that closest characterizes his perceived quality. The labels act as anchors, so multiple judgments of the same or different stimuli can be

related to each other. Examples of such scales are *Absolute Category Rating (ACR)* scales and *Continuous Category Rating (CoCR)* scales. In the latter case, judgments between two labels can also be expressed.

Based on the judgments, a *Mean Opinion Score (MOS)* can be derived by averaging all judgments per stimulus (ITU-T Recommendation P.800.2, 2013). The MOS is assumed to reflect the judgment of an *average actor* for this stimulus and the resulting experience. Judgments of an experience can either be taken while experiencing (these being denoted as *momentary judgments*) or after an experience. Taking momentary judgments might affect the perceptual, quality formation, and descriptive processes, as the assessment task must be conducted in parallel. However, momentary judgments allow the investigation of the impact of varying performance within a stimulus. In fact, this allows the investigation of the noticeability of varying performance. Assessing the perceived quality of an experience after this experience is denoted as *retrospective judgment* (Weiss et al., 2014). A retrospective judgment is based on recallable characteristics of an experience. It has been observed that not all parts of an experience affect a retrospective judgment of this experience in a similar manner (cf. Section 2.3).

It must be noted that a judgment cannot be considered absolute in terms of being universal, as it is the result of the quality formation process and description process within the actor. These processes are affected, for example, by differences in the so-called internal reference, which may lead to biases (Zielinski et al., 2008; Pitrey et al., 2011). For example, narrowband speech stimuli are judged to be better if no wideband stimuli are presented in the same experiment (Köster et al., 2015).

2.1.4 *Prediction of Perceived Quality*

Beyond understanding the underlying processes in detail, one major goal of research on QoE is the algorithmic prediction of judgments. This is especially important, as the evaluation by humans is an expensive procedure and thus limits the number of evaluations. In addition, continuous evaluation, as required for network monitoring, is not feasible in this manner.

An algorithmic predictor is called a *model*. A model maps the input, often a stimulus or an abstract reduced presentation of a stimulus to the expected judgment. The necessary information to create a model can be derived with subjective experiments, i. e., selecting *representative* stimuli and letting participants judge the perceived quality. A model can be applied for automatic evaluation, which is for example useful for the development of new coding algorithms. Particularly in the case of telecommunication services, models have been developed for planning purposes, such as the *E-Model* (ITU-T Recommendation G.107, 2015).

In fact, a model is inherently limited by the underlying data which lead to the selection of the model's parts and parameters. A model must be carefully applied to the implied restrictions resulting from the used data. Using it outside its designed scope may result in invalid predictions.

2.2 RETROSPECTIVE JUDGMENTS OF EXPERIENCES

For a retrospective judgment of an experience, one person can only rely on information available to him. While experiencing, a person needs to encode and memorize information in order to recall this information for a later judgment. As early as the perception stage, the amount of information is reduced, since only a subset of information can be encoded successfully into the perceptional memory and afterwards into the working memory (Raake, 2006b, p. 8f.). While these types of memory have only a very limited storage time, in the range of seconds to minutes, the information that becomes encoded into the long-term memory may remain accessible for several years. Memorized information decays, reducing the amount of *original* information even further. In addition, recall from long-term memory is not a perfect process. On recall, the original information is complemented by additional, potentially unrelated, information (cf. Schacter et al., 2003). In fact, unrelated, recalled information might even prevent the recall of original information (cf. Schacter et al., 2003). Here, prior knowledge and also prior experiences affect the actual recalled information.

Two aspects are important with regard to experiences and their retrospective judgments. First, how are experiences stored and in which manner is the information about individual experiences grouped together? Second, what effects or so-called biases determine retrospective judgments of an episodic experience? Based on the concept of episodic memory, this leads to the definition of the *usage episode* that forms the basis for my work on multi-episodic perceived quality.

2.2.1 *Episodic Memory*

Personal experiences and related information are stored in the *episodic memory* (Tulving, 1972). This memory stores items of information and their spatial-temporal relationship (Tulving, 1972, p. 385). The items are grouped together by specific events, or so-called episodes. These are encoded based on their autobiographical reference to pre-stored content (Tulving, 1972, p. 385f.). In addition, temporal and spatial information is stored. This includes information about the situation and feelings (Tulving, 1972, p. 385f.). An episode has an explicit start and end while retaining a temporal order of information (Conway and Pleydell-Pearce, 2000, p. 262). Based on recallable information about

an episode, this episode can be re-experienced and thus enable *mental* time travel. This is denoted as memory-vividness (Conway and Pleydell-Pearce, 2000). An episode is successfully memorized and retrieved if the perceptual properties and their approximate temporal relationship to other episodes can be described (Conway and Pleydell-Pearce, 2000).

The episodic segmentation process, i. e., constitution in the memory of an actor, is expected to happen during the actual experience (Ezzyat and Davachi, 2011; Kurby and Zacks, 2008). Here, the goal of an actor and the temporal proximity of individual events are considered of special importance (Black and Bower, 1979). It must be noted that first-time episodes and repeated episodes are considered different with regard to memorization (Conway and Pleydell-Pearce (2000) referencing Barsalou (1988)). The latter are linked by a shared theme, but tend to retain less specific information about the individual episodes (Robinson, 1992).

The characteristic of the episodic memory affects retrospective judgments due to the grouping of information and thus the ability to recall.

2.2.2 *Effects on Retrospective Judgments*

For a *retrospective judgment* of an experience, a person must rely on information about this experience that has been *a)* perceived and *b)* memorized, and can be *c)* recalled while conducting this judgment. With regard to retrospective judgments, major work has been done for the judgment of pain, showing several, even counterintuitive, effects. It has been found that not all parts of an experience affect a retrospective judgment equally. These effects are described in the following.

PRIMACY AND RECENCY

Primacy and recency have first been observed for sequential learning (Murdock Jr., 1962). In a sequential learning task, it has been found that the likelihood of recalling specific items afterwards depends on the position of an item. Items at the very beginning and very end have an increased likelihood of being recalled correctly. These effects are denoted as *primacy effect* and *recency effect*, respectively. The former denotes an increased likelihood of recalling items from the beginning and the latter the increased likelihood of recalling final items.

With regard to retrospective judgments of an individual experience, a recency effect is often observed. In an episode with varying pain, the retrospective judgment is lower (less painful) if less pain occurs at the end (Kahneman et al., 1993; Redelmeier and Kahneman, 1996). This could also be observed in cases where an episode was *extended*

by a less painful ending. An effect of primacy has not been observed for retrospective evaluation of pain, indicating that the beginning of such an experience has no increased importance for the retrospective judgment.

PEAK

For retrospective judgments of an experience, it has been observed that an outstanding part is overrepresented in the retrospective judgment. This is denoted as *peak effect*. Such an effect has been observed in retrospective assessment of pain (Kahneman et al., 1993; Redelmeier and Kahneman, 1996). Here, a spike in *momentary pain* results in a severely worse retrospective judgment. It seems as if exceptional parts of an experience can either be memorized better or are more likely to be recalled. Peak effects are most often considered for *negative* peaks, e. g., outstanding pain, but rarely for positive peaks, e. g., outstanding pleasure.

DURATION NEGLECT

In retrospective judgments, a duration neglect has been found. Here, it has been observed that the actual duration of an experience has only a reduced to no impact on a retrospective judgment. This has been mainly observed for retrospective judgments of pain (Fredrickson and Kahneman, 1993; Ariely, 1998).

This overview shows that retrospective judgments of experiences are affected by characteristics of the episodic memory. Here, items of information might not be recallable, not recalled precisely, or not considered with equal importance. The observed effects/biases indicate that not all parts of an experience are equally important for a retrospective judgment. Rather, the formation process of retrospective judgments seems to assign a special importance to individual parts. The observed biases have been used as the basis for investigating retrospective judgments of perceived quality, which is presented in the following.

2.3 PERCEIVED QUALITY AND MACROSCOPIC FLUCTUATIONS

The performance of telecommunication services is in general not constant, but rather varies over time. Performance fluctuations can occur due to varying network transmission, but also due to applied lossy compression. With regard to perceived quality, only those performance fluctuations must be considered that affect the quality formation process of a user.

Fluctuations that affect the quality formation process are distinguished as *microscopic* and *macroscopic* (Raake, 2006a, p. 72).[3] This differentiation focuses on performance fluctuations in one stimulus or episode. The former are fluctuations that are not perceived as variation in perceived quality. Such fluctuations are often rather short, such as non-bursty packet loss in a *Voice-over-IP (VoIP)* call (cf. Raake, 2006a, p. 72). In contrast, macroscopic fluctuations are perceived and judged as variation in perceived quality. An example of macroscopic fluctuations is a noticeable change in video encoding bandwidth.

The impact of *macroscopic* performance fluctuations on retrospective judgments of the perceived quality has already received some attention for telecommunication services. Although some approaches have been undertaken, the impact of *varying perceived quality* on a retrospective judgment and the prediction of such judgments are not yet completely solved. An overview on the state of the art is given in the following, starting with the assessment methods for varying *macroscopic* performance. Subsequently, an overview on observed effects is given, followed by a short presentation of modeling approaches of retrospective judgments.

2.3.1 *Quality Assessment of* Macroscopic *Fluctuations*

The perceived quality for *macroscopic* performance fluctuations can be assessed by requesting a user to judge the quality of this experience in retrospection. A retrospective judgment can be used to deduce the actual experience. However, especially in the case of longer experiences, a retrospective judgment may not contain all desired information about an experience. A final retrospective judgment can be complemented by *momentary* judgments and *intermediate* retrospective judgments.

For momentary judgments, the *current* perceived quality is assessed continuously during the experience. This allows the investigation of the noticeability of fluctuations, which might not be deducible from a retrospective judgment alone. This method is called *Single Stimulus Continuous Quality Evaluation (SSCQE)* and is standardized for video quality assessment (ITU-R Recommendation BT.500-13, 2012), but has also been applied for the evaluation of speech-only stimuli (e. g., Gros and Chateau, 2001). While experiencing, the momentary perceived quality should be judged by adjusting a slider. The position of the slider should reflect the current perceived quality. It has been observed that a reduction in performance almost instantaneously leads to a reduction in momentary judgments, but that adaptation due to improvements are delayed (e. g., Hands and Avons, 2001; Gros and

3 Raake (2006a) distinguishes microscopic and macroscopic with regard to packet-loss behavior for speech telephony. The notation used here is generalized to be independent of the actual source of fluctuations.

Chateau, 2001; Hamberg and de Ridder, 1999). Borowiak and Reiter (2013) extended the SSCQE method by not assessing momentary judgments. Instead, a participant is allowed to react to macroscopic fluctuations by adjusting the performance to the desired level.

The impact of macroscopic fluctuations can also be assessed by intermediate retrospective judgments. Here, a stimulus is split into individual parts. Each part is presented and a retrospective judgment, representing an intermediate judgment, is taken individually. The intermediate judgments allow a fine-grained analysis of the impact of the fluctuations.

With regard to the investigation of macroscopic fluctuation, the impact of varying user behavior is an issue. A MOS can only be derived using those judgments, which are based on identical or very similar stimuli and thus are assumed to lead to similar experiences. Because perception and experiences are influenced by an actor's behavior, varying usage behavior limits the applicability of the MOS. This can be overcome by either limiting the user behavior completely, i. e., permitting passive consumption and assessment only or by enforcing a certain behavior. The latter can be achieved by providing instructions to participants or letting them solve a task that can only be solved in a limited number of ways. For the evaluation of conversational speech telephony, for example, *Short Conversation Scenarios (SCSs)* have been developed. Here, the information that is to be exchanged is defined. Although a conversational structure is suggested, the exact timing is not enforced, and thus the assessment of macroscopic performance fluctuations is limited. An alternative is the method of *simulated conversations* (Weiss et al., 2009; Berger et al., 2008). This method enforces a *realistic* and reproducible user behavior including speaker changes. It is standardized as ETSI 102506 (ETSI, 2011). Here, a telephone conversation is split into individual parts of listening and speaking. Speaking parts and listening parts are then concatenated alternatingly in a meaningful order to create a simulated conversation. For speaking parts, predefined questions should be answered by the participant. This has been done orally as well as written. This should enable an otherwise passive listener to feel as though he is taking part in a real conversation. This method allows the presentation of similar stimuli, except for the exact speaking phases, to multiple participants including precisely timed degradations.

2.3.2 *Effects on Retrospective Judgments*

The impact of macroscopic performance fluctuations on retrospective judgments has been investigated mainly for video transmission and speech telephony. Here, similar effects to retrospective judgments of general experiences have been observed (cf. Section 2.2). Most often, a recency effect and in some cases a peak effect were observed. Du-

ration neglect has received only limited attention, but could be observed in some cases. A primacy effect has not been observed for perceived quality.

Although effects could be observed, this is not always the case, and the reasons for this are still under investigation. Fundamental work was conducted by Hands and Avons (2001). Their work is based on the *belief-adjustment model* (Hogarth and Einhorn, 1992). This model explains the occurrence of recency effect, primacy effect, and duration neglect for the integration of new information into one's belief. For 30 s video sequences, Hands and Avons (2001) could observe a recency effect as well as a duration neglect. A duration neglect was shown by presenting either 5 s or 10 s of reduced performance, but no impact on retrospective judgments was observed. This effect occurred although participants were able to assess the duration closely. A recency effect could be observed only if no intermediate judgments were taken. If momentary judgments were taken additionally, a recency effect could not be observed. This indicates that the momentary judgments affect the quality formation process due to the presence of the explicit assessment. Hamberg and de Ridder (1999) also observed both effects for video sequences of up to 180 s while varying impairment duration from 2 s to 10 s. With regard to shorter stimuli, effects on retrospective judgments are rarely observed. In fact, such effects seem to diminish if the length of an experience is reduced. For example, Ninassi et al. (2009) did not find a recency effect for 8 s videos.

Beside video transmission, the impact of macroscopic performance fluctuations has been investigated for (speech) telephony. Here, a recency effect could also be observed (e. g., Rosenbluth, 1998; Hamberg and de Ridder, 1999; Gros and Chateau, 2001; Gros et al., 2004; Belmudez, 2015; Weiss et al., 2009; Lewcio, 2014), whereas a negative peak effect has been less often observed (e. g., Weiss et al., 2009; Belmudez, 2015; Lewcio, 2014). The work of Weiss et al. (2009), Lewcio (2014), and Belmudez (2015) is based on the method of *simulated conversations* (ETSI, 2011). Here, the perceived quality of a simulated conversation is complemented by judgments of the individual listening parts. Analyzing the relationship between the intermediate judgments and the retrospective judgment, enables the investigation of potential effects. In addition, Rosenbluth (1998) investigated and observed a duration neglect for speech telephony.

With regard to macroscopic performance fluctuations and their impact on a retrospective judgment of the perceived quality, the state of the art is rather limited. Although recency effect, peak effect, and duration neglect have been observed, it is not known under which circumstances these occur. In fact, the characteristics of these effects are not yet fully understood, e. g., length of the recency effect. One reason for this is the incomparability of the conducted experiments.

Therefore, effects were observed repeatedly, but exact characteristics can hardly be derived.

2.3.3 *Prediction of Retrospective Judgments*

One practical goal of research on QoE is the prediction of retrospective judgments. Retrospective judgments can be predicted using either momentary or intermediate judgments as well as predictions of these judgments.[4] An alternative is to omit the prediction for these judgments and use a parametric description of the complete stimulus to directly predict the final retrospective judgment.

The *baseline model* for temporal integration is based on the assumption that no effects occur, i. e., that all individual parts of an experience are equally important. This can be represented by the *unweighted* arithmetic mean of all momentary or intermediate judgments. This model can be improved by accounting for observed effects that result in a deviation between the prediction and the judgment that is to be predicted. The baseline model can be extended by using a *weighted arithmetic mean* and a weight function. Here, a recency effect can be modeled by increasing the weight of later parts (Rosenbluth, 1998; Weiss et al., 2009; Hamberg and de Ridder, 1999). In a similar manner, a peak effect can be modeled. However, the implemented prediction models in the state of the art for retrospective judgments of single stimuli or single episodes are very specific to the experimental findings.

2.3.4 *Conclusion*

Retrospective judgments of perceived quality show effects similar to the retrospective judgments of experiences in general. However, the findings with regard to perceived quality remain so far inconclusive, as effects are regularly observed but rarely quantified. Here, the major focus lies on a *sufficient* precise prediction independent of the underlying reason. For example, a recency effect could be regularly observed, but it is not (yet) known under which circumstances it occurs, e. g., the minimal duration of an experience tending to show a recency effect. Furthermore, it is not known if recency is affected by the usage situation or modality (e. g., is visually presented content affected in a similar way to auditory content?) etc. In addition to a recency effect, a peak effect could be observed while a duration neglect only received limited attention.

In fact, research on QoE has been and will probably remain mainly technology-driven. In particular, the wide variety of applications, tech-

4 If the momentary or intermediate judgments are different to the final retrospective judgment, for example using a different scale or assessing something different, these judgments must be first transformed before predicting the retrospective judgment.

nology, and fast-pacing technological changes limit the comparison and derivation of knowledge about the formation process of judgments on perceived quality. Nevertheless, the state of the art shows that not all parts of an experience affect a retrospective judgment equally.

2.4 MULTI-EPISODIC PERCEIVED QUALITY

Services are in general used on a regular basis by a user (cf. Geerts et al., 2010). The experiences of a usage episode lead to a perceived quality in the user of this very episode. The definition of *usage episode* in the context of multi-episodic perceived quality is derived from the concept of episodic memory (cf. Section 2.2), focusing on usage of telecommunication systems. In addition, goal achievement as a requirement is added, following the concepts of utility and expected utility by Kahneman (2000).

Definition 2.6 (Usage Episode) *A usage episode is a distinct, meaningful, and self-contained interaction by a user with a service or system to achieve his goal(s).*

Multi-episodic perceived quality results from a *formation process*, combining prior experiences and their perceived quality. In fact, prior experiences can affect the *quality formation process* and thus influence the perceived quality of later experiences. Also, a user's behavior towards a service might change due to perceived quality affecting usage frequency, task-solving strategies, or even lead to abandoning the service completely. Multi-episodic perceived quality is thus the result of a sequential process. Here, the order of usage episodes and their individual experiences affect the final judgment. Investigations on multi-episodic perceived quality can therefore only be undertaken in experiments adhering to a between-subject design. Here, every participant is only exposed to a single *multi-episodic* condition.[5] In the following, the word *episode* is used as a synonym of usage episode.

In the field of QoE, multi-episodic perceived quality has so far only received limited attention. For practical use, it is often sufficient to understand the relationship between different performance parameters and the impact on perceived quality in a time frame ranging from seconds up to some minutes. Thus, an influence of time, tasks, and other factors on perceived quality is often neglected. However, understanding the formation process of multi-episodic perceived quality enables, for example, a service provider to serve a better service for his users.

5 A within-subject design might applicable for the investigation of multi-episodic perceived quality if it can be ensured that the presentation order of different *multi-episodic* conditions does not affect the multi-episodic judgment for each condition.

2.4.1 *Excursus: Multi-episodic usage in UX*

Effects of multi-episodic usage while focusing on changes in user behavior have been evaluated in *User Experience (UX)*. QoE and UX conceptually overlap as both focus on experiences in general and experiences with technology (cf. Wechsung and De Moor, 2014). UX, stemming from usability, focuses on the interaction with technology and how interactions affect usage as well as emotions towards used technology.[6] Multi-episodic evaluation is an important aspect of UX, because user behavior towards technology changes as a user learns how to use the technology and which tasks are well-suited to its use. The multi-episodic concept is described by Roto et al. (2011, p. 8). However, they failed to put it into context with prior work and omitted their definitions.

Two major experiments that assessed the use of products over a longer usage period have been conducted. Karapanos et al. (2009) investigated in one experiment how the expectations and behavior towards a new smart phone change in a usage period of four weeks. This is extended by Kujala et al. (2011) with the *UX curve* method. Here, a participant evaluates his experiences with a product or service in retrospect. The participant draws a line reflecting how his satisfaction changed over time and annotates the vertices with the reasons for the changes. This also includes recalled adaptations of usage behavior. Using this method, one experiment was conducted, evaluating the changes of satisfaction and user behavior with a new smart phone over a usage period of one year. The experiments of Karapanos et al. (2009) and Kujala et al. (2011) showed that the usage and emotions towards the product under investigation change over time. In the beginning, interactions are more playful and exploratory, but over time becoming more task-oriented and productive.

Although multi-episodic integration is not a key aspect for UX, it is an important aspect especially regarding the adaptation of new products and new services.

2.4.2 *Average Perceived Quality (Duncanson, 1969)*

Initial work in the direction of multi-episodic perceived quality for telecommunication services was performed by Duncanson (1969). Duncanson asked regular users of an oversea speech telephony service about their experience with the said service. He investigated if there is a difference between *a)* the *perceived quality of a just finished call* with average performance and *b)* the *assumed quality* of a call with average performance. For this experiment, Duncanson used a 4-point

6 For a longer discussion on similarities and difference between QoE and UX see Wechsung and De Moor (2014) and also Hassenzahl (2008).

ACR scale.[7] It could be shown that judgments for case a) yield a higher score than for case b). Duncanson concludes "that ratings of single, recent telephone calls yield results different from ratings of subjectively averaged, past telephone calls of the same type" (Duncanson, 1969, p. 116).

Although Duncanson studied episodic perceived quality and assumed quality of a usage episode with average performance, this experiment revealed an important aspect of multi-episodic perceived quality. The results indicate that the formation process of the assessed episodic perceived quality leads to a lower judgment than expected. This suggests that the formation process does not equally weight all prior experiences with said service, but rather focuses on experiences with low performance.

2.4.3 *Assessment of Multi-episodic Perceived Quality*

Prior work in the direction of multi-episodic evaluation focuses on use, experience, and perception of a service or product under realistic conditions. This is true for UX as well as as for initial work on multi-episodic perceived quality (i. e., Möller et al., 2011).

In the following section, first aspects are presented which are known to affect perceived quality and are thus important to consider for multi-episodic assessment. Subsequently, the two potential assessment methods for multi-episodic perceived quality are presented.

ASPECTS

Perceived quality for short stimuli as well as complete usage episodes is investigated by exposing multiple participants to the same stimulus or very similar stimuli and assessing the resulting perceived quality. The calculation of a MOS, independent of the type of judgment or scale, results in the description of the *average perceived quality*. With regard to multi-episodic perceived quality and also perceived quality in general, a MOS can only be calculated with judgments that result from the same condition or very similar conditions. In fact, this implicitly assumes that effects/biases are similarly for different participants with regard to the formation process of multi-perceived quality, i. e., the same effects with a similar characteristic occur for each participant. In experiments on perceived quality, this assumption of *temporal consistency* is in general neglected, as the time frames under investigation are rather small, i. e., usually in the range of seconds up to some minutes. For the investigation of multi-episodic perceived quality, this must be taken into account as a potential confounding factor. In fact, it is not yet known if the formation process is *universal*

7 Duncanson (1969) applied a 4-point ACR scale with the labels excellent (4), good (3), fair (2), and poor (1).

or affected by differences between individuals. In the following, factors are discussed that might affect multi-episodic perceived quality and thus should be kept constant to apply a MOS evaluation.

Perceived quality is influenced by *prior knowledge* about the service under consideration (cf. Section 2.1.2). This includes knowledge about the specific service and also about the type of service in general. For example, information on the reliability of telephony services as well as knowledge about potential degradations may set expectations. Also, promises about the performance by a service provider are considered as knowledge, e. g., performance promises about a service in an advertisement. The availability of such information might affect expectations, attribution of degradations, and *assumed quality*. Therefore, this might affect the formation process of multi-episodic perceived quality.

The quality formation process is also affected by the *task*, its *importance*, and the *task-solving strategy*. Depending on the actual tasks, some degradations might not even be noticeable. For example, high delay in a telephone conversation with rare turn-taking will result in a better perceived quality than in a conversation requiring more turn-taking (e. g., Egger et al., 2010; Schoenenberg, 2015). Degradations might also suggest an adaptation of the task-solving strategy, e. g., reduced turn-taking due to high delay in a telephone conversation (Schoenenberg, 2015). Also, the task-solving strategy might be adjusted over multiple usage episodes with the result that task-solving becomes more efficient. In such a case, latter episodes will require less time and thus possibly affect the multi-episodic quality formation process. Also, the importance of a task might affect the multi-episodic quality formation process. A higher importance of a task can increase the likelihood of memorizing and recalling information about this specific usage episode. This task and its episode might thus have a higher impact on multi-episodic perceived quality. For example, an interviewee in a telephone job interview is likely to recall this specific episode and describe his perceived quality in difference to a common call with a friend. Therefore, tasks, as well as the task-solving strategies and their importance, should be kept constant to avoid an undesired impact, i. e., not explainable, on multi-episodic perceived quality. In addition, tasks that yield a comparable task-solving duration should be selected.

In addition to prior knowledge and tasks, the actual *usage pattern* might also affect multi-episodic judgments. A usage pattern describes when, for what, and how a service is used. This includes how often a user is exposed to the service's performance and thus new experiences affecting the perceived quality are acquired. A usage pattern can be regular with a defined frequency but also irregular. In fact, different usage patterns might result in differences of multi-episodic judgments. This might be due to differences in memorization, but

also due to decay of memorized information. Besides the usage pattern, the experienced performance seems to be of greater importance. In terms of multi-episodic perceived quality, this refers to the performance and its variation of all individual episodes.

All these aspects might affect the formation process of multi-episodic perceived quality and thus might lead to different observable effects. However, so far, the potential impact of these aspects on multi-episodic judgments is not yet known.

ASSESSMENT METHODS

Multi-episodic perceived quality can be investigated using two diametric methods which differ in the behavioral freedom of users. On the one hand, users are allowed to interact freely with the service under investigation. In this setting, each user can decide individually when, how, how often, and for which tasks to use this service, i. e., the individual user behavior is not limited. On the other hand, a similar user behavior can be enforced by defining when, how, and for which tasks the service should be used. In the following, both methods are presented in detail, and the implications are discussed.

FREE-USE METHOD Multi-episodic perceived quality can be investigated by observing the interactions of a user with a service without restricting his usage behavior. Here, users are free to select when and how to use a service, to abandon it, and also to acquire new information about it. This approach has for example been taken by Duncanson (1969). In fact, this allows the observation of *real* users of a *real* service. In such a case, information about a user's usage pattern, tasks, goals, expectations, and prior knowledge is often hard to acquire. This lack of information might be overcome by sampling a large number of users. Allowing users to decide when and how a service is used enables the investigation of service usage in an ecologically valid situation, i. e., how users would behave in their current situations. This method offers ecological validity, but limits reproducibility.

DEFINED-USE METHOD Limiting the user's freedom by defining when and how to interact with a service allows to present the same condition to multiple participants. For multi-episodic perceived quality, this requires the definition of the usage pattern including task and performance for each usage episode. This ensures that all participants are exposed in a similar manner to the service. A task can be artificial, i. e., a participant does not necessarily engage in such a task on his own, and should lead to a similar task-solving strategy. For telecommunication services, this could be a telephone call with defined content or listening to a recorded telephone call.

To limit an influence of prior experiences and information about a service, it must be ensured that users have similar knowledge about the service and similar prior exposure to the service. This can be achieved by a careful selection of users or by creating a *new service*. In fact, a new service only prevents the existence of prior experiences with this service, but cannot avoid the impact of prior knowledge, for example, about typical degradations for such a service type in general. This assessment method is here denoted as defined-use method. Here, the usage pattern as well as the task and the performance per usage episode are defined. Such a set is denoted as *multi-episodic condition*. Exposing multiple participants to the same multi-episodic condition allows to derive a MOS. The impact of varying performance on multi-episodic perceived quality can then be investigated by changing only the performance between otherwise identical conditions. This method limits ecological validity, enabling however in the process reproducibility.

DISCUSSION Both assessment methods are diametrically opposed. Free-use allows studying multi-episodic perceived quality with existing users by observing their interaction with said service and optionally gathering direct feedback. This non-intrusive method is, however, in general limited by the obtainable information about a user. In addition, the performance of a *real* service can hardly be controlled and the desired degradations might not occur reliably enough. In fact, modifying the performance of a service intentionally might introduce ethical issues if users are not informed about this. This is not an issue for the defined-use method. Here, participants are instructed when and how to use a service including specific tasks. This limits ecological validity. In addition, this method requires that a service is available that can provide the desired performance in a precise and reliable manner. This means that it must be possible to introduce degradations when defined and prevent degradations when none are defined. This is a challenging task for telecommunication services in particular due to the necessary data transmission and often uncontrollable network performance. In addition to technological challenges in the deployment of such a service, a participant needs to solve each task at a specific point in time. This is not an issue if multi-episodic perceived quality is studied in one session, e. g., several consecutive usage episodes. Investigating multi-episodic perceived quality over longer time spans spanning several days, weeks, or months increases the required effort for a user to follow the defined usage pattern. Longer time spans also increase the likelihood that a defined usage pattern cannot be fulfilled by an individual participant, as each participant must integrate the usage pattern into his daily life. However, investigating multi-episodic perceived quality beyond one session is

necessary in order to derive knowledge about the influence of time between usage episodes.

2.4.4 *Application of the Defined-use Method (Möller et al., 2011)*

Initial work on multi-episodic perceived quality with the defined-use method was performed by Möller et al. (2011). In the following, first the experimental design and then the results are presented. This work forms the basis for the here presented experiments (see Chapter 3, Chapter 4, and Chapter 5).

DESIGN

In this experiment, pairs of two participants used a video telephony service (Skype) over a usage period of 12 days, i. e., each pair experienced one multi-episodic condition. Möller et al. (2011)[8] followed the premise that the service works in general, i. e., it provides the best achievable performance, but sometimes usage episodes are only presented with reduced performance. The service needed to be used twice per day, i. e., the first call between 6 h and 15 h and the second call between 15 h and midnight. For each of these 24 calls, one SCS needed to be solved (ITU-T Recommendation P.805, 2007). SCSs suggest a conversational structure to limit the behavioral freedom of caller and callee.[9] Furthermore, each pair of participants was allowed to use the service for private communication if desired.

After finishing a call, each participant rated the perceived quality of this usage episode (*episodic judgment*). After the 2nd defined call on the 2nd, 7th, and 12th day, the perceived quality of all experienced usage episodes *up to that point* was assessed (*multi-episodic judgment*). The episodic judgments and multi-episodic judgments were both taken on the 7-point CoCR scale (ITU-T Recommendation P.832, 2000, p. 18).[10] This scale is shown in Figure 2.1.

Figure 2.1: 7-point CoCR scale with German labels; labels from left-to-right: extremely bad (0), bad (1), poor (2), fair (3), good (4), excellent (5), and ideal (6) (ITU-T Recommendation P.851, 2003, p. 19) (own illustration).

8 It must be noted that Möller et al. (2011) uses the term *service quality* as synonym to multi-episodic perceived quality, rather than service quality in terms of Parasuraman et al. (1985).

9 A detailed explanation on SCS is given in Section 3.1.5.

10 For a conversion between the discrete 5-point ACR scale and the 7-point CoCR scale see Köster et al. (2015).

In this experiment, performance was defined per usage episodes by limiting the maximum transmission bandwidth symmetrically, i. e., for both participants, for audio and video. This is expected to avoid macroscopic fluctuations within a usage episode and thus avoid potential but still unknown effects on multi-episodic perceived quality. Three performance levels were applied: *High Performance (HP)*: 500 kbit/s, *Medium Performance (MP)*: 150 kbit/s, and *Low Performance (LP)*: 32 kbit/s.[11] Performance of the service was varied on a per day basis, i. e., all usage episodes in a day were presented with the same performance level. Möller et al. (2011) tested five multi-episodic conditions (see Table 2.1). Here, a condition defines the time for each usage episode, task, and also performance level. In each condition, one or two performance levels were used only.

It must be noted that condition 5 is different, because only MP and LP are applied, and thus MP represents the best performance for this condition. Depending on the prior experiences of the participants, this might affect quality judgments, as LP can only be compared to MP. The first two days are presented in each condition with the best performance level for this condition.

The experiment was conducted by the participants in their home environments using their own private computer and broadband Internet access. This limits the necessary effort for participants and allows them to integrate the experiment into their daily life. For the same reason, each pair of participants was required to know each other beforehand. In this experiment, participants were equipped with a headset and a webcam to avoid an impact of varying terminal equipment (Möller et al., 2011).

Table 2.1: Conditions applied by Möller et al. (2011) with performance level per day. Non-HP episodes are in **bold** (LP) and *italic* (MP).

Condition	Performance Level (per day)						
	1..2	3	4	5..8	9	10	11..12
1	HP	HP	HP	HP	HP	HP	HP
2	HP	**LP**	HP	HP	HP	HP	HP
3	HP	**LP**	HP	HP	**LP**	HP	HP
4	HP	**LP**	**LP**	HP	**LP**	**LP**	HP
5	*MP*	**LP**	*MP*	*MP*	**LP**	*MP*	*MP*

11 A detailed description about the *precise* technical parameters (e. g., codecs, bandwidth distribution, resolution) resulting from the bandwidth limitations was not published by Möller et al. (2011).

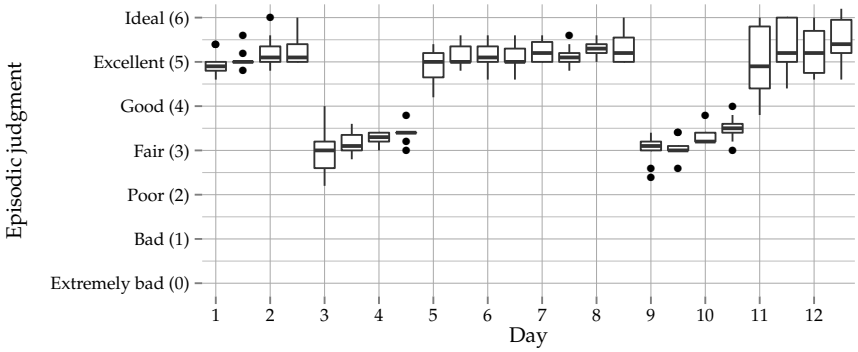

Figure 2.2: Box plot of episodic judgments for condition 4 of Möller et al. (2011) (own illustration).

RESULTS

This experiment was conducted with 58 participants. Two participants were removed from further data analysis, as the required transmission bandwidth was not achieved. This left 10 participants for each condition 2..5 and 16 participants for condition 1. In the following, the data of Möller et al. (2011) is analyzed. Results are reported as MOS ranging from 0 to 6, with standard deviation in brackets.[12] Following, first the episodic judgments and then the multi-episodic judgments are analyzed. MOS for all episodic judgments and multi-episodic judgments per condition are shown in Table 2.2. An exemplary box plot of episodic judgments is shown for condition 4 in Figure 2.2.

EPISODIC JUDGMENTS For all conditions, the episodic judgments reflect the applied episodic performance level, i. e., HP and MP are judged to be better than LP. It can thus be concluded that the technical setup worked as desired. However, comparing the episodic judgments for HP episodes between the conditions 1..4 shows significant

Table 2.2: Episodic judgments and multi-episodic judgments by Möller et al. (2011). Reported as MOS with standard deviation in brackets.

Cond.	Episodic judgments			Multi-episodic judgments		
	HP	MP	LP	2nd day	7th day	12th day
1	5.5 (0.7)	-	-	5.4 (0.3)	5.3 (0.5)	5.7 (0.8)
2	5.2 (0.8)	-	3.9 (1.1)	5.3 (0.7)	5.1 (0.7)	5.5 (0.5)
3	4.9 (1.0)	-	4.0 (1.0)	5.1 (0.6)	4.8 (0.6)	4.9 (0.8)
4	4.9 (0.8)	-	3.7 (0.8)	5.1 (0.3)	4.7 (0.2)	4.8 (0.6)
5	-	4.8 (0.8)	3.9 (1.1)	4.9 (0.3)	4.8 (0.4)	4.9 (0.6)

12 For details about the applied statistical tests see Section 4.3.

differences ($H(3) = 69.2911$, $p < 0.001$). A post-hoc test (pairwise Wilcoxon rank-sum test with Holms' correction) shows that the episodic judgments for HP are not significantly different between condition 3 and condition 4 ($p = 0.137$), but all other conditions are significantly different to each other ($p \leqslant 0.04$). Episodic judgments of LP are not significantly different between conditions 2..4 ($H(2) = 4.8033$, $p = 0.0906$). The observed difference for episodic judgments of HP are relatively small and are likely an artifact of the between-subject design.

For condition 2..4, a significant difference between the two performance levels is observed ($W = 90159.00$, $p < 0.001$). It must be noted that although HP and LP were judged to be different, LP was still judged approximately as *fair*. It thus does not seem to be perceived as a *severe* degradation. With regard to condition 5, the episodic judgments of LP episodes are close to conditions 2..4. In fact, MP judgments are only slightly lower than HP judgments for conditions 1..4. As MP and HP were not presented together, this indicates that either the scale was used differently between conditions or that both performance levels were perceived as similar.

The results show that the defined-use method could be applied successfully, as episodic judgments could be taken reliably even in a field experiment. Möller et al. (2011) noted that episodic judgments of HP episodes have a tendency to slowly increase over the usage period. Here, an increase of 0.3 pt of MOS is reported over all five conditions. In addition, it is reported that the judgments of HP episodes that follow LP episodes seem to be negatively affected by the preceding LP episodes. Möller et al. (2011) suggested a recovery period of up to two episodes for episodic judgments. A statistical analysis is omitted due to the relative small (potential) difference between episodic judgments, the high standard deviation, and the small number of participants.

MULTI-EPISODIC JUDGMENTS With regard to multi-episodic judgments, the result of this experiment are limited. Here, the multi-episodic judgments are rather close, ranging only from 4.7 (0.2) to 5.7 (0.8). A significant difference between the multi-episodic judgments per condition is only found for condition 4 ($H(2) = 8.4114$, $p = 0.0149$). In all other conditions, no significant difference is observed. For condition 4, a post-hoc test (pairwise Wilcoxon rank-sum test with Holms' correction) shows that the multi-episodic judgment after the 2nd day is significantly different from the judgment after the 7th day ($p = 0.003$). However, the judgment after the 2nd day and the 7th day are not significantly different from the judgment after the 12th day (each $p = 0.450$). In fact, an absolute comparison with regard to multi-episodic judgments between the five conditions conducted must be omitted due to the observed differences for episodic

judgments between the conditions, i. e., an artifact that might be attributed to the between-subject design. It must thus be concluded that only condition 4 resulted in an observable, although rather small, effect on multi-episodic judgments due to the presentation of the 3rd and 4th day in LP.

DISCUSSION Möller et al. (2011) successfully applied the defined-use method for the assessment of multi-episodic perceived quality in a field experiment with a usage period of 12 days. The episodic judgments are consistent with the desired performance levels, showing that the used service could provide the performance levels in a reliable manner. However, the multi-episodic judgments only provide limited insight, as the differences between conditions are rather small. This might be due to the fact that LP did not represent a *severe* degradation, or that the number of degraded usage episodes was too low. Alongside this, a difficulty with this experiment is that no useful details about the technical parameters were provided. This is likely due to the use of (proprietary) technology provided by Skype. In fact, no information about resolution, frame rate, codecs, audio signal bandwidth, echo cancellation, audio coding bandwidth, or video coding bandwidth etc. were reported. Also, recordings or monitoring data about the actual network transmission is not available. Another factor that might have influenced the results is the usage of a widely known service. This might have influenced expectations, due to prior knowledge and experiences with Skype. Furthermore, participants were allowed in this experiment to use the service for personal communication beside the defined usage pattern. Thus, some participants might have used the service more often than others. This might have affected their episodic judgments and also multi-episodic judgments. Nevertheless, Möller et al. (2011) showed the applicability of the defined-use method for investigations in field experiments. Although participants used the service in their home environment on their own, i. e., in settings that were uncontrolled and unknown to the researchers, the episodic results show that perceived quality can be assessed successfully with this method.

2.5 CONCLUSION

Multi-episodic perceived quality has so far only received limited attention. The work of Duncanson (1969) and Möller et al. (2011) form a basis for the investigation of the formation process of multi-episodic perceived quality. In particular, the defined-use method developed and tested by Möller et al. (2011) seems suitable for this investigation. However, neither Duncanson (1969) nor Möller et al. (2011) were able to investigate the formation process of multi-episodic perceived quality in detail.

3

TOWARDS MULTI-EPISODIC PERCEIVED QUALITY

Initial work on multi-episodic perceived quality with the defined-use method was conducted by Möller et al. (2011). Although only limited results were obtained with regard to multi-episodic perceived quality, the results of this experiment show that this assessment method, i. e., presenting one multi-episodic condition to several participants, can be successfully applied in field experiments. Based on this and prior work on effects of retrospective judgments, I designed and conducted a series of experiments to examine the formation process of multi-episodic perceived quality. These experiments form the basis for the development of prediction models for multi-episodic judgments. In the following, I first present the here considered aspects for the application of the defined-use method. This includes judgments, performance levels, usage periods, service types, and tasks. Subsequently, I describe the hypotheses I am going to investigate about the formation process of multi-episodic perceived quality.

3.1 CONSIDERED ASPECTS FOR THE INVESTIGATION

3.1.1 *Initial Experiences*

Möller et al. (2011) investigated multi-episodic perceived quality for a service which provides the first episodes with highest performance. Reduction in performance was only presented for later usage episodes. This allows a participant to familiarize himself with the highest performance level and his resulting perceived quality. When reduced performance occurs, the user can compare his current experience with his prior experiences with the service. This should avoid that a participant needs to compare his current experience, resulting from a reduced performance, to prior experiences with another services to derive a judgment. In the experiment by Möller et al. (2011), at least four episodes were presented with the highest performance level. For my experiments, I follow this approach.

In the experiment of Möller et al. (2011), no anchor stimuli were presented to participants before conducting the multi-episodic part of this experiment. Thus, each participant could only rely on his individual prior experiences for his judgments of the episodic perceived

quality and the multi-episodic perceived quality. This might lead to an effect that the highest performance is judged to be better after reduced performance episodes have been experienced, because the latter might lead to an adjustment of the internal reference. However, such an effect has not been observed in this experiment. One reason for this might be the use of Skype and participants being required to have prior experiences with this service. The impact of prior experiences with a service and thus potential, unknown effects on the formation process of multi-episodic perceived quality can be avoided by creating a *new* service that is not known to participants in advance. Besides creating a new service, the evaluation of perceived quality of short anchor stimuli is conducted here with each participant before presenting the multi-episodic condition. Although this might set expectations for the performance to be experienced, it gives participants a shared basis for their judgments. The perceived quality assessment of anchor stimuli is denoted in the following as *training*.

3.1.2 *Judgments*

Perceived quality judgments are taken similarly to Möller et al. (2011). For each usage episode, the retrospective perceived quality is assessed. These so-called *episodic judgments* allow the determination of how the performance of each usage episode was perceived. Episodic judgments are taken directly after finishing a usage episode.

The investigation of multi-episodic perceived quality will be based on retrospective judgments, which are in the following denoted as *multi-episodic judgments*. For a multi-episodic judgment, a participant is requested to assess his *perceived quality of all prior usage episodes with the service*. It is assumed that differences in performance between multi-episodic conditions manifest as consistent differences in the multi-episodic judgments. Here, *multi-episodic condition* refers to a set of usage episodes and their defined occurrence over the usage period with a defined performance per usage episode and a task per usage episode. Similar to episodic judgments, multi-episodic judgments are acquired after finishing a usage episode. If a multi-episodic judgment is required, it is assessed after the episodic judgment. This should prevent an influence on the episodic judgment due to the assessment of the multi-episodic perceived quality. In fact, assessing multi-episodic perceived quality directly after the episodic judgment might increase the impact of this very episode on the multi-episodic judgment. It is not yet known, however, if such an effect occurs and must be left for future work.

Following Möller et al. (2011), the 7-point *Continuous Category Rating (CoCR)* scale is used for the episodic judgments and multi-episodic judgments (cf. Figure 2.1, page 22). The 7-point CoCR allows for more fine-grained judgments than the 5-point ACR scale. This might enable

the observation of small differences between conditions but might also introduce small deviations, as more rating possibilities are provided. Using the same scale for both judgments enables a direct comparison between both judgments without requiring a conversion between scales.

3.1.3 *Performance Levels*

For the investigation of multi-episodic perceived quality, three performance levels are used. These are denoted as *High Performance (HP)*, *Medium Performance (MP)*, and *Low Performance (LP)*. All three performance levels should be clearly distinguishable with regard to perceived quality. HP denotes the highest performance and should lead to higher episodic judgments than MP. Both should be judged to be better than LP. In line with Möller et al. (2011), performance levels are selected, so almost no macroscopic fluctuations of perceived quality occur. This avoids potential effects due to within-episodic fluctuations, as their impact on episodic judgments and multi-episodic judgments are not yet fully understood (cf. Section 2.3). The presentation of all three performance levels in one condition allows the investigation of the existence of a peak effect, which cannot be quantified using two performance levels alone (cf. Section 2.4). HP and LP are presented here in all conditions per experiment to be presented, so the judgments of the conditions are directly comparable and are not affected potentially by differences in worst and best performance levels. This is useful for the investigation of the impact of the necessary between-subject design.

The results of Möller et al. (2011) show that episodic judgments are in line with the defined performance levels, because these show a clear difference of episodic judgments between HP and LP. However, the impact on multi-episodic judgments due to LP usage episodes was very limited (cf. Section 2.4.4). This indicates that the selected performance level for LP did, in fact, produce degradations that were perceived and judged, but were not severe enough to produce observable effects on multi-episodic judgments. Thus, LP must be selected in such a way that degradations are *severe* enough to produce *observable* effects on multi-episodic perceived quality. As the inability to fulfill a task due to overly severe degradations and the impact on perceived quality due to frustration is so far unknown, successful task fulfillment is required throughout this thesis for all applied performance levels.

3.1.4 *Usage Periods*

Möller et al. (2011) applied the defined-use method in a usage period of 12 days. Whereas Möller et al. (2011) focused on a usage pe-

riod of multiple days, multi-episodic perceived quality also occurs in one session if a session consists of multiple usage episodes. Studying multi-episodic perceived quality in one session alone allows one to conduct experiments in a controlled laboratory environment. Typical experiments on perceived quality do not exceed 90 min to avoid an influence of fatigue (cf. Schatz et al., 2012). In fact, limiting the usage period to such a short time frame reduces the required effort and thus allows the investigation of a higher number of multi-episodic conditions. Furthermore, the environment and equipment can be kept constant for all participants and thus does not have to be considered as confounding factors.

In addition to the reduced effort for the investigation of multi-episodic perceived quality, the findings form a meaningful starting point for the investigation of multi-episodic perceived quality over several days. Three usage periods are investigated in this thesis: one session, 6 days, and 14 days.

3.1.5 *Service Types*

In this thesis, two types of telecommunication services are used. Here, services are of special interest that are frequently used and enable rather short usage episodes. Different service types must be considered, as it is not yet known if the formation process of multi-episodic perceived quality is affected by the service type. For each service type, a generic task needed to be selected in such a way that solving one such task in one interaction results in a usage episode. In the following, I present the service types and tasks which I will use for the investigation of multi-episodic perceived quality.

Speech Telephony

Speech telephony services provide live communication between two or more remote parties for spoken interaction. This is a well established and, in fact, classic telecommunication service. The quality perception and underlying influence factors for speech telephony are well understood, and standardized evaluation methods have been developed for the evaluation of perceived quality (e. g., ITU, 1992).

TWO-PARTY CONVERSATION

Speech telephony is most often used for communication between two remote parties who engage in a conversation. A telephony conversation is an interactive exchange of information with changing roles of speaker and listener between caller and callee (Hopper, 1992). The interaction behavior of caller and callee can affect the perceived quality for both parties (e. g., Schoenenberg et al., 2014; Egger et al., 2010).

Methods have been developed to achieve a comparable interaction behavior in a conversation and thus limit the impact of different behavior on perceived quality. Most prominent are the *Short Conversation Scenarios (SCSs)* in which caller and callee need to solve a typical two-party telephony task together (Möller, 2000, p. 76). Here, caller and callee need to exchange a defined set of information while a conversational structure is suggested. A common situation is mimicked in which the caller has a demand with specific requirements, which he tries to fulfill by initiating the conversation and informing the callee about his demand. Based on this information, the callee selects an appropriate predefined option or information and presents this to the caller. If this fulfills the requirements of the caller, a second information transfer is initiated. Here, the callee provides information to the caller, so that the caller can finally fulfill his requirement. For this method, standardized scenarios are defined in the ITU-T Recommendation P.805 (2007). The standardized SCSs usually result in a conversational duration of 3 min to 7 min. This allows the investigation of the whole range of degradations for speech telephony. Here, also those degradations that *affect* the usage behavior can be evaluated. Furthermore, an active conversation allows to investigate the impact of degradations in a setting in which speech telephony is actually used. Besides the advantages, the evaluation of perceived quality in an active conversation requires a great effort. First, a service/system must be available that can provide the desired performance levels. This is especially a problematic in a field experiment. Second, variations in user behavior can affect quality perception and thus quality judgments. This might be problematic for the investigation of multi-episodic perceived quality.

THIRD-PARTY LISTENING

Perceived quality of speech telephony can be assessed to a certain degree in a passive situation. Here, a participant listens to a recorded conversation and thus is not an actual part of this conversation, i. e., his behavior cannot affect the conversation. If the recording of a two-party conversation is presented, this is denoted as third-party listening (ITU-T Recommendation P.832, 2000, p. 13). This is, in fact, an artificial situation, as it cannot occur in a two-party conversation. Here, only a monologue of one conversational partner might occur as part of a conversation. For multi-party conferencing, however, this is a likely situation. In either case, using recordings of a complete two-party conversation is here considered to represent a usage episode if it contains a meaningful conversation.

The elimination of user behavior allows the use of recordings of conversations and thus the presentation of the exact same stimuli to multiple participants. If the desired degradations do not affect the

behavior of caller and callee, the degradations can even be inserted via post-processing the recordings.

A listening-only experiment has one major limitation besides the inability to assess the impact of degradations on the speaking phase (Guéguin et al., 2008). In fact, a passive observer is not forced to follow a conversation, as he does not have to react to a conversational partner. This can be avoided by applying a task which requires following the conversation. For conversations based on SCSs, a note-taking task can be applied. Indeed, for SCS the actual task of caller and callee is to exchange a specific set of information, i. e., each one needs to answer specific questions. These questions are similar for the standardized SCSs and can be used as a task in a listening-only situation while presenting recordings of SCSs. This forces participants to follow the conversation to successfully solve the task.

For the assessment of multi-episodic perceived quality, third-party listening has some advantages over two-party conversation although the usage situation is artificial. Most importantly, the task can be solved by a participant alone. This eliminates the need for a conversational partner, and the same stimulus (i. e., a conversation including degradations) can be presented to multiple participants. Here, even the duration of usage episodes can be defined beforehand. Also, the technical complexity is reduced, as neither live transmission nor live processing is required.

Entertainment Media Consumption

Telecommunication services for media consumption provide a unidirectional transmission of (multi-)media content to a user on his request. This can be unimodal content, such as audio, and multi-modal content, such as audiovisual content.

A typical usage scenario is the provision of media content for entertainment purposes. Services that provide media content on demand are denoted as *Audio-on-Demand (AoD)* for audio-only content and as *Video-on-Demand (VoD)* for audiovisual content. For such a service, a user can select from the available content that item which he currently desires. The desired item is then transmitted to him. While the media selection procedure is often interactive, actual media consumption provides only limited interactivity. Here, a user might be allowed to pause, seek, or abort the consumption. In fact, the interactivity of an on-demand service can be completely limited by presenting predefined content and not allow in-presentation interaction. In contrast to telephony, media entertainment focuses on the consumption of pre-produced content. This allows the use of high-end recording equipment and adequate post-processing. Thus, limiting in general the sources of severe degradations to the transmission and the reproduction.

Media-on-demand services are well-suited for investigating multi-episodic perceived quality. First, such a service can be set up in a non-interactive way and thus avoid the impact of varying user behavior. Second, content can be provided that is of interest for participants and thus motivates them to participate in such an experiment. Besides this, the content can be pre-processed in advance, limiting technical complexity.

3.2 HYPOTHESES

Based on prior work on retrospective experiences (cf. Section 2.2) and quality assessment (cf. Section 2.3), I developed 7 hypotheses to investigate the formation process of multi-episodic perceived quality. The experimental investigation of these hypotheses will be used for the implementation of a prediction model for multi-episodic judgments in Chapter 6. The major goal here is to investigate if models based upon the average of prior episodic judgments are sufficient or more sophisticated model are required due to observable effects. The hypotheses are presented in the following.

3.2.1 Hypothesis: *Number of Consecutive LP Episodes*

Hypothesis 1 (H1) *Increasing the number of LP episodes before a multi-episodic judgment decreases this judgment.*

The presentation of LP episode(s) is expected to result in a reduction in multi-episodic judgments compared to the presentation of these usage episode(s) in HP. When presenting all episodes in HP, the multi-episodic judgment should be sufficiently reflected by averaging all prior episodic judgments (cf. Möller et al., 2011). The more LP episode(s) are presented, the higher should be the expected reduction of multi-episodic judgments. Here, a lower boundary is expected to be set by the episodic judgments for LP episodes. Based on this hypothesis, the impact of LP episodes can also be quantified, which is important for the implementation of a prediction model.

3.2.2 Hypothesis: *Position of LP Episode(s)*

Hypothesis 2 (H2) *The more HP episodes are presented directly before a multi-episodic judgment, the lower is the negative impact of earlier presented LP episodes.*

A *recency effect* has been observed in sequential learning and retrospective judgments of episodic experiences. With regard to perceived quality, an effect of recency could be observed in stimuli with macroscopic performance fluctuations if stimuli were long enough. If an effect of recency occurs for multi-episodic perceived quality, then usage

episodes with close temporal proximity to a multi-episodic judgment have a higher impact than episodes that occurred earlier. By varying the position of LP episode(s) before a multi-episodic judgment, the existence of a recency effect can be investigated. If a recency effect occurs, conditions that present LP episode(s) closer to the multi-episodic judgment will result in a lower multi-episodic judgment than those conditions that present more HP episodes following the same number of LP episode(s).

3.2.3 Hypothesis: *Non-Consecutive vs. Consecutive LP Episodes*

Hypothesis 3 (H3) *The presentation of non-consecutive LP episodes leads to a higher reduction of multi-episodic judgments than presenting the same number of LP episodes consecutively.*

Here, it is investigated whether the number of performance changes between episodes affects multi-episodic judgments. It is expected that increasing the number of changes results in a higher reduction, as the performance is less predictable for participants. This can be investigated by presenting the same number of LP episodes either consecutively or non-consecutively before a multi-episodic judgment. If the effect is small or non-existing, then it might be hidden by a recency effect. Indeed, keeping the number of HP and LP episodes constant, LP episodes must be separated by HP episode(s) in non-consecutive cases. Thus, some LP episodes are presented earlier than in consecutive cases.

3.2.4 Hypothesis: *Strength of Degradation*

Hypothesis 4 (H4) *The lowest experienced episodic performance has an increased impact on multi-episodic judgments, whereas less severe degradations are less important.*

The so-called peak effect, which has been observed in retrospective assessment of episodic experiences, denotes a higher impact of the worst part of an experience and a lower impact of less displeasing parts of an experience on a retrospective judgment of this experience (cf. Section 2.2). Such an effect has been observed for perceived quality affecting the quality formation process (cf. Section 2.3). The existence of such an effect has not been investigated for multi-episodic perceived quality. It can be investigated by presenting more than two performance levels and analyzing their impact on multi-episodic judgments. In fact, these performance levels must result in a different perceived quality. If a peak effect exists, then the episode(s) presented with the worst performance level should have a higher impact on the multi-episodic judgment than episodes which provide a better episodic experience. This is investigated here by introducing a

third performance level denoted as MP in addition to HP and LP. Here, MP must be selected, so it achieves a better perceived quality than LP but is inferior to HP. This allows the determination of the impact of MP episodes on multi-episodic judgments compared to LP episode(s) and thus the investigation of the existence of a peak effect.

3.2.5 Hypothesis: *Recovery after LP Episodes*

Hypothesis 5 (H5) *Presenting additional HP episodes after a negatively affected multi-episodic judgment results in an increase of the following multi-episodic judgment.*

This hypothesis is similar to H2, but focuses on the recovery after a negatively affected multi-episodic judgment. Recovery can be investigated by presenting only HP episodes after the negatively affected multi-episodic judgment. This should result in an increase of the additional multi-episodic judgment. If enough HP episodes are presented, this final judgment should reach a similar level as if no LP episodes were presented at all.

3.2.6 Hypothesis: *Duration of LP Episodes*

Hypothesis 6 (H6) *LP episodes with a much longer duration result in a higher reduction of multi-episodic judgments than shorter LP episodes.*

For retrospective judgments of episodic experiences with macroscopic fluctuations, a *duration neglect* could be observed (cf. Section 2.2). However, such an effect must not necessarily occur for multi-episodic perceived quality. In fact, in H1, the number of LP episodes and the impact on multi-episodic judgments is investigated. Thus, for an increasing number of LP episodes, a longer overall duration of LP is experienced. This, however, leaves open to conjecture whether the formation process of multi-episodic perceived quality relies *a)* on the overall duration of LP episode(s), *b)* the number of LP episode(s), or *c)* both. In case of a) and c), episodic judgments alone would not contain all the required information for the prediction of multi-episodic judgments. A duration neglect can be investigated by comparing the results of conditions that are similar with regard to performance level while varying the duration of the LP episode(s).

3.2.7 Hypothesis: *Services are Judged Independent*

Hypothesis 7 (H7) *The multi-episodic judgment for one service is not affected by the presentation of a second service in the same usage period.*

Multi-episodic perceived quality can only be assessed in retrospect by evaluating prior, recallable experiences of a service and derive the

judgment. If in the same usage period multiple services have been used, this might be problematic. The multi-episodic judgment of a service could be affected by the presence of other service(s). Here, an assessor might fail to attribute perceived quality correctly to a service, or the other service(s) might affect expectations. For investigating multi-episodic perceived quality in one session alone, this is not problematic, as the exposure to other services within a multi-episodic condition can be controlled. However, for a usage period spanning several days, this can hardly be prevented. It is therefore important to understand if the use of other services affects the multi-episodic judgment of a service to be judged.

3.3 CONCLUSION

In this chapter, I first presented the aspects that might affect the formation process of multi-episodic perceived quality. Subsequently, I presented the service types, the tasks, and the performance levels that will be used for the investigation of multi-episodic perceived quality. Then, I presented my hypotheses of the quality formation process for multi-episodic judgments. These form the basis of the experimental investigation. With regard to modeling, most important seems to be H1, as this hypothesis is expected to result in a large effect. All other hypotheses provide knowledge about edge cases of the formation process of multi-episodic perceived quality. In the following chapter, the experiments on multi-episodic perceived quality in one session are presented. In Chapter 5, I present the experiments with usage periods of multiple days.

4

MULTI-EPISODIC PERCEIVED QUALITY IN ONE SESSION

Four experiments were conducted for the evaluation of the seven hypotheses on multi-episodic perceived quality in one session. The session duration is limited here to 45 min. These experiments follow the same experimental procedure but differ in usage situation, task, and service type. An overview on the four experiments is given in Table 4.1. In fact, in each experiment only a subset of hypotheses is investigated. E1 and E2a are the major experiments, focusing most importantly on H1 and H2. Both experiments differ in usage situation, i. e., two-party conversation vs. third-party listening. E2b and E3 are especially designed for the investigation of H7 and H6, respectively.

Table 4.1: Conducted one-session experiments: E1, E2a, E2b, and E3.

Experiment	Service Type(s)	Task	Episodes	Hypothesis
E1	Telephony	Two-party conversation (SCS)	6, 9	H1, H2, H4, H5
E2a	Telephony	3rd-party listening (SCS)	6	H1, H2, H3
E2b	Telephony and VoD	3rd-party listening (SCS) and movie	12	H7
E3	AoD	Audio book	6	H6

In the following, first the experimental design that is shared between the four experiments is presented. Here, the multi-conditions for the investigation of the seven hypotheses are presented. Then, the performance levels, procedure, and content are presented. In the second part of this chapter, the results of the experiments are presented per hypothesis.

4.1 DESIGN

For the investigation of multi-episodic perceived quality in one session, service types were selected that are based on speech interaction

or speech content. Using unimodal rather than multi-modal service types omits considering multi-modal integration for the quality formation process. In E1, E2a, and E2b, a speech telephony service is used. Two-party conversation is applied in E1, whereas third-party listening is used in E2a and E2b. An AoD service is used in E3, presenting a speech-only audio book. For the investigation on the impact of a second service (H7), a multi-modal service (i. e., a VoD service) is used for E2b in addition.

The four experiments consisted of 6 episodes except for one condition of E1. The minimal duration per episode was selected to be at least 2 min. In fact, all experiments except E1 used media consumption and thus the duration per episode could be defined beforehand.

Per experiment except E1, two performance levels were used, i. e., *High Performance (HP)* and *Low Performance (LP)*. In E1, also *Medium Performance (MP)* was presented in addition. For all experiments, the first three episodes were presented in HP. This enables participants to experience the service in a well-working setting similar to Möller et al. (2011). Non-HP episodes were only introduced per service for the 4th, 5th, and 6th episode.

In all experiments, multi-episodic judgments were taken on the 7-point CoCR scale after every third episode with the service. Thus, the first multi-episodic judgment[13] was taken after presenting only HP episodes and therefore should be similar between all conditions. In fact, this allows to investigate the potential impact of the applied between-subject design. Furthermore, this judgment can be used as a reference to assess the impact of the presented LP episode(s) on following multi-episodic judgments. This judgment is in the following denoted as the *reference*.

4.1.1 *Conditions*

Overall 11 conditions were created that allow the investigation of the seven hypotheses in detail. All conditions are shown in Table 4.2. The impact of varying episodic performance on multi-episodic perceived quality can be evaluated by comparing the multi-episodic judgments between the conditions. In addition, an influence of the different usage situations, i. e., two-party conversation vs. third-party listening, can be investigated by comparing episodic judgments as well as multi-episodic judgments of the same condition between E1 and E2a.

All hypotheses except H5 are investigated with regard to the multi-episodic judgment after the 6th episode. H5 is investigated with regard to the multi-episodic judgments after the 3rd, 6th, and 9th ep-

13 Throughout this thesis the term *multi-episodic judgment* is used in singular if referring to measurements after the *same* (in terms of time) usage episode even when describing different multi-episodic conditions. Plural is used if the judgments were taken after different episodes (in terms of time).

Table 4.2: One-session experiments: overview on conditions Non-HP episodes are in **bold** (LP) and *italic* (MP).

Condition	Episodic performance				
	1-3	4	5	6	7-9
C0	HP	HP	HP	HP	-
C1	HP	**LP**	HP	HP	-
C2a	HP	HP	**LP**	HP	-
C2b	HP	HP	**LP**, long	HP	-
C3	HP	HP	HP	**LP**	-
C4	HP	**LP**	**LP**	HP	-
C5a	HP	HP	**LP**	**LP**	-
C5b	HP	HP	**LP**	**LP**	HP
C6	HP	**LP**	**LP**	**LP**	-
C7	HP	HP	**LP**	*MP*	HP
C8	HP	**LP**	HP	**LP**	-

isode. C0, which presents only HP episodes, was only conducted in E2b for the VoD service. For all other experiments, this condition was omitted. Here, the reference is used as approximation for the multi-episodic judgment after the 6th episode of C0.

H1, i. e., increasing the number of LP episodes reduces the following multi-episodic judgment, can be investigated by comparing the multi-episodic judgment after the 6th episode for C3, C5[14], and C6 as well as C2a and C4. C3, C5, and C6 present an increasing number of LP episodes directly before this multi-episodic judgment, whereas C2a and C4 present the last episode before this multi-episodic judgment in HP.

H2 focuses on the position of the LP episodes towards the following multi-episodic judgment. Based on the recency effect, it is expected that increasing the number of HP episodes before a multi-episodic judgment reduces the negative effect of previously presented LP episodes. This can be investigated by comparing the multi-episodic judgment after the 6th episode of C1, C2a, and C3 as well as C4 and C5 for one and two LP episodes.

In H3, it is assumed that consecutive LP episodes are preferred over the same number of non-consecutive LP episodes, because the performance varies less often. This should lead to higher multi-episodic judgments for consecutive cases. This hypothesis can be investigated by comparing the multi-episodic judgment after the 6th episode of C8 with C4 and C5.

14 With regard to the multi-episodic judgment after the 6th episode, it is not differentiated between C5a and C5b, because both conditions are identical until this judgment.

H4 focuses on the investigation of a peak effect, i. e., multi-episodic judgments are more affected by the lowest episodic performance than less severe degradations. This is investigated by introducing the performance level MP in addition to HP and LP. MP should provide a perceived quality worse than HP but better than LP. Comparing the multi-episodic judgment after the 6th episode of C7 with C5 and C2a, allows to investigate this hypothesis. These three conditions present the 5th episode in LP, but differ in the performance level of the 6th episode. This episode is either presented in MP, HP, or LP. If a peak effect occurs for multi-episodic perceived quality, the result of C7 should be closer to C2a.

H5 is closely related to H2. Here, the recovery between two multi-episodic judgments due to the presentation of additional HP episodes is investigated. This can be evaluated with C5b and C7. Both conditions are extended by an additional block of three HP episodes. Here, the multi-episodic judgment after the 6th episode should show a larger negative effect than the judgment after the 9th episode.

H6 focuses on the impact of the duration of one LP episode on the following multi-episodic judgment. Here, a higher reduction is expected if a longer LP episode is presented. This is evaluated by comparing C2a and C2b, which both present the 5th episode in LP. All episodes in C2a and C2b have a similar duration except the 5th episode of C2b. This episode is twice as long in case of C2b. Doubling the duration is expect to result in a measurable effect if the duration is not neglected. The 5th episode is used for this, so the difference in duration between episodes is less obvious to participants.

In H7, it is hypothesized that multi-episodic perceived quality is judged on a per-service basis, i. e., usage of other service(s) in the same usage period does not affect the multi-episodic judgments of the judged service. This can be be investigated by presenting a second service in the same session. Following the experimental approach of sequential usage episodes, this can be investigated by presenting the two services alternatingly. This is investigated in E2b by presenting a speech telephony service in C5 and a VoD service. Here, the VoD service is either presented in C0, C4, or not at all. This service is presented in two conditions to investigate if its multi-episodic conditions affect the episodic judgments as well as the multi-episodic judgments of the speech telephony service.

4.1.2 *Performance Levels*

In each of the four experiments, a speech-only service was used. By using very similar parameters for the performance levels, the results of conditions shared between experiments can be compared. This allows drawing conclusions about the impact of the usage situation

and service type on the formation process of multi-episodic perceived quality.

Three performance levels (HP, MP, and LP) that result in different episodic judgments needed to be selected. Here, LP should produce a measurable effect on multi-episodic judgments. Thus, LP should not only be noticeable different, but rather a *severe* reduction in performance. However, it must be ensured that task fulfillment remained possible. All performance levels should lead to *constant* impairments rather than macroscopic fluctuations. For digital telecommunication services this can be achieved in the compression stage, e. g., selecting and configuring a codec.

HP should provide the state-of-the-art performance. For speech telephony, this is at the time of this writing the transmission in wideband with proper loudness, but without temporal clipping, noise, echo, or other negative factors. Here, the codec G.722 (Mode 1) is selected (ITU-T Recommendation G.722, 2012). This codec provides a similar perceived quality to uncompressed wideband.

For LP, the speech signal is coded with LPC-10[15]. This codec is designed for low bit rate radio transmission while providing intelligibility rather than natural reproduction. The re-synthesized speech signal sounds very unnatural and is described as robotic and muddy with a hissing background noise. As LPC-10 is not used for speech telephony, it is not an ecological valid degradation. Although this might limit generalizability of experimental results, LPC-10 is useful for the investigation of multi-episodic perceived quality, as it allows to create a severe reduction in performance while maintaining intelligibility and thus tasks remain solvable. Furthermore, LPC-10 can be used as baseline condition for future work, as a patent-free, open-source implementation is available.

For MP, G.711 (ITU-T Recommendation G.711, 1988) is selected. This codec is often used as reference for narrowband speech telephony. Compared to G.722, the difference between narrowband and wideband on perceived quality is measurable. In comparison to LPC-10, the re-synthesized speech signal contains far less artifacts.

As LPC-10 is rarely compared to G.711 or G.722, an objective evaluation using *Perceptual Objective Listening Quality Assessment (POLQA)* (ITU-T Recommendation P.863, 2014) was conducted. POLQA estimates the MOS for the 5-point ACR scale. Köster et al. (2015) presented a transformation function that allows to transform judgments from the 5-point ACR scale to the here used 7-point CoCR scale. For the evaluation of LPC-10, 12 s German speech samples have been processed with all three codecs. The resulting speech signals were then evaluated with POLQA in super-wideband mode. The results are shown in Table 4.3. It must be noted that LPC-10 is very different to G.722 and G.711. LPC-10 only achieves a MOS of 1.9 on the 7-point CoCR. In fact,

15 LPC-10 is also known as FS-*1015* and STANAG *4198*.

Table 4.3: Performance levels: Comparison of the selected codecs for HP, MP, and LP with POLQA. The prediction was transformed to the 7-point CoCR scale (Köster et al., 2015).

Performance	Signal bandwidth	Codec	POLQA
HP	50..7000 Hz	G.722, Mode 1	4.0
MP	300..3400 Hz	G.711	3.3
LP	300..3400 Hz	LPC-10	1.9

G.722 results in higher MOS than G.711, but the difference is smaller than between G.711 and LPC-10. However, for the investigation of H4, the actual differences between the three performance levels are not important as long as the episodic judgments are different.

The end-to-end delay is an additional factor influencing the perceived quality. This is important for E1, due to the use of a conversational task. A one-way delay of up to 100 ms is rarely noticeable and thus not perceived as a degradation (ITU-T Recommendation G.107, 2015, p. 9). LPC-10 requires a look-ahead of 90 ms while a frame duration of 22.5 ms is applied. This results in a coding delay of up to 112.5 ms. G.711 and G.722 provide a algorithmic delay of 0.125 ms and both codecs can be used with a minimal frame duration of 10 ms. Thus, LPC-10 might introduce a noticeable difference alone due to the additional coding delay. This might result in an additional reduction in perceived quality for two-party conversations.

The VoD service for E2b was presented on a tablet computer. Here, a *Nexus 7 (2013)* was used. This device provides a 7 inch screen with a resolution of 1920x1200 px. For HP, content is downscaled to a resolution of 1280x720 px and encoded with H.264 (25 FPS, 5 Mbit/s, two-pass). For LP, the video signal is degraded by setting the QP factor to 50. This results in a constant blockiness. The audio channel is encoded with AAC (48 kHz, stereo) for HP and LP. Degrading the video channel only is chosen to prevent a direct comparison of degradations between the two services. However, both services were presented on the same device.

In all experiments, a diotic representation was provided using a pair of headphones. The experimental setups are described in the Appendix i (p. iii).

4.1.3 *Procedure*

The one-session experiments consisted of 4 stages with an overall duration of up to 90 min. In the *first stage*, participants were informed about the experiment, the task(s), and the experimental procedure. In this stage, participants reported their demographic data.

In the *second stage*, participants were checked for normal hearing. Here, an adapted Békésy audiometry was conducted, covering the spectrum for wideband speech telephony.[16] A participant is considered to have hearing capabilities within normal limits if his hearing sensitivity is between 0 dB HL and 25 dB HL (Roeser, 2007, p. 256). For E2b, audiometry was skipped due to the increased duration of the multi-episodic part as well as in E3 due to the focus on the perception of duration.

In the *third stage*, typical speech telephony degradations were presented. Here, participants assessed the perceived quality of 27 speech stimuli with a duration of approximately 8 s. In each stimuli, two sentences were presented by one human speaker. These stimuli contained different signal bandwidths (wideband, narrowband), codecs (G.711, G.722, GSM-FR, LPC-10), white noise (fullband, no codec), and random packet loss (2%, and 5%; G.722, Mode 1, PLC: zero insertion). In E2b, also degradations for the VoD service were presented. 22 videos with an approximate duration of 8 s were used. Here, only the QP factor was varied for H.264 (0, 42, 47, and 50). After the presentation of each stimulus, the perceived quality of this stimulus was assessed using the 7-point CoCR scale.

In the *fourth stage*, one multi-episodic condition was presented. Here, usage episodes were presented sequentially. Directly after finishing an episode, participants judged the episodic quality on the 7-point CoCR scale (episodic judgment). In addition, the experienced degradations could be described in qualitative form (German: "Falls Störungen aufgetreten sind, dann beschreiben Sie diese bitte."; English: "If degradations occurred, then please describe them."). After every 3rd episode per service, a multi-episodic judgment was taken using the same scale. Episodic judgments were presented with the question "How would you judge the overall quality of the just finished episode?" and multi-episodic judgments with "How would you judge the overall quality of all episodes so far?". These questions were modified, i. e., replacing the term episode to telephone call or similar, to customize them for each experiment. The used questions in German for each experiment are presented in the Appendix ii (p. ix). The 7-point CoCR scale and questions were inspired by Möller et al. (2011).

4.1.4 *Content*

To achieve comparable two-party conversations in E1, one SCS (ITU-T Recommendation P.805, 2007) per episode needed to be solved. Here, the presentation order of SCSs was kept constant for all conditions. For later use in E2a and E2b, the conversations of all participants of

16 Audiometry was conducted with an *Ear 1.7* audiometer (*http://www.ear20.de/ear_1_7.html*) connected to a *Sennheiser HDA 200* for 125 Hz, 250 Hz, 500 Hz, 1 kHz, 2 kHz, 4 kHz, and 8 kHz.

E1 were recorded. For E2a and E2b, one recording for each of the episodes 1..6 was selected. These were required to show normal task-solving behavior. The duration of the selected conversations ranged from 128 s to 194 s with an average of 153 s.

For the VoD service used in E2b, scenes of a sitcom were selected. Scenes were taken from *The Big Bang Theory* (Season one, Blu-ray version, German). The scenes were selected to be meaningful and self-contained, and thus each one could represent a usage episode. These scenes were selected, so the duration of the two-party conversational recordings were closely matched. For each episode one scene was presented. These ranged from 134 s to 198 s with an average duration of 166 s. For E3, an audio book was selected. Here, Isabel Allende's *"City of the Beasts"* was used, which is spoken by a male German native speaker.[17] Usage episodes varied in duration from 174 s to 199 s with an average duration of 184 s. The usage episode with doubled duration was 362 s long.

4.2 PARTICIPANTS

All four experiments were conducted at Technische Universität Berlin, Germany. Participants were required to have normal hearing (see Section 4.1.3). In E2b, normal vision was also mandatory (vision aid allowed).

E1 was conducted from April 2014 to December 2015 with 78 female and 51 male participants aging from 18 to 53 years ($\mu = 26.7$, $\sigma = 5.9$). E2a was conducted from August 2014 to March 2015 with 75 female and 40 male participants aging from 18 to 50 years ($\mu = 26.2$, $\sigma = 4.6$). E2b was conducted in January 2016 with 20 female and 13 male participants aging from 19 to 34 years ($\mu = 26.6$, $\sigma = 4.3$). E3 was conducted in October and November 2015 with 20 female and 16 male participants aging from 18 to 32 years ($\mu = 25.0$, $\sigma = 3.9$). Participation in E1 was compensated with 20 EUR, in E2a and E3 with 10 EUR, and in E2b with 15 EUR.

The audiometry conducted in E1 and E2a showed normal hearing for all participants. Thus, no participant needed to be excluded due to impaired hearing.

All participants were individually checked for inconsistent episodic judgments. A participant is considered inconsistent if more than two episodic judgments exceed the 1.5 × *interquartile range* per usage episode of this condition. The interquartile range is, in fact, a rather conservative criteria. However, due to the lack of ground truth on episodic judgments in multi-episodic assessment, participants should only be removed if severe differences occur. This criteria was met for no participant, and thus none was excluded from the data analysis.

17 The audio book is sold as 24 CD collection: Isabel Allende: *Die Stadt der wilden Götter / Im Reich des goldenen Drachen / Im Bann der Masken*, ISBN: 978-86717-191-5.

Table 4.4: Participants per condition for E1, E2a, E2b, and E3. For E2b the conditions of the VoD are shown as the telephony service was always presented in C5a.

Condition	Experiments			
	E1	E2a	E2b	E3
-	-	-	11 (no VoD)	-
C0	-	-	11	-
C1	18	12	-	-
C2a	15	15	-	16
C2b	-	-	-	20
C3	13	13	-	-
C4	11	15	11	-
C5a	-	24	-	-
C5b	16	-	-	-
C6	13	21	-	-
C7	43	-	-	-
C8	-	15	-	-

Table 4.4 gives an overview on the number of participants per multi-episodic condition for the four experiments. Every condition was at least assessed by 11 participants. A larger number of participants were used for C7 (E1) to precisely quantify the impact of recovery (H5).

4.3 DATA ANALYSIS

In the following, the results of the experiments are analyzed. First, episodic judgments are inspected with regard to consistency between the conditions for each experiment. Second, the impact of the different usage situations in E1 and E2a are investigated. Finally, multi-episodic judgments are evaluated with regard to the hypotheses (cf. Chapter 3).

All results are reported as MOS ranging from 0 to 6 with standard deviation in brackets (for the scale see p. 22). In addition, a statistical evaluation is conducted. Two unpaired samples are evaluated with a *Wilcoxon rank-sum test*. More than two unpaired samples are compared with a *Kruskal-Wallis test*. If this test shows significant differences, then a post-hoc test is conducted. Here, a *pairwise Wilcoxon rank-sum test with Holms' correction* is used. Paired samples are evaluated with a *Wilcoxon signed-rank test*. For all tests, a significance level of 5% is applied. Nonparametric tests are applied due to the rather small sample size for all conditions.

Table 4.5: One-session experiments: episodic judgments. Reported as MOS with standard deviation in brackets.

Experiment	Service	Episodic judgment		
		HP	MP	LP
E1	Telephony (Conversation)	4.2 (0.7)	3.3 (0.8)	1.5 (0.7)
E2a	Telephony (Listening)	4.2 (0.8)	-	1.8 (0.8)
E2b	Telephony (Listening)	4.0 (0.7)	-	1.6 (0.6)
	VoD	4.7 (0.7)	-	1.7 (0.7)
E3	AoD	4.7 (0.6)	-	1.0 (0.6)

4.3.1 *Episodic Judgments*

The episodic judgments of all conditions are compared to investigate the potential influence of the between-subject design. In prior work a negative influence on the episodic judgment of HP episodes could be observed if these directly follow LP episode(s) (Möller et al., 2011). Such an effect could not be observed in any of the here presented one-session experiments.

Table 4.5 shows the episodic MOS for all four experiments by performance level and service. In the following, the differences in episodic judgments between the conditions are investigated for each experiment. Box plots for all conditions per experiments are presented in the Appendix iii (p. xi).

EXPERIMENT E1 In E1, significant differences between episodic judgments of HP episodes are found ($H(6) = 22.4404$, $p = 0.001$). A post-hoc test shows that C2a (MOS: 3.9(0.6)) and C7 (MOS: 4.3(0.8)) are significantly different ($p < 0.001$). For the episodic judgments of LP episodes, no significant differences between the conditions are found ($H(6) = 8.9823$, $p = 0.1746$). Although HP episodes are significantly different between the conditions, an analysis did not yield a reason for this difference. It might be an artifact of the between-subject design or due to the larger sample size of C7. The episodic judgments are significantly different between HP and LP ($W = 129765.50$, $p < 0.001$). A comparison of MP with HP and LP is done later in the evaluation of H4, as it was only presented in C7. It can thus be concluded that the results of E1 are consistent.

EXPERIMENTS E2A AND E2B For E2a, episodic judgments of HP episodes are significantly different between the conditions ($H(6) = 13.9445$, $p = 0.0303$). However, conducting a post-hoc test shows

no significant differences between the conditions for episodic judgments of HP episodes ($p \geqslant 0.066$). For episodic judgments of LP episodes, no significant differences between the conditions are found ($H(6) = 4.317$, $p = 0.6339$). The episodic judgments for HP and LP are significantly different ($W = 98931.00$, $p < 0.001$).

In E2b, the episodic judgments of the telephony service are not significantly different between the three conditions for HP ($H(2) = 2.7806$, $p = 0.249$) and not for LP ($H(2) = 2.3006$, $p = 0.3165$). For the VoD service, episodic judgments of HP episodes are not significantly different ($W = 1739.50$, $p = 0.078$). The episodic judgment of HP and LP episodes are significantly different for the telephony service ($W = 8635.50$, $p < 0.001$) and the VoD service ($W = 2419.50$, $p < 0.001$).

Comparing E2a and E2b with regard to the speech telephony service shows that the judgments of LP episodes are not significantly different ($W = 7836.50$, $p = 0.123$), but a significant difference is found for HP ($W = 35993.00$, $p = 0.015$). This might be due to the differences in the used audio equipment, i. e., used pair of headphones and also sound cards including loudness calibration (cf. Appendix i on p. iii), or the presence of the VoD service. However, the actual reason could not be determined. With regard to episodic judgments both experiments yield consistent results, but differences between the experiments were found.

EXPERIMENT E3 In E3, no significant differences between the two conditions for episodic judgments of HP ($W = 4338.50$, $p = 0.325$) and LP are found ($W = 111.50$, $p = 0.123$). It must be noted that judgments of the LP episode are not significantly different although both conditions differ in the duration of this episode. C2a (normal duration) resulted in a MOS of $0.8(0.6)$, whereas C2b (doubled duration) resulted in a MOS of $1.2(0.5)$. It is thus concluded that the duration did not affect the episodic judgment of this single LP episode. Thus, a duration neglect for episodic judgments is observed. The two performance levels HP and LP are significantly different ($W = 6477.50$, $p < 0.001$).

IMPACT OF USAGE SITUATION: EXPERIMENTS E1 VS. E2A By comparing the episodic judgments between E1 and E2a, a potential impact of the usage situation can be investigated. Between the two experiments, the episodic judgments for HP are not significantly different ($W = 171422.50$, $p = 0.677$). With regard to LP, both experiments show a significant difference ($W = 14751.00$, $p < 0.001$).

This indicates that the usage situation without macroscopic performance fluctuations only seems to affect the judgments of LP episodes. For two-party conversation, a 0.3 pt lower MOS is observed than for third-party listening. This might be due to reduced intelligibility and thus higher effort, but also due to differences in user behavior, e. g., need to repeat information. However, the effect is rather small.

Table 4.6: One-session experiments: multi-episodic judgments after the 3rd usage episode. Reported as MOS with standard deviation in brackets.

Experiment	E1	E2a	E2b (telephony)	E2b (VoD)	E3
Multi-episodic judgment	4.3 (0.7)	4.3 (0.8)	4.0 (0.9)	4.4 (0.8)	4.6 (0.6)

4.3.2 *Multi-episodic Judgments*

In the following, the results for multi-episodic judgments are evaluated. First, the consistency between the conditions is analyzed with regard to the multi-episodic judgment after the 3rd episode. Then the hypotheses on multi-episodic judgments, i. e., differences between the multi-episodic conditions, are evaluated.

4.3.2.1 *Consistency*

In all four experiments, the first three episodes of a service were presented in HP. Thus, the multi-episodic judgment taken directly after this episode should be similar for all conditions of an experiment.

For this judgment, no significant differences are observed for E1 ($H(6) = 3.1883$, $p = 0.7849$), E2a ($H(6) = 6.7373$, $p = 0.3458$), and E3 ($W = 188.50$, $p = 0.368$). For E2b, neither significant differences are observed for the speech telephony service ($H(2) = 0.2865$, $p = 0.8666$) nor the VoD service ($W = 59.50$, $p = 0.973$). This indicates that as long as only HP episodes are presented, the between-subject design did not affect this multi-episodic judgment. Table 4.6 shows the multi-episodic judgment after the 3rd episode for the four experiments. Although not directly comparable due to differences between experiments, it must be noted that the potential differences between the experiments are rather small.

4.3.2.2 *H1: Number of Consecutive LP Episodes*

In H1, it is assumed that increasing the number of LP episodes before a multi-episodic judgment results in a decrease of this judgment. This hypothesis can be evaluated by comparing the multi-episodic judgment after the 3rd episode as *reference* with the multi-episodic judgment after the 6th episode of C3, C5, and C6 as well as C2a and C4. C3, C5, and C6 present one to three LP episodes directly before this multi-episodic judgment, whereas C2a and C4 present one or two LP episode followed by one HP episode. This is investigated in E1 and E2a. The multi-episodic judgment after the 6th episode is shown in Table 4.7.

Table 4.7: One-session experiments: multi-episodic judgments after the 6th usage episode for H1. Reported as MOS with standard deviation in brackets.

Condition	LP episode(s)	Multi-episodic judgment	
		E1	E2a
Reference (HP only)	-	4.3 (0.7)	4.3 (0.8)
C3	6	3.1 (0.8)	3.5 (0.5)
C5a and C5b	5..6	2.3 (0.9)	2.4 (0.9)
C6	4..6	2.1 (0.7)	2.5 (0.8)
C2a	5	3.2 (0.6)	3.5 (0.5)
C4	4..5	2.9 (0.5)	3.0 (0.9)

For E1, the reference, C3, C5b, and C6 are significantly different ($H(3) = 53.2096$, $p < 0.001$). A one-sided post-hoc test finds that the reference is significantly different from these three conditions ($p < 0.001$). Also, C3 and C5b ($p = 0.033$) as well as C3 and C6 are significantly different ($p = 0.018$). No significant difference is found between C5b and C6 ($p = 0.261$). The reference, C2a, and C4 are significantly different ($H(2) = 30.4285$, $p < 0.001$). A one-sided post-hoc test finds that the reference is significantly different to C2a and C4 ($p < 0.001$) and that C2a is significantly different from C4 ($p = 0.039$).

For E2a, the reference, C3, C5a, and C6, are also significantly different ($H(3) = 65.084$, $p < 0.001$). A one-sided post-hoc test finds that the reference is significantly different from these three conditions (C3: $p = 0.002$; C5a and C6: $p < 0.001$). Also, C3 and C5a ($p < 0.001$) as well as C3 and C6 ($p < 0.001$) are significantly different. No significant difference is found for C5a and C6 ($p = 0.727$). The reference, C2a, and C4 are significantly different ($H(2) = 28.9757$, $p < 0.001$). A one-sided post-hoc test finds that the reference is significantly different to C2a and C4 ($p < 0.001$), and C2a is significantly different than C4 ($p = 0.031$).

With regard to the number of LP episodes before a multi-episodic judgment, both experiments yield similar findings. These show a decrease in the multi-episodic judgment if the number in LP episodes is increased from zero to one and from one to two. If no HP episodes follow, a decrease of approximately 1 pt for each presented LP episode was observed. However, it must be noted that both experiments show consistently no decrease from two to three LP episodes. In this case, the multi-episodic judgment remains *above* the episodic judgments of LP episodes. In both experiments, a decrease of approximately 1 pt is also observed if one HP episode follows one LP episode(s). For two HP episodes, a decrease of less than 0.5 pt is observed in both experiments.

Table 4.8: One-session experiments: multi-episodic judgments after the 6th usage episode for H2. Reported as MOS with standard deviation in brackets.

Condition	LP episode(s)	Multi-episodic judgment	
		E1	E2a
C1	4	3.7 (0.6)	3.6 (0.5)
C2a	5	3.2 (0.6)	3.5 (0.5)
C3	6	3.1 (0.8)	3.5 (0.5)
C4	4..5	2.9 (0.5)	3.0 (0.9)
C5a and C5b	5..6	2.3 (0.9)	2.4 (0.9)

It must thus be concluded that H1 is only partly true. A decrease can be observed if up to two LP episodes are presented. Presenting a third LP episode does not seem have an effect on the multi-episodic judgment. In fact, the multi-episodic judgment remains above the episodic judgments of LP episodes, i. e., the difference is approximately 0.6 pt. This indicates that prior presented HP episodes still attribute to the final multi-episodic judgment.

In fact, both experiments show similar effects in the evaluated conditions. However, it must be noted that in absolute numbers the multi-episodic judgment of E2a seems to be higher than for E1 except for the reference. This might be due to the difference usage situation, i. e., two-party conversation vs. third-party listening.

4.3.2.3 *H2: Position of LP Episode(s)*

In H2, it is hypothesized that presenting HP episodes after one or two LP episode(s) reduces the negative impact on the following multi-episodic judgment, i. e., after the 6th episode. C1, C2a, and C3 present each one LP episode with either two, one, or no HP episode(s) before this judgment. C4 and C5 present two consecutive LP episodes, whereas one or no HP episode are presented before this judgment. Table 4.8 shows the multi-episodic judgment after the 6th episode for these conditions.

For E1, the multi-episodic judgments are significantly different between C1, C2a, and C3 ($H(2) = 7.0608$, $p = 0.0293$). A one-sided post-hoc test shows that C1 is significantly different from C2a and C3 (each $p = 0.032$), but C2a and C3 are not significantly different ($p = 0.444$). In this experiment, C4 and C5b are significantly different ($W = 126.00$, $p = 0.031$, one-sided).

For E2a, C1, C2a, and C3 are not significantly different ($H(2) = 0.6447$, $p = 0.7245$), whereas C4 and C5a are significantly different ($W = 249.50$, $p = 0.023$, one-sided).

Table 4.9: One-session experiments: multi-episodic judgments after the 6th usage episode for H3. Reported as MOS with standard deviation in brackets.

Condition	LP episode(s)	E2a
C4	4..5	3.0 (0.9)
C8	4 and 6	2.5 (0.6)
C5a	5..6	2.4 (0.9)

With regard to H2, both experiments yield similarities and differences for varying the position of one or two LP episodes. For E1, presenting the LP episode(s) earlier reduces the impact on the directly following multi-episodic judgment. Thus, a recency effect is observed. In fact, this effect could for one LP episode only observed if more than one HP episode was presented after the one presented LP episode. For E2a, an effect of position could only be observed for two LP episodes but not for one LP episode. This indicates that the usage situation affects the multi-episodic formation process. However, the actual reason for this could not be deduced.

Thus, H2 can be accepted for two-party conversations, but only partly accepted for third-party listening.

4.3.2.4 *H3: Non-consecutive vs. Consecutive LP Episodes*

In H3, it is assumed that the presentation of consecutive LP episodes yields a better multi-episodic judgment than the same number of LP episodes presented non-consecutively. This hypothesis is investigated in E2a only. It can be evaluated by comparing the final multi-episodic judgment of C4 and C5a, which both present two LP episodes consecutively, with C8. C8 presents the 4th and 6th episode in LP. Table 4.9 shows the multi-episodic judgment after the 6th episode for these three conditions.

The final multi-episodic judgment of these three conditions is not significantly different ($H(2) = 4.3146$, $p = 0.1156$). In fact, the final multi-episodic judgment of C5a and C8 is rather close compared to C4. This might indicate a higher impact of the very last episode in terms of a recency effect.

Thus, H3 must be rejected, i. e., the non-consecutive presentation is not judged differently from a consecutive presentation of LP episodes. It must be concluded that either such an effect does not exists, or it is too small to be observed in the conducted experiment. However, an outstanding importance of the performance level of the last episode is suggested.

Table 4.10: One-session experiments: multi-episodic judgments after the 6th usage episode for H4. Reported as MOS with standard deviation in brackets.

Conditions	6th usage episode		Multi-episodic judgment
	Performance	Episodic judgment	
C2a	HP	3.7 (0.6)	3.2 (0.6)
C7	MP	3.3 (0.8)	3.1 (0.7)
C5b	LP	1.7 (1.0)	2.3 (0.9)

4.3.2.5 *H4: Strength of Degradation*

In H4, it is assumed that the lowest episodic performance (here LP) has a stronger negative effect on a following multi-episodic judgment, whereas *lesser degraded* episode(s) (here MP) yield a smaller effect or even no effect at all. This is denoted as peak effect (cf. Section 2.2.2). This is only evaluated in E1 by comparing C2a, C5b, and C7. These present the 5th episode in LP, but differ in the performance level of the 6th episode. Table 4.10 shows the multi-episodic judgment after the 6th episode as well as the episodic judgment of this episode.

The episodic judgments for this episode are significantly different between the three conditions ($H(2) = 27.0301$, $p < 0.001$). A one-sided post-hoc test shows a significant difference between LP and MP ($p < 0.001$) as well as LP and HP ($p < 0.001$). However, HP and MP are not significantly different ($p = 0.07$).

The final multi-episodic judgment of these three conditions is significantly different ($H(2) = 13.3662$, $p = 0.0013$). A one-sided post-hoc shows significant differences for C2a and C5b ($p = 0.006$) as well as C5b and C7 ($p = 0.002$). C2a and C7 are not significantly different ($p = 0.481$).

With regard to the interpretation, it is problematic that the episodic judgments for HP and MP are not significantly different. In fact, it is indicated that MP episodes were perceived and judged slightly worse than HP. This leaves three interpretations. Either this is an artifact of the between-subject design, HP and MP were not different enough, or a peak effect could be observed. Thus, H4 must be left unanswered.

4.3.2.6 *H5: Recovery after LP Episodes*

In H5, the recovery of multi-episodic judgments is investigated. This is investigated in E1 by comparing C5b and C7. Both present 9 episodes while non-HP is presented for the 5th and 6th episode. Table 4.11 shows the three multi-episodic judgments of these conditions. For each condition, the three multi-episodic judgments are signifi-

Table 4.11: One-session experiments: multi-episodic judgments after the 3rd, 6th, and 9th usage episode for H5. Reported as MOS with standard deviation in brackets.

Conditions	Episodic performance	Multi-episodic judgment		
		3rd	6th	9th
C5b	4:HP, 5:LP, 6:LP	4.1 (0.6)	2.3 (0.9)	3.6 (0.6)
C7	4:HP, 5:LP, 6:MP	4.4 (0.8)	3.1 (0.7)	4.1 (0.6)

cantly different (C5b: $H(2) = 25.4401$, $p < 0.001$; C7: $H(2) = 46.8101$, $p < 0.001$). For C5b, a paired post-hoc test shows significant differences between all three multi-episodic judgments (3rd vs. 9th: $p = 0.003$; 6th vs. 9th: $p = 0.003$; 3rd vs. 9th $p = 0.010$). A paired post-hoc test for C7 yields similar findings (3rd vs. 6th: $p < 0.001$; 6th vs. 9th: $p < 0.001$; 3rd vs. 9th: $p = 0.013$). Between these two conditions, the multi-episodic judgment after the 6th episode ($W = 146.00$, $p < 0.001$) as well as the multi-episodic judgment after the 9th episode ($W = 182.50$, $p = 0.006$) are significantly different.

Both conditions show an increase in the final multi-episodic judgment due to the three additional HP episodes. In fact, both conditions still show a significant difference between the multi-episodic judgments after the 3rd and 9th episode. Thus, a negative effect of the presented non-HP episodes is still present, as the final the multi-episodic judgment remains lower than the first multi-episodic judgment.

It must be concluded that the multi-episodic judgment recovers if three additional HP episodes are presented, and thus H5 is accepted. However, from the two conditions alone it cannot be deduced if a recency effect occurred or the increased number of HP episodes alone lead to the increase.

4.3.2.7 *H6: Duration of one LP Episode*

In E3, the impact of varying duration of one LP episode on the following multi-episodic judgment is investigated (H6). It is hypothesized that an increased duration of one LP episode will lead to a higher reduction of the following multi-episodic judgment. In this experiment, C2a and C2b present the 5th episode in LP. While all other episodes are presented with a similar duration, the 5th episode is presented in C2b with approximately the doubled duration. In C2a, this episodes has a duration of 180 s and in C2b 360 s. The episodic judgment for the 5th episode and the final multi-episodic judgment for these two conditions are shown Table 4.12.

Between the two conditions, the episodic judgment of the 5th episode is not significantly different (cf. Section 4.3.1). Also, the multi-episodic judgment after the 6th episode is not significantly different ($W = 181.00$, $p = 0.507$). Thus, H6 has to be rejected, i. e., a neglect of

Table 4.12: One-session experiments: multi-episodic judgments after the 6th usage episode for H6. Reported as MOS with standard deviation in brackets.

Conditions	5th usage episode		Multi-episodic judgment
	Duration	Episodic judgment	
C2a	similar	0.8 (0.6)	3.9 (0.6)
C2b	doubled	1.2 (0.5)	3.7 (0.6)

duration is observed, as even doubling the duration did not produce a measurable effect. In fact, the duration seems to be neglected for the episodic judgment and the final multi-episodic judgment. It can thus be concluded that the episodic judgment sufficiently describes the impact of a LP episode on following multi-episodic judgments.

4.3.2.8 *H7: Services are Judged Independent*

In E2b, it was investigated if the multi-episodic judgments of two services are independent (H7). This is investigated by presenting a VoD service in addition to the already used speech telephony service. Both services needed to be used alternatingly. The speech telephony service was always presented in C5a. The VoD service was presented in C0, C4, or not presented at all. The episodic judgments and multi-episodic judgments for the two services are shown in Table 4.13.

The final multi-episodic judgment of the VoD service is significantly different between C0 and C4 ($W = 107.50$, $p = 0.002$). This shows that the two LP episodes affect the multi-episodic judgment of this service. With regard to the speech telephony service, no significant difference on the final multi-episodic judgment is found ($H(2) = 0.2952$, $p = 0.8628$).

As the final multi-episodic judgment of the speech telephony service is not different if a VoD service is present or not, it can be concluded that the speech telephony service was assessed based on the

Table 4.13: One-session experiments: multi-episodic judgments after the 6th usage episode for H7. Reported as MOS with standard deviation in brackets.

Conditions VoD	Multi-episodic judgment	
	Telephony	VoD
-	2.6 (0.8)	-
C0	2.7 (0.7)	4.7 (0.6)
C4	2.7 (0.8)	3.5 (0.8)

episodes conducted with this service alone. In fact, the participants could only use the content, degradation, and presentation modality to differentiate the two services, as both services were presented on the same device. Nevertheless, participants were able to attribute the episodes and their experience to each service. Thus, H7 can be accepted.

4.4 DISCUSSION AND CONCLUSION

The four experiments show that multi-episodic perceived quality can be assessed in one session. Here, the formation process of multi-episodic perceived quality for one session of up to 45 min was investigated. Although the selected degradation for LP (LPC-10) is not actually used for speech telephony, its presentation resulted in observable effects on episodic judgments and multi-episodic judgments. However, the selection of LPC-10 needs to be considered with regard to generalizability of the results. It might be that some of the observed effects occurred due to the use of LPC-10. Nevertheless, the selected performance levels enabled to investigate the formation process of multi-episodic perceived quality.

It could be observed consistently that increasing the number of LP episodes leads to a reduction in the final multi-episodic judgment (H1). In addition, an effect of saturation is observed if two or three LP episodes are presented. Here, the final multi-episodic judgment did not decrease further although more LP episodes were experienced. In fact, a saturation was expected to occur if the final multi-episodic judgment reaches the same level as the episodic judgments of LP episodes. However, the multi-episodic judgment remained above this threshold. This indicates that previous experiences with HP episodes were still present for the multi-episodic judgment. Such an effect has so far not been found for retrospective assessment of perceived quality.

In addition, the occurrence of a recency effect was investigated in E1 and E2a (H2). Here, the position of one and two LP episode(s) towards the following multi-episodic judgment was varied. In E1, the presentation of HP episode(s) before the final multi-episodic judgment had a positive effect on this judgment. In E2a, this could only be observed for two LP episodes. This indicates that the usage situation affects the multi-episodic formation process. In fact, participants experienced a listening-only situation in E2a, whereas in E1 a two-party conversation was used. The latter might be more mentally demanding. Thus, participants might forget the experienced LP episode faster if HP episode(s) are presented afterwards. In E2a, the impact of consecutive vs. non-consecutive LP episodes was investigated, and the results are conclusive (H3). The three conditions yielded similar multi-episodic judgments and thus H3 cannot be accepted.

In E1, a (negative) peak effect was investigated (H4). However, the results are inconclusive and thus this hypothesis must be left unanswered. In E1, also recovery of multi-episodic judgments was investigated (H5). Here, three HP episodes were presented after two non-HP episodes. The results show that the negative effect reduces for the final multi-episodic judgment due to the three additional HP episodes.

In E3, the duration of one LP episode was varied. It was expected that presenting this episode with doubled duration would lead to a higher reduction of the final multi-episodic judgment than normal duration (H6). However, no significant difference has been observed between the two conditions. This indicates that the actual duration of an episode does not seem to affect the formation process of multi-episodic perceived quality. Thus, H6 must be rejected. In fact, this might be related to use of constant performance rather than macroscopic fluctuations. In the latter case, the actual duration might be easier to memorize and recall, and thus a duration neglect might not occur.

In E2b, it was investigated if the presentation of another service affects the multi-episodic quality formation process of a primary service (H7). The results show that episodic judgments and multi-episodic judgments were assessed independently per service.

In conclusion, the results show that multi-episodic judgments in one session are affected by several effects. Here, a recency effect, a duration neglect, and a saturation effect could be observed. The existence of a peak effect could neither be proven nor disproven. The experimental results, beyond analysis of occurring effects, allow the evaluation of potential prediction models. For this the large number of conditions in E1 and E2a are well suited.

5

MULTI-EPISODIC PERCEIVED QUALITY IN MULTIPLE DAYS

The integration of perceived quality of usage episodes into a multi-episodic perceived quality is not limited to one session alone. Rather usage episodes with a service often occur regularly and might cover multiple days or even longer usage periods. In difference to consecutive usage in one session, the time that passes between episodes might be longer. However, it is so far unknown how and if at all the time between episodes affects the quality formation process of multi-episodic perceived quality. A feasible option to cover such time spans in an experiment is to conduct it as a field experiment. Here, field experiment refers to a subjective experiment conducted with participants in their home environment over several days. This enables repeated-use of a service for rather short episodes while limiting the effort for participants. However, this approach has two inherent limitations. First, the usage environment cannot be controlled and is unlikely to be identical between participants. Second, participants cannot be directly supervised during the experiment. Both might be confounding factors. In addition, technical systems must be robust while providing the desired performance level per usage episode.

For the investigation of multi-episodic perceived quality with a usage period of multiple days, three experiments were conducted denoted as E4, E5, and E6. These experiments apply the defined-use method. Following Möller et al. (2011), each usage episode was presented individually, i. e., one episode per session. E4 and E5 focused on general feasibility of investigating multi-episodic perceived quality in a field experiment. These experiments covered a usage period of 14 days in which participants used two services on a daily basis. Besides feasibility, E4 focused on the reproduction of the results of Möller et al. (2011). In E5, the adaptation speed of multi-episodic judgments was planed to be investigated. Based on E4, E5, and the one-session experiments, E6 was designed. Here, H1, H2, and H3 were investigated. In this experiment only one service was used in a usage period of 6 days.

In all three experiments, two performance levels were applied. These are also denoted as HP and LP. Performance levels were applied day-wise, i. e., usage episodes on the same day with the same service were

Table 5.1: Multiple days: overview on experiments.

Experiment	Service types	Tasks	Days	Participants
E4	Telephony and VoD	Two-party conversation (SCS)	14	20
E5	VoD and AoD	Audio book and Video consumption	14	21
E6	AoD	Audio book	7	95

presented with the same performance level. The performance levels were different between the three experiments.

Table 5.1 gives an overview of the experiments including service types, tasks, and number of participants. In the following, the experiments are presented one after the other. The results are reported as MOS ranging from 0 to 6 with standard deviation in brackets. A statistical evaluation is conducted by applying the nonparametric tests as for the one-session experiments (cf. Section 4.3).

5.1 EXPERIMENT E4

E4 was inspired by Möller et al. (2011).[18] The experiment of Möller et al. (2011) indicated only little differences for multi-episodic judgments between the evaluated multi-episodic conditions (cf. Section 2.4.4). This might be due to the limited number of degraded episodes or the introduced degradations were not severe enough. The goal of E4 was the application of the defined-use method in a similar usage period to Möller et al. (2011) and to gather practical knowledge about how to conduct a field experiment. In addition, conducting E4 had two goals: First, a larger difference of multi-episodic judgments between conditions was desired to investigate the formation process. For this *a)* the number of LP episodes was increased and *b)* the performance parameters selected, so the perceptual difference between HP and LP was increased. Second, two services were used. This enabled to investigate the formation process in presence of two services.

5.1.1 *Design*

In this experiment, a speech telephony service and a VoD service were selected. The two services needed to be used daily over a usage period of 14 days. The telephony service was used by a pair of participants together. Each pair needed to solve two SCSs per day: the first one between 6 am and 1 pm and the second one between 3 pm and

18 The results of E4 are published in Guse and Möller (2013).

midnight. The VoD service needed to be used daily for one episode between 1 pm and midnight. Here, *Friends Season 3* episode 1..8 were used. Each original episode was split into two parts while not disturbing the storyline. The duration of these parts ranged from 12 min to 17 min. For each usage episode one part was presented.

In this experiment, two conditions were investigated. In the first one, all usage episodes were presented in HP (denoted as C0). In the second condition, denoted as C9, both services were presented in HP for the first two days followed by two days LP and so forth. This resulted in 8 days presented HP versus 6 days presented in LP. While C0 replicates the baseline condition of Möller et al. (2011), C9 should lead to a reduction in multi-episodic judgments.

After finishing an episode, participants assessed the episodic perceived quality on the 7-point CoCR scale (cf. Figure 2.1). The multi-episodic perceived quality for each service was assessed on the 4th, 7th, 10th, and 14th day directly after finishing the daily episode with the VoD service (same scale).

For the speech telephony service, the speech codec G.722 was used with a packet length of 20 ms for HP and for LP. Neither FEC nor elaborate algorithms for PLC were applied. For PLC, only zero insertion was used. For LP, a random packet loss rate of 5% was inserted in addition. For the VoD service, the video was encoded with H.264 at a resolution of 720x576 px and the audio channel encoded with AAC (48 kHz, 165 kbit/s, stereo). The two performance levels for this service differed only in the video encoding bandwidth. For HP, a bandwidth of 2 Mbit/s and for LP a bandwidth of 125 kbit/s was applied.

In this experiment, participants used their own computer. For audio reproduction and recording, a *Logitech PC120* headset was provided to each participant. The speech telephony service was implemented as a VoIP service, whereas the VoD service was only simulated. The latter only played local preprocessed content. The video was always played in full screen independent of the actual screen size or resolution. A complete description of the technical setup can be found in the Appendix i (p. iii).

On the day before starting the 14 day usage period, an introductory session was conducted with each pair of participants. In this session, participants received detailed instructions for this experiment and provided their demographic data. Also, one episode with each of the two services need to be solved. These were presented in HP.

5.1.2 *Participants*

E4 was conducted in Berlin, Germany from August to September 2012. For each of the two conditions, 5 pairs successfully finished this experiment with an average age of 22.9 years ($\sigma = 3.1$), consisting of 11 male and 9 female participants. Of the 280 scheduled calls, 5 were

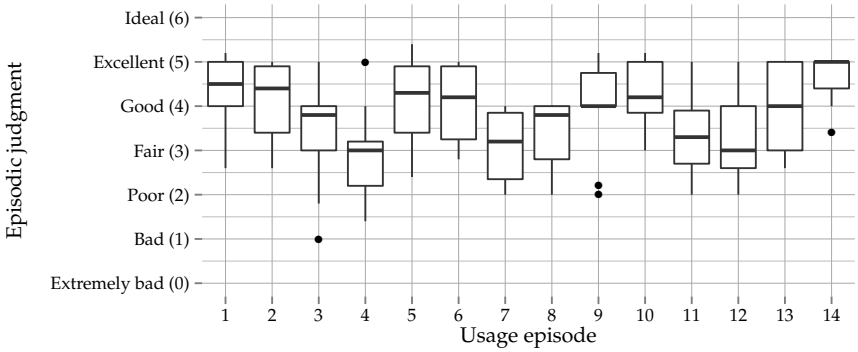

Figure 5.1: Multiple days (E4): episodic judgments for the VoD service in C9.

not conducted and 18 episodic judgments missing. For the VoD service, all episodes were conducted, but 13 episodic judgments and one multi-episodic judgment were missing. This is, in fact, an issue with regard to the evaluation, i. e., deriving a MOS, as desired by applying the defined-use method, because not all participants assigned to a multi-episodic condition were exposed to the same multi-episodic condition. However, due to the small sample size and the missing ground truth, no participant was removed from the data analysis.

5.1.3 *Data Analysis*

First, the episodic judgments are evaluated for the VoD service followed by a detailed analysis of the multi-episodic judgments. An analysis of the speech telephony service is omitted here, as the implemented setup failed to provide the desired performance levels. Thus, a MOS evaluation could not be conducted, as participants were not exposed to the same multi-episodic condition. Figure 5.1 shows the box plot for the episodic judgments for the VoD service in C9. In the following, first the episodic judgments are analyzed followed by the analysis of the multi-episodic judgments.

EPISODIC JUDGMENTS

For the VoD service, the episodic judgments for HP episodes between the two conditions are significantly different ($W = 5153.00$, $p = 0.018$). C0 resulted in a MOS of 4.2 (0.9) and C9 resulted in 4.0 (0.7). However, the difference is rather small and likely an artifact due to the between-subject design. For C0, a slight increase in episodic judgments, i. e., only presented HP episodes, is indicated from the first episode towards the final episode of this multi-episodic condition. The first episode is judged with 4.1 (0.7), whereas the last episode reaches 4.4 (0.6). However, this increase is not significant ($W = 24.50$, $p = 0.100$).

Table 5.2: Multiple days (E4): multi-episodic judgments for the VoD service. Reported as MOS with standard deviation in brackets.

Day	Multi-episodic judgment		
	C0	C9	Statistical difference
4	4.0 (0.6)	3.5 (1.4)	$W = 57.00$, $p = 0.622$
7	4.0 (0.6)	3.8 (1.3)	$W = 48.00$, $p = 0.909$
10	4.1 (0.6)	3.9 (1.0)	$W = 49.50$, $p = 0.743$
14	4.1 (0.6)	4.0 (1.0)	$W = 48.50$, $p = 0.939$

The episodic judgments of LP for C9 resulted in 3.3 (1.0). For C9, the two performance levels are significantly different ($W = 3566.50$, $p < 0.001$). However, the difference is smaller than desired as only a reduction of approximately 1 pt is introduced.

MULTI-EPISODIC JUDGMENTS

The multi-episodic judgments for the VoD service are shown in Table 5.2. For C0, no significant differences in multi-episodic judgments are found between the four measurements ($H(3) = 0.7099$, $p = 0.8709$). Also for C9, no significant differences are observed ($H(3) = 0.9195$, $p = 0.8207$). Even between the two conditions, no significant differences between multi-episodic judgments are found (cf. Table 5.2).

5.1.4 *Discussion*

E4 was only partly successful. It could be shown that the defined-use method could be applied successfully and practical knowledge about how to conduct a field experiment acquired. However, the results are limited with regard to multi-episodic perceived quality. Most prominent is the failure of the speech telephony service. This system did not provide the desired performance levels, as it failed to cope with temporary network limitations. This technical failure prevented that participants were exposed to the same multi-episodic condition, and thus a MOS evaluation could not be conducted. In addition, some pairs of participants found it very difficult to conduct two calls daily with each other. Here, it was problematic to find the time slots and embed them into the participants' daily life.

Also the results for the VoD service are limited although the system worked as desired. With regard to the multi-episodic judgments, no significant differences between the two conditions could be found. In fact, the final judgments are nearly identical between the two conditions. This indicates that the applied performance for LP was not severe enough. This is also indicated by the rather small difference

between episodic judgments for HP and LP. Nevertheless, the results of C0 are as expected, i. e., multi-episodic judgments are at a similar level as the episodic judgments. This is in line with Möller et al. (2011). In addition, a slight increase of episodic judgments over the usage period is indicated. The reason for this is unknown and could not be deduced. With regard to the implementation of a prediction model for multi-episodic judgments, the the gathered data is insufficient due to the failure of the speech telephony service and the limited number of multi-episodic conditions.

5.2 EXPERIMENT E5

Based on the practical insights of E4, in E5 only media consumption was used.[19] Here, a VoD service and an AoD service were used. Similar to E4, the services needed to be used over a usage period of 14 days. In this experiment, the impact of increasing the number of performance changes between episodes on multi-episodic judgments was investigated. The number of LP episodes (i. e., 6) was kept constant. In this experiment, two multi-episodic conditions were evaluated: C9 (as in E4), which presents three changes of performance from HP to LP, and C10, which presents four changes. C10 presents the 2nd, 4th, 5th, 9th, 12th, and 13th day in LP and all other days in HP.

Each of the two services needed to be used once per day. The AoD service needed to be used between 7 am and 1 pm, and the VoD service needed to be used between 5 pm and 11 pm. The two services were presented to participants in different multi-episodic conditions, i. e., C9 for AoD and C10 for VoD as well as C10 for AoD and C9 for VoD.

For this experiment, all necessary equipment, i. e., a mobile phone and a pair of headphones, was provided to the participants. This eliminated differences of equipment as confounding factor and also reduced the complexity of the setup. For the AoD service, an audio book with speech-only content was used. For HP, this was encoded with AAC (44.1 kHz, 320 kbit/s, stereo) and for LP in GSM-FR. Here, the audio book *City of the Beasts* from Isabelle Allende was used as in E3. The audio book was cut into individual scenes with a duration of 12 min to 17 min per episode. For the VoD service, scenes of *The Big Bang Theory* (Season one, BluRay version, German) were used. Here, meaningful scenes were cut, resulting in a duration of 8 min to 12 min per episode. The audio signal was treated similar to the AoD service. The video signal was encoded at the native resolution of the mobile device (800x480 px; 4.3 inch) with H.264 (25 FPS, two-pass). For HP, a video encoding bandwidth was set to 3 Mbit/s and for LP to 0.25 Mbit/s. Degradations on audio and video were inserted to increase the perceptual difference between HP and LP. Details of the technical setup are described in the Appendix i (p. iii).

19 The results of E5 are published in Guse et al. (2014).

Episodic judgments were taken on the 7-point CoCR scale after every episode. Multi-episodic judgments per service were taken at the 2nd, 5th, 8th, 11th, and 14th day. This judgment was taken after finishing the daily episode with the VoD service.

On the day before starting the 14 day usage period, an introductory session was conducted with every participant. Similar to E4, one HP episode with each service was presented in the introductory session.

5.2.1 *Participants*

This experiment was conducted in Berlin, Germany from June until August 2013. 21 participants (11 female, 10 male) took part in this experiment with an average age of 27.8 years ($\sigma = 4.0$). 11 participants were assigned to the AoD service with C9 and the VoD service with C10 and 10 vice versa. Participants received 40 EUR as compensation.

Out of the overall 294 usage episodes, 9 were not conducted for the AoD service and 13 not conducted for the VoD service. 6 multi-episodic judgments were missing. Similar to E4, no participant was removed from the data analysis due to the lack of ground truth and the small sample size.

5.2.2 *Data Analysis*

In this experiment, no issues with the technical setup were observed. In the following, the data of all participants is analyzed, starting with the episodic judgments followed by the multi-episodic judgments.

EPISODIC JUDGMENTS

The AoD service resulted for HP episodes in a MOS of 4.7 (0.7) and for LP episodes in 2.1 (0.9). Significant differences of the episodic judgments between the two conditions are neither observed for HP ($W = 3133.50$, $p = 0.533$) nor for LP ($W = 1938.50$, $p = 0.671$). The episodic judgments are significantly different between HP and LP ($W = 19348.00$, $p < 0.001$).

The VoD service resulted for HP episodes in a MOS of 4.8 (0.6) and for LP episodes in 2.1 (0.9). Neither episodic judgments of HP ($W = 3058.00$, $p = 0.716$) nor LP ($W = 1582.50$, $p = 0.155$) are significantly different between the two conditions. The episodic judgments are significantly different between the two performance levels ($W = 19212.50$, $p < 0.001$).

For HP episodes, both services are judged significantly different ($W = 11207.00$, $p = 0.031$). In fact, the difference is rather small (approximately 0.1 pt). It is thus assumed not to affect the multi-episodic hypotheses testing. For LP, no significant difference is found ($W = 7521.50$, $p = 0.885$). In Figure 5.2, box plots of the episodic

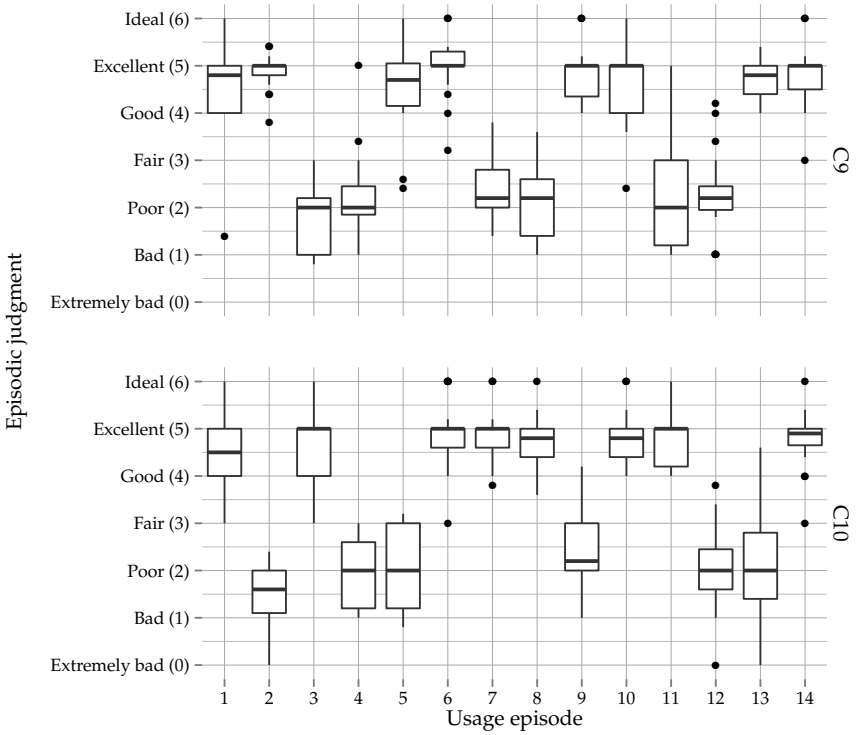

Figure 5.2: Multiple days: box plots of episodic judgments in E5 for C9 (top) and C10 (bottom).

judgments for both conditions are presented without discriminating by service.

In this experiment, no increase of episodic judgments for HP episodes could be observed as reported by Möller et al. (2011) and found in E4. However, statistical testing is omitted due to the small sample size.

MULTI-EPISODIC JUDGMENTS

The multi-episodic judgments for both services and conditions are shown in Table 5.3. For the AoD service, no significant differences between multi-episodic judgments are observed for C9 ($H(4) = 6.8676$, $p = 0.1431$). For C10, a significant difference is found ($H(4) = 10.2563$, $p = 0.0363$). However, a post-hoc test does not find significant differences for C10 ($p \geqslant 0.055$). For the AoD service, comparing the multi-episodic judgments between the two conditions does not indicate large effects (see Table 5.3).

For the VoD service, no significant differences between multi-episodic judgments are found for C9 ($H(4) = 3.2427$, $p = 0.5181$). For C10, significant differences are found ($H(4) = 22.8848$, $p < 0.001$). A post-hoc test shows that the multi-episodic judgment of the 2nd day is significantly different to all following multi-episodic judgments

Table 5.3: Multiple days (E5): multi-episodic judgments. Reported as MOS with standard deviation in brackets.

Day	Multi-episodic judgment			
	AoD		VoD	
	C9	C10	C9	C10
2	4.0 (0.7)	4.4 (0.8)	3.7 (0.4)	5.2 (0.3)
5	3.0 (0.7)	3.7 (0.7)	3.7 (0.4)	3.4 (0.7)
8	3.6 (0.9)	3.6 (0.6)	3.9 (0.4)	3.4 (0.9)
11	3.6 (0.8)	3.3 (0.6)	3.8 (0.4)	3.5 (1.0)
14	3.6 (0.9)	3.5 (0.6)	3.5 (0.5)	4.0 (0.7)

($p < 0.006$). Between the two conditions, only the multi-episodic judgment of the 2nd day is significantly different ($W = 0.00$, $p < 0.001$).

5.2.3 *Discussion*

In E5, the implemented system was able to provide the desired performance levels in a reliable manner as reflected by the episodic judgments. As long as only HP episodes were presented, the following multi-episodic judgment remained on a similar level as the episodic judgments for both services (C10). This is in line with Möller et al. (2011) and also E4. It is notable that the episodic judgments are not very different between the two services. A reason for this might be that both services presented the audio modality with the same codec configuration and thus similar degradations.

As desired, the presentation of LP episodes affected the following multi-episodic judgments. However, the two conditions resulted only in limited variation of multi-episodic judgments after the first LP episode was presented. Except for the first multi-episodic judgment of the VoD service, no differences between the two conditions could be observed. In C9, even the presentation of three consecutive days in HP indicated only a *slight*, non-significant increase. This suggests a longer integration interval for the multi-episodic judgments than three days, i. e., prior experiences still affected the judgments. That no difference between the two conditions could be observed, indicates that the number of day-wise performance changes did not affect the multi-episodic quality formation process and thus suggests that H3 is false. The results of E5 can be used for model verification only, as the two conditions yielded very similar results.

5.3 EXPERIMENT E6

E6 is a follow-up to the one-session experiments, but extends the usage period to 6 days with one usage episode per session. The main goal of this experiment was to investigate if the effects observed in one session can also be observed in longer usage periods. Besides interesting knowledge, this is an important aspect for the implementation of prediction models. In E6, a AoD service is used to investigate a subset of the presented hypotheses (cf. Section 3.2). Here, the impact of the number of LP episodes (H1), the position of LP episodes (H2), and consecutive versus non-consecutive LP episodes (H3) were investigated.

5.3.1 *Design*

The experimental design of E6 is similar to the one-session experiments. On each of the 6 days, the AoD service needed to be used twice per day. For the investigation of the three hypotheses, six multi-episodic conditions were created. Here, the performance was varied on a per day basis, i. e., the two episodes of the same day were presented with the same performance level. The multi-episodic conditions are similar to the one-session experiments and are thus denoted with the *same* abbreviations (see Section 4.1.1). In all conditions, the first three days were presented in HP. LP episodes were only presented from the 4th day to the 6th day. The conditions are C1, C3, C4, C5, C6, and C8. C1 and C3 present either the 4th or the 6th day in LP. C4, C5, and C8 present two days between the 4th and the 6th day in LP. C6 presents all usage episodes on these three days in LP.

For the investigation of H1, the results of C3, C5, and C6 can be compared. H2 can be evaluated by comparing the results of C1 and C3 as well as C4 and C5. Finally, H3 is evaluated by comparing the results of C8 with C3 and C5.

As in E3 and E5, the audio book *City of the Beasts* from Isabel Allende was used. In E3, episodes with a duration of approximately 3 min were used, whereas E5 used 12..17 min. For E6, a duration of 6..8 min was chosen. This should enable participants to focus on the content while limiting their effort. Again, the audio book was cut, so individual scenes were self-contained. The scenes were presented in the chronological order (one scene per episode).

Before starting the 6 day usage period, an introductory session was conducted. Here, participants received all necessary information about the experiment and demographic data was collected. Then, participants were presented typical speech telephony degradations, i. e., training, in the same way as in the one-session experiments (cf. Section 4.1.3). Finally, participants used the AoD service for two episodes to ensure that they understood how to use the AoD service.

During the multi-episodic part of this experiment, the first episode of a day needed to be conducted between 7 am and 1 pm and the second episode between 3 pm and 10 pm. Episodic judgments and multi-episodic judgments were taken on the 7-point CoCR scale. Multi-episodic judgments were collected after the second episode of the 3rd day and the 6th day. In fact, the judgment after the 3rd day is thus only based on HP episodes. Thus, it can be used as reference point to assess the impact of the presented LP episodes on the final multi-episodic judgment. This is similar to the one-session experiments and in the following denoted also as *reference*. After each episode, two content-related questions were presented to force participants to pay attention to the content. Here, the correct answer out of three options needed to be selected. This allows to evaluate if participant had experienced an episode and thus could answer the questions correctly.

On the day after finishing the usage period, a final interview was conducted. Here, participants were interviewed about issues with the technical system.

In this experiment, participants used their own computer and their own pair of headphones. Participants could access the AoD service via the Internet using a HTML5-capable web browser. For HP, the source material (CD, 44.1 kHz, stereo) was encoded with MP3 (192 kbit/s). This was necessary to enable media distribution via Internet, as uncompressed audio data is not suited for this. In fact, the bitrate was selected to produce no audible impairments for the used speech-only content. For LP, the content was first encoded with LPC-10 before encoding it finally with MP3. A detailed description of the implemented system is given in the Appendix i (p. iii).

5.3.2 *Participants*

E6 was conducted in Berlin from September until November 2015. Participants were required to have normal hearing capabilities. This experiment was conducted with 57 female and 38 male participants aging from of 18 to 33 years ($\mu = 25.8$, $\sigma = 4.0$). Participants received 20 EUR as compensation. In this experiment, all usage episodes were conducted and all questionnaires filled. Here, participants were informed by email, when a usage episode should be conducted, and a digital questionnaire system was used.

Similar to the laboratory experiments, each participant was individually checked for inconsistent episodic judgments. A participant is considered inconsistent if more than two episodic judgments exceed the 1.5 × *interquartile range* of the performance levels of this condition. None of the participants fulfilled this criteria. In addition, the content-related questions were evaluated. Here, it was required that participants should answer at least 50% of the questions correctly to assume they followed the experimental instructions. Out of the 24 questions,

Table 5.4: Multiple days (E6): overview on conditions. Reported as MOS with standard deviation in brackets.

Condition	LP days	Participants	HP	LP
C1	4	16	4.5 (0.9)	1.3 (0.8)
C3	6	15	4.7 (0.7)	1.3 (0.8)
C4	4..5	14	4.5 (0.7)	1.3 (0.5)
C5	5..6	18	5.0 (0.8)	0.9 (0.5)
C6	4..6	15	4.6 (0.6)	1.2 (0.7)
C8	4 and 6	16	4.5 (0.6)	1.2 (0.7)

participants answered on average 20.8 questions ($\sigma = 2.7$) correctly. One participant who participated in C8 is excluded from data analysis, because only 8 questions were correctly answered.

5.3.3 *Data Analysis*

In the following, the data of E6 are analyzed. First, the potential impact of the between-subject design is evaluated. Then, the multi-episodic judgments are evaluated with regard to the three investigated hypotheses.

5.3.3.1 *Consistency*

An impact of the between-subject design is investigated by evaluating the episodic judgments of HP and LP. Table 5.4 shows the episodic judgments for all conditions. A significant difference is found for HP ($H(5) = 33.4978$, $p < 0.001$). A post-hoc tests shows that C5 is significantly different to all other conditions ($p < 0.05$). In addition, C3 is significantly different to C4 and C8 ($p < 0.02$). For LP, significant differences between conditions are also found ($H(5) = 18.2748$, $p = 0.0026$). A post-hoc tests shows that C5 is significantly different to C1, C4, and C8 ($p < 0.05$). It must be noted that episodic judgments of C5 resulted in the highest MOS for HP and in the lowest MOS for LP. A detailed analysis did not yield a reason for the difference, and it is thus considered an artifact due to the between-subject design. Box plots for all conditions can be found in the Appendix iii (p. xi).

For the multi-episodic judgment after the 3rd day, no significant differences between conditions are observed ($H(5) = 5.2111$, $p = 0.3907$). This indicates that as long as only HP episodes were presented, the between-subject design did not affect multi-episodic judgments.

Table 5.5: Multiple days (E6): multi-episodic judgments after the 6th day for H1. Reported as MOS with standard deviation in brackets.

Condition	LP episode(s)	Multi-episodic judgment
Reference (HP only)	-	4.7 (0.6)
C3	6	3.6 (0.6)
C5	5..6	2.5 (1.0)
C6	4..6	2.4 (0.7)

5.3.3.2 *H1: Number of LP Episodes*

In H1, it is assumed that increasing the number of LP episodes before a multi-episodic judgment results in a decrease of this judgment. This hypothesis can be evaluated by comparing the final multi-episodic judgment between the *reference*, C3, C5, and C6. These present none, one, two, and three days in LP before the multi-episodic judgment. Table 5.5 shows the final multi-episodic judgment and the reference for these conditions.

C3, C5, C6, and the reference are significantly different ($H(3) = 68.3657$, $p < 0.001$). A post-hoc test finds that the reference is significantly different to all three conditions ($p < 0.001$). In addition, C3 and C5 ($p = 0.002$) as well as C3 and C6 are significantly different ($p < 0.001$). For C5 and C6, no significant difference is found ($p = 0.366$).

The results show that increasing the number of LP days directly before the multi-episodic judgment results in a reduction of this judgment. Here, a reduction of approximately 1 pt for none to one and one to two days is observed. However, no further decrease can be observed if three days are presented in LP. It must be noted that the multi-episodic judgment remains 1 pt higher than the episodic judgments of LP episodes. With regard to H1, the results are in line the one-session experiments E1 and E2a. For both usage periods, the multi-episodic judgment decreases until two episodes/days were presented in LP. Then, the judgment remains on the same level above the episodic judgments of LP. Thus, H1 can only be partly accepted. However, the underlying reason for the observed saturation could not be derived from this experiment.

5.3.3.3 *H2: Position of LP Episode(s)*

In H2, it is assumed that presenting HP episodes after LP episodes reduces the negative impact on the directly following multi-episodic judgment. This can be investigated for one day in LP with C1 and C3 as well as for two days in LP with C4 and C5. Table 5.6 shows the final multi-episodic judgment for these four conditions.

Table 5.6: Multiple days (E6): multi-episodic judgment after the 6th day for H2. Reported as MOS with standard deviation in brackets.

Condition	LP episode(s)	Multi-episodic judgment
C1	4	4.1 (0.7)
C3	6	3.6 (0.6)
C4	4..5	3.0 (1.1)
C5	5..6	2.5 (1.0)

With regard to the final multi-episodic judgment neither C1 and C3 ($W = 138.50$, $p = 0.136$, one-sided) nor C4 and C5 ($W = 151.00$, $p = 0.087$, one-sided) are significantly different. Still, in both cases an effect of position is indicated and thus a recency effect might have been observed. However, the results of E6 are inconclusive and thus H2 can neither be accepted nor be rejected.

5.3.3.4 *H3: Consecutive vs. Non-consecutive LP Episodes*

In H3, it is assumed that consecutive LP episodes yield a better multi-episodic judgment than the same number of LP episodes presented non-consecutively. This hypothesis can be evaluated by comparing C4 and C5 with C8. C4 and C5 present each two days LP consecutively, whereas C8 presents the 4th and the 6th day in LP. Table 5.7 shows the final multi-episodic judgment for these conditions. These three conditions are not significantly different with regard to the final multi-episodic judgment ($H(2) = 2.3809$, $p = 0.3041$).

Thus, H3 must be rejected, because no significant difference is observed. It must thus be concluded that the non-consecutive case is judged not different than the consecutive case. In fact, the slight improvement in the multi-episodic judgment of C8 compared to C5 can also be explained by a recency effect due to the earlier presentation of the first LP episodes. This observation is similar to E2a.

Table 5.7: Multiple days (E6): multi-episodic judgment after the 6th day for H3. Reported as MOS with standard deviation in brackets.

Condition	LP episode(s)	Multi-episodic judgment
C4	4..5	3.0 (1.1)
C8	4 and 6	2.7 (0.4)
C5	5..6	2.5 (1.0)

5.3.4 *Discussion*

In this experiment, three hypotheses were investigated complementing the one-session experiments. In fact, an absolute comparison between E6 and the one-session experiments cannot be conducted due to the differences in the experimental designs. Nevertheless, in E6 similar effects could be observed or were indicated.

For H1, a decrease in the final multi-episodic judgment could be observed if up to two days were presented in LP. Presenting three consecutive days in LP did not result in a further decrease. This closely resembles the findings of E1 and E2a.

With regard to varying the position of one or two days in LP, a recency effect is indicated (H2). For both cases, a (non-significant) increase of the multi-episodic judgments is indicated if HP episodes are presented directly before the final judgment. This is in line with the findings of E1 and partly with E2a. In the latter, no effect could be observed for one LP episode but for two consecutive LP episodes.

H3 must be rejected based on E6. Similar to E2a, the final multi-episodic judgment remains rather close. The slight difference that is indicated between the conditions could also be explained by a recency effect. In fact, this might even be an anomaly due to the between-subject design. It must be concluded that increasing the number of performance level changes did not produce an observable effect.

5.4 CONCLUSION

E6 was conducted based on the practical knowledge gained in E4 and E5. E4 and E5 showed that multi-episodic perceived quality can be investigated in field experiments by applying the defined-use method. Although the results were very limited, it was found that a media-on-demand service can be set up in a reliable manner. Besides avoiding technical complexity, also employing pairs of participants has been found difficult in a field experiment (E4). Especially, the effort for each pair of participants to conduct the episodes together seems to be limiting. These limitations could be overcome by using media-consumption tasks. In addition, this allows to avoid the impact of varying user behavior on the duration of usage episodes. This lead to the design of E6. This experiment was also inspired by the one-session experiments.

For H1, very similar results to the one-session experiments were observed. A decrease in the final multi-episodic judgment could be observed if more LP episodes were presented. In addition, also an effect of saturation could be observed. This indicates that the integration interval of the formation process is longer than 3 days, as previous HP episodes still affected the final multi-episodic judgment. With regard to H2, i. e., impact of varying position of LP episodes, an

effect is indicated (non-significant). Here, a recency effect seems to occur. This is line with the results of E1 and E2a.

With regard to H3, the results are similar to E2a. Here, no significant difference is observed between the non-consecutive and consecutive presentation of two LP days. The indicated difference might also be attributed to a recency effect. Thus, the non-consecutive presentation seems to be judged similar to the consecutive presentation or the effect size seems to be rather small. Therefore, H3 must be rejected.

It must be noted that in E6 effects were observed or indicated that are similar to the one-session experiments. This indicates that the presentation of episodes itself has a higher impact on the quality formation process than the actual time between episodes. However, a direct comparison of the results of these experiments cannot be conducted due to the differences in the experimental designs, i. e., tasks, service types, and duration of usage episodes.

In difference to E4 and E5, E6 provides a large data set (6 conditions, 95 participants). In the following chapter, this data set forms the basis for the development and the evaluation of models for the prediction of multi-episodic judgments based on episodic judgments.

6

PREDICTION OF MULTI-EPISODIC JUDGMENTS

An initial approach on the prediction of multi-episodic judgments has been conducted by Möller et al. (2011). The model[20] proposed here is to average all *prior* episodic judgments to predict a multi-episodic judgment. Möller et al. (2011) evaluated the precision of this predictor with their 14 day experiment (cf. Section 2.4.4). Instead of evaluating the prediction accuracy for MOS, the prediction accuracy for each individual participant was evaluated for all three multi-episodic judgments taken in this experiment. Here, multi-episodic judgments were taken after the 2nd, 7th, and 14th day. With regard to the five conducted multi-episodic conditions, it could be shown that the predictor is precise for the 2nd day and precision decreases for later multi-episodic judgments.

This chapter starts with an overview of the effects observed in the here presented experiments that seem relevant for the prediction of multi-episodic judgments. Subsequently, model types are presented that seem suited for the prediction. The prediction accuracy for these model types is evaluated using only E1, E2a, and E6. These experiments yielded large and in itself consistent data sets.[21] The other here presented experiments (E2b, E3, E4, and E5) are not considered suitable for the implementation of a prediction model, as only a very limited set of conditions was investigated. In difference to Möller et al. (2011), who predicted multi-episodic judgments using episodic judgments per individual participant, the goal is here the prediction of the multi-episodic MOS based on the episodic MOS. Predicting individual judgments might be desirable, but the conducted experiments did not collect sufficient data to be able to explain differences between individual participants. In fact, the defined-use method was applied, so a MOS could be derived that reflects the judgment of the *average actor*.

The here presented modeling is necessarily limited to the development of best fitting predictor(s), as no data sets for cross-validation

20 The term *model* is used in the following in terms of curve fitting. A model is here a mathematical function, which might be parameterizable, that maps an input space to an output space.

21 An approach towards modeling based on the experiment of Möller et al. (2011) and E5 is published in Guse et al. (2014). However, these two experiments are omitted here for the development of prediction models due to the limited effects on multi-episodic judgments.

were available. Furthermore, the conducted experiments differ in the experimental design and also observed effects. Thus, models created for one experiment cannot be verified validly using data from the other experiments.

6.1 EFFECTS ON MULTI-EPISODIC JUDGMENTS

In the conducted experiments, several effects have been observed. First, it must be noted that multi-episodic judgments are very similar to episodic judgments if no degraded episodes are presented. This has been observed by Möller et al. (2011) and also in the here presented experiments. If no LP episodes were presented, the experiment of Möller et al. (2011) and E4 indicated a slight increase of episodic judgments from the first to the last episode in a multi-episodic condition. However, the indicated effect is rather small and the actual reason for this could not be deduced. It is thus not considered for the implementation of a prediction model.

The largest observed effect on multi-episodic judgments is the decrease due to an increased number of LP usage episodes (H1). This has been investigated in E1, E2a, and E6. Here, a decrease is observed until saturation occurred, i. e., the multi-episodic judgment did not decrease further (saturation effect). This occurred if more than two LP episodes/days were presented. In fact, the final multi-episodic judgment remained above the episodic judgments of LP. This shows that the final multi-episodic judgment is still affected by previous HP episodes, i. e., the integration interval is longer than 3 episodes or 3 days, respectively. In addition to the integration interval, this indicates that one or all of the three LP episodes/days had a reduced impact on the final multi-episodic judgment.

In addition, a position effect has been observed (H2). This has been investigated in E1, E2a, and E6. An impact of position and thus a recency effect could be observed in E1 and E2a. Here, the impact of LP episode(s) on the following multi-episodic judgment decreased the more HP episodes were presented afterwards. While E1 showed such an effect in both cases, i. e., for one and two LP episodes, it was only present in E2a for two LP episodes. This might be attributed to the passive usage situation. In E1, a two-party conversation was used, whereas E2a applied a third-party listening task. Although not statistically significant, a recency effect was indicated in E6 for both cases, i. e., one or two days presented in LP.

In E2a and E6, the impact of non-consecutive LP presentation was also investigated (H3). An effect could not be observed in both experiments and thus the number of performance changes is assumed to be neglectable for the prediction of multi-episodic judgments. In fact, the small, potential difference might also be explainable by a recency effect.

A peak effect is not considered for modeling, as such an effect could not be observed in E1 (H4). Even if in C7 a peak effect occurred, the influence on the multi-episodic judgment seems to be rather small.

For one session, also the recovery of negatively affected multi-episodic judgments was investigated (H5). This was studied with two conditions in E1. Here, 9 episodes were presented while non-HP episodes were presented as the 5th and the 6th episode. The multi-episodic judgment after the 6th episode showed a negative effect, and the judgments after the 3rd and 9th episode were significantly different. Thus, in both conditions a negative impact is still present in the final multi-episodic judgment. This finding can be explained by a recency effect or by the increased number of HP episodes.

For one session, a duration neglect was investigated in E3 (H6). Here, doubling the duration of one LP episode did not negatively affect the following multi-episodic judgment. In fact, even the episodic judgment was not affected negatively. It is thus concluded that a duration neglect for episodic judgments as well as multi-episodic judgments was observed. Therefore, it can be concluded that the duration of an episode without macroscopic fluctuations does not have to be accounted for the prediction of multi-episodic judgments.

Finally, in E2b the independence of multi-episodic judgments of two services has been investigated (H7). Here, the presentation of a second service did not lead to a measurable effect on the multi-episodic judgments of the first service. This was found for the presentation of the second service with and without LP episodes. Thus, the presence of a second service must not be considered for the prediction of multi-episodic judgments of the first service that is under multi-episodic assessment.

6.2 TYPES OF MODELS

For the implementation of a prediction model, the modeling approach of Möller et al. (2011) is extended. They proposed to the use the average of all prior episodic judgments to predict multi-episodic judgments. This provides an accurate prediction for earlier multi-episodic judgments. However, prediction accuracy decreases if later multi-episodic judgments are to be predicted.

In this thesis, models based upon the weighted average are proposed to increase prediction accuracy. Such models allow to assign an individual weight to each episodic judgment and thus model the individual impact on the multi-episodic judgment to be predicted. In the following, episodic judgments are denoted as e_i and multi-episodic judgments are denoted as m_n. The index i denotes the episode and the index n denotes the episode after which a multi-episodic

judgment was taken. The weight for an episodic judgment is denoted as a_i. Thus, the prediction model for $\hat{m}_n$ is defined as:

$$\hat{m}_n = \frac{\sum_{i=1}^{n} a_i * e_i}{\sum_{i=1}^{n} a_i}. \tag{6.1}$$

The weighted average is in itself a rather simple model, as it is only parametrized with a weight function. However, selecting a suitable weight function is not a simple task, because overfitting is an issue. Following Occam's Razor, a lower degree of freedom for a weight function is preferable. Here, two weight functions are proposed that both can account for a recency effect.

The first weight function is a window function, which is in the following denoted as *WI*. This function is parametrized by the window parameter w. All episodic judgments in this window are assigned a weighting factor of 1 and all episodes before a weighting factor of 0 (Equation 6.2).

$$WI : a_i = \begin{cases} 1, & \text{if } i - n + w > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.2}$$

Here, w is limited to $w \in \mathbb{N}$ and $0 < w \leqslant n$. Setting $w := n$, this model type becomes the average over all prior episodic judgments, i. e., the model proposed by Möller et al. (2011).

In fact, a static window is a rather unlikely case for the formation process of multi-episodic judgments, as the importance of episodic judgments is considered only as binary. To overcome this a linear function (denoted as *LI*) is proposed. Here, the weight for usage episodes decreases linearly for an increasing distance to the multi-episodic judgment (Equation 6.3).

$$LI : a_i = \begin{cases} i - n + w, & \text{if } i - n + 2 * w > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6.3}$$

Here, w is also limited to $w \in \mathbb{N}$ and $0 < w \leqslant n$. However, the actual window is increased by $2 * w$, so the very first episodic judgment can be considered with a maximum weight of $\geqslant 0.5$.[22] Limiting w for both models to the same set, allows to compare the accuracy of both models directly.

Employing a weighted average using the two presented weight functions is expected to increase prediction accuracy, because a recency effect can modeled. However, for one observed effect, a weighted average with the two proposed weight functions will necessarily produce a deviation. In case of the observed saturation effect, both weight

22 Please note that normalization (required due to the weight function) is handled by the weighted average itself.

functions will be too negative. Here, the presentation of the 4th, 5th, and 6th episodes/days in LP (C6) resulted in a similar multi-episodic judgment than the presentation of the 4th episode in HP and the 5th and 6th episodes/days in LP (C5). The modeling approach for this specific case is presented and evaluated in Section 6.3.3.

PREDICTION ACCURACY In the following, a *model* denotes the combination of the selected weight function with the selected value for the parameter w. The prediction accuracy of a model is evaluated using the *Root Mean Square Deviation (RMSD)*:

$$\text{RMSD} : \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{x}_i)^2}\,. \tag{6.4}$$

In a perfect case a RMSD of zero, i. e., no deviation, can be achieved. If two models achieve a very similar RMSD, then the one with the smaller w is preferable, because this model requires less historic information to achieve a similar prediction accuracy. Overall, a model is preferable that explains a higher number of conditions rather than individual conditions only.

6.3 EVALUATION

In the following, the evaluation of the prediction accuracy for the proposed model types is done individually for E1, E2a, and E6. This is necessary as all three experiments were different with regard to usage situation and usage period. This evaluation is conducted using the episodic MOS to predict the multi-episodic MOS. First, the multi-episodic judgment for HP-only episodes, i. e., the reference, is evaluated. Following, the prediction accuracy is evaluated for multi-episodic judgments affected by LP episode(s). Finally, the potential improvement of accounting for a saturation effect is investigated.

6.3.1 *One Session: E1 and E2a*

EXPERIMENT E1 In E1, a two-party conversation was investigated with 6 and 9 episodes. With regard to the multi-episodic judgment after the 3rd episode, i. e., HP only, both model types perform very similar. Figure 6.1 shows the RMSD for both model types with regard to w for each condition. The black dashed line represents the average RMSD over all conditions. For both model types the accuracy remains nearly independent of w. It is notable that the accuracy depends on the condition, i. e., C7 yields a very low RMSD, whereas C1 is far higher. This is likely an artifact due to the between-subject design.

With regard to the multi-episodic judgment after the 6th episode, both model types perform slightly differently. Figure 6.2 shows the

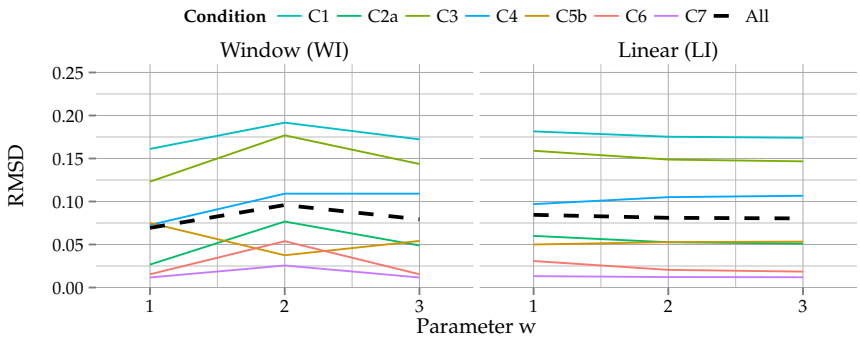

Figure 6.1: One session (E1): multi-episodic prediction accuracy for HP only episodes (3rd usage episode).

RMSD for the 6th episode for both model types. Here, WI performs better while increasing w to 4 (0.23). However, the prediction accuracy depends on the considered condition. For example, C3 is far off for $w = 1$ and improves until $w = 3$, whereas C4 is best predicted with $w = 5$. LI outperforms WI in prediction accuracy and is, furthermore, more robust. For LI, the best accuracy is achieved for $w = 2$. All conditions except C4 and C6 yield here the minimal RMSD. In fact, C4 and C6 reach their minimum at $w = 3$. Considering all conditions, the minimal RMSD for LI is achieved at $w = 2$ (0.12). This is close to the prediction accuracy for presenting HP episodes only.

With regard to the recovery (H5), two conditions were investigated. The RMSD is shown for both weight functions in Figure 6.3. WI is very precise for $w = 4$ but is otherwise far off. LI performs well for C5b with $w \geqslant 3$, whereas it performs very similarly for all w in case of C7. Thus, LI is preferable, as it provides a higher robustness for the selection of w.

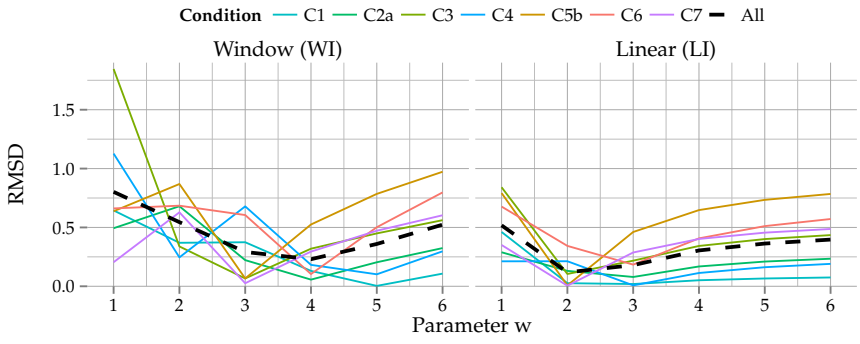

Figure 6.2: One session (E1): multi-episodic prediction accuracy for all conditions (6th usage episode).

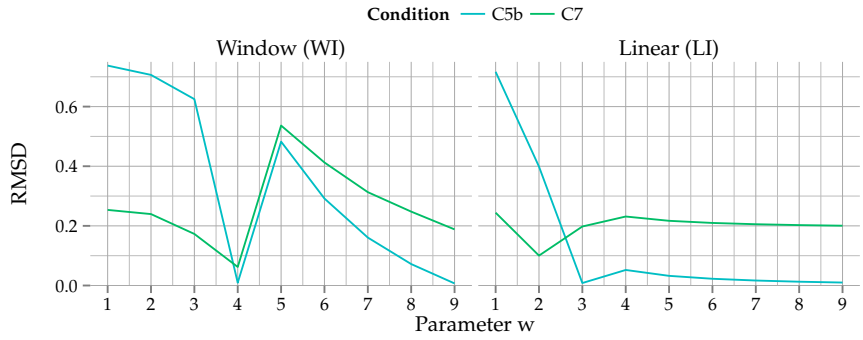

Figure 6.3: One session (E1): multi-episodic prediction accuracy for recovery (9th usage episode).

EXPERIMENT E2A E2a complements E1 with a passive usage situation while it shares the performance levels. With regard to the prediction of HP-only episodes, the results are slightly different compared to E1. The prediction accuracy for the multi-episodic judgment after the 3rd episode is shown in Figure 6.4. Here, the prediction accuracy improves if w is increased. However, this is mainly due to C4. The reason for this could not be determined. Omitting C4 shows similar results compared to E1, i. e., the prediction accuracy is not affected by the selection of w. The RMSD is similar for all w (~0.1). Here, no difference between WI and LI is observed.

With regard to the prediction of the multi-episodic judgment after the 6th episode, the results closely resemble E1. The RMSD is shown in Figure 6.5. For WI, the minimal RMSD is reached at $w = 4$ (0.18) while $w = 3$ (0.23) is rather close. For LI, the minimal RMSD is achieved at $w = 2$ (0.13). In fact, the found parameters are very similar to E1, and the RMSD behaves similarly for individual conditions.

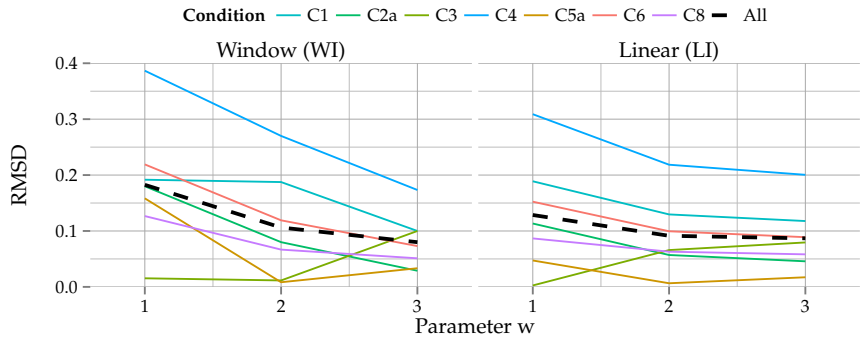

Figure 6.4: One session (E2a): multi-episodic prediction accuracy for HP only episodes (3rd usage episode).

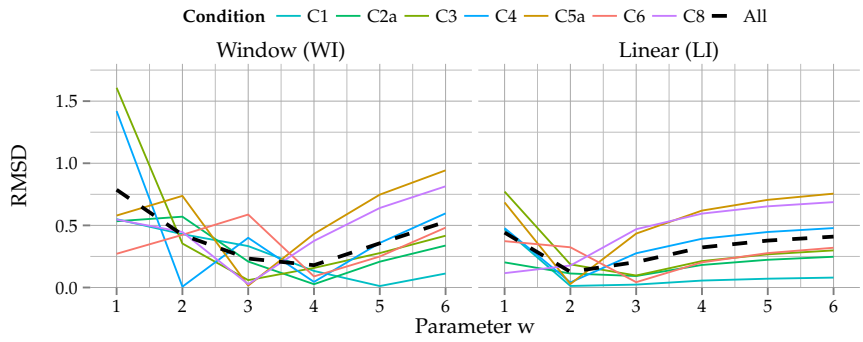

Figure 6.5: One session (E2a): multi-episodic prediction accuracy for all conditions (6th usage episode).

CONCLUSION For the prediction of multi-episodic judgments in one session, both weight functions perform similarly well. For both experiments, it is notable that similar values for w were found for each of the two weight functions. With regard to the prediction accuracy, LI is preferable over WI due to the better prediction accuracy. Moreover, LI seems to be more robust against improper selection of w. For both experiments, a minimal RMSD could be achieved for $w = 2$ in case of LI. For the prediction of the multi-episodic judgment after the 9th episode, which has only been investigated in E1, LI ($w = 3$) performs best. However, as only two conditions with regard to recovery were investigated (H5), adjusting w seems improper.

6.3.2 *Multiple Days: E6*

In E6, a usage period of six days was investigated for an AoD service. This service needed to be used twice per day. In this experiment, the first three days (six episodes) were presented in HP. Afterwards, the multi-episodic perceived quality was assessed. Prediction accuracy for this judgment is shown in Figure 6.6. Here, the prediction accuracy improves for an increasing w. This is more prevalent for WI than for LI. WI achieves its minimal RMSD with $w = 6$, i. e., all prior episodes. LI provides only a marginal decrease for $w \geqslant 3$.

With regard to the prediction of the multi-episodic judgment of the 6th day, both weight functions perform differently. Figure 6.7 shows the RMSD for both weight functions. While LI reaches a minimal RMSD at $w = 4$, WI achieves its minimal RMSD not until $w = 8$. In addition, LI achieves a minimal RMSD of 0.15, and WI only achieves a minimal RMSD of 0.26. With regard to E6, LI is preferable to WI, as a higher prediction accuracy is achieved. Furthermore, LI requires a smaller w while it provides a higher robustness for choosing w.

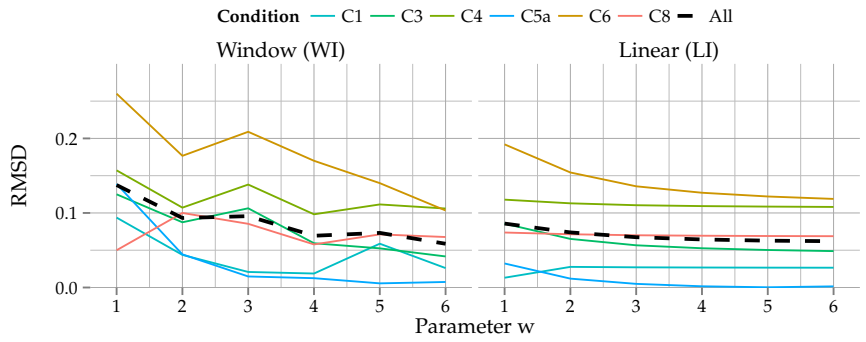

Figure 6.6: Multiple days (E6): multi-episodic prediction accuracy for HP only episodes (3rd day, i. e., 6th usage episode).

6.3.3 *Saturation Effect*

In all three experiments, a saturation effect could be observed. Here, the final multi-episodic judgment remained on the same level independent if two or three LP episodes/days were presented, i. e., C5[23] and C6 were not judged differently. In both cases, the multi-episodic judgment remained above the episodic judgments for LP episodes (approx. 1 pt). In fact, C5 and C6 only differ in the performance level of the 4th episode/day. For C5, this episode/day is presented in HP, whereas C6 presented this episode/day in LP. As both conditions were not judged differently, this suggests that the difference in performance level of the 4th episode/day does not seem to affect the formation process of the multi-episodic judgment. With regard to the applicability of the weighted average, this is problematic, as a model needs to produce a similar prediction based on different inputs, i. e., one HP episode/day and two LP episodes/days (C5) versus three LP episodes/days (C6). As the multi-episodic judgment remained above

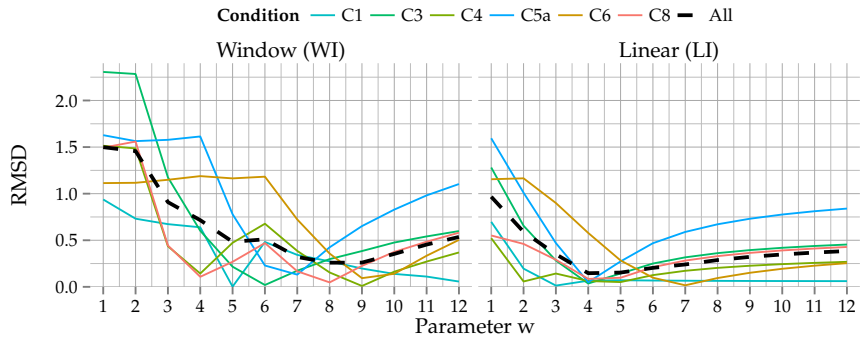

Figure 6.7: Multiple days (E6): multi-episodic prediction accuracy for all conditions (6th day, i. e., 12th usage episode).

23 In the following, C5 refers to C5b in case of E1 and to C5a in case of E2a.

the level of episodic judgments of LP episodes, at least the last four episodes/days must be considered in the case of C6, whereas C5 only requires at least three episodes/days. For all three experiments, LI and WI require a larger w to reach the best prediction accuracy for C6 compared to all other conditions. For E1 and E2a, $w = 3$ (LI) and $w = 4$ (WI) are required for C6, whereas C5 achieves its lowest prediction accuracy already at $w = 2$ (LI) and $w = 3$ (WI). For E6, $w = 7$ (LI) and $w = 9$ (WI) are required for C6, whereas C5 achieves its lowest prediction error at $w = 4$ (LI) and $w = 7$ (WI). It must be noted that the optimal value(s) of w for LI as well as WI in case of C5 are identical to the optimal w for all conditions. Thus, the overall prediction performance for both weight functions can be improved if w can be reduced to this value in case of C6. Although the underlying reason for the observed saturation effect could not be deduced, this can be achieved by *modifying* the episodic judgment(s) of the 4th episode/-day for C6, so it resembles C5. This modification can be achieved by replacing these judgment(s) by the average of episodic judgments of HP episodes. It is thus here proposed to change e_4 in case of C6 for E1 and E2a:

$$\tilde{e}_4 = 1/3 \sum_{i=1}^{3} e_i \,. \tag{6.5}$$

For E6, the episodic judgments of e_7 and e_8 are adjusted as follows:

$$\tilde{e}_{[7,8]} = 1/6 \sum_{i=1}^{6} e_i \,. \tag{6.6}$$

This is equivalent to extending $\hat{m}_n$ in the case of C6 with the addition of the term (here only shown for E1 and E2a):

$$\left(\sum_{i=1}^{3} e_i/3 - e_4\right) \cdot a_4 / \sum_{i=1}^{n} a_i \,. \tag{6.7}$$

This modified version of C6 is denoted as *C6 (adjusted)*. In the following, the prediction accuracy of the two weight functions for C5, C6, and C6 (adjusted) is presented. For E1 and E2a, this is shown in Figure 6.8 and Figure 6.9, respectively.

For E1 and E2a, this adjustment results in a shift of the minimal RMSD for the two weight functions. In case of WI, the minimal RMSD shifts from $w = 4$ to $w = 3$ for both experiments. This is also observed for LI. Here, the minimal RMSD shifts from $w = 3$ to $w = 2$. It is notable for both experiments that this adjustment leads to a similar shape of RMSD in case of C5 and C6 (adjusted). Furthermore, the minimal RMSD is reached earlier, but for increasing w the RMSD is worse for C6 (adjusted) than for C6.

With regard to E6, a similar observation is made. The minimal RMSD shifts from $w = 9$ to $w = 6$ for WI and for LI from $w = 7$ to $w = 4$. Here, C6 (adjusted) also resembles C5 closely (see Figure 6.10).

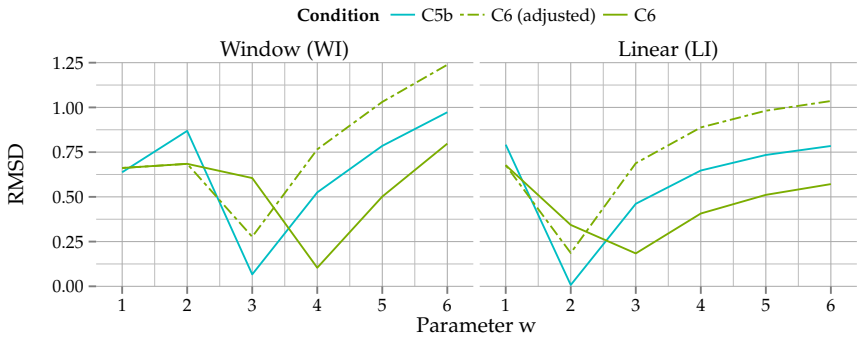

Figure 6.8: One session (E1): multi-episodic prediction accuracy for saturation effect (6th usage episode).

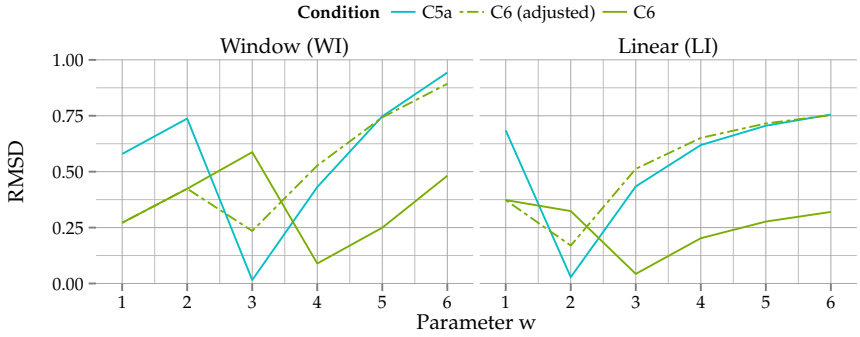

Figure 6.9: One session (E2a): multi-episodic prediction accuracy for the saturation effect (6th usage episode).

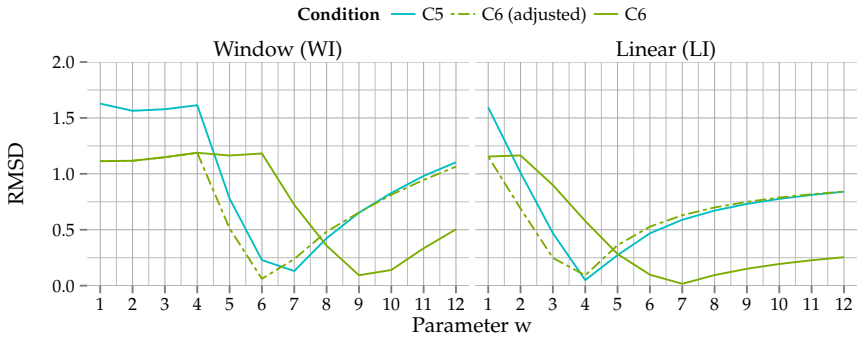

Figure 6.10: Multiple days (E6): multi-episodic prediction accuracy for the saturation effect (6th day, i. e., 12th usage episode).

It can thus be concluded that the saturation effect can be accounted for by using the proposed algorithm. For all three experiments, this algorithm resulted in a reduction of w. Although the RMSD is increased in some cases, w is reduced to the optimal solution considering all conditions without the saturation adjustment (see Section 6.3.1). For all conditions, applying the adjustment leads to a reduction of the RMSD (E1: 0.02, E2a: 0.03, and E6: 0.08). In fact, this overall reduction is rather small with regard to all conditions.

6.4 CONCLUSION

In this chapter, two model types based on the weighted average for predicting the multi-episodic MOS using the episodic MOS were presented. It could be shown that a weighted average using either a window function or a linear function enabled to predict the multi-episodic MOS. Both weight functions perform similarly if only HP usage episodes were presented. In fact, if LP usages episodes were presented, both outperform the unweighted average of all prior episodic judgments, i. e., a smaller w achieves a better prediction accuracy. Comparing WI and LI, the latter shows a better prediction accuracy and higher robustness for choosing the w. This is observed in all three experiments. One interesting observation is made with regard to the one-session experiments. For both weight functions, the highest prediction accuracy is achieved for the same w in case of E1 and E2a, i. e., LI: $w = 2$ and WI: $w = 4$. This suggests that the usage situation must actually not be considered in this case for predicting the multi-episodic MOS.

Although no reason for the observed saturation effect could be derived, accounting for it with the described algorithm improves the overall prediction accuracy for all three experiments.

In fact, it must be noted that the weighted average with the rather simple weight functions (one degree of freedom) enabled a decent prediction accuracy for one session as well as a usage period of 6 days. Nevertheless, the conducted modeling is inherently limited to the data of the conducted experiments, and models could not be verified as no data sets were available that could be used for cross-validation.

7

CONCLUSION

The perceived quality of telecommunication services, especially with regard to repeated-use, is an important aspect for service providers, because it might affect the user's satisfaction and also future-use behavior. The formation processes of retrospective judgments of general experiences as well as perceived quality are known to be affected by several biases such as recency effect, peak effect, and duration neglect (cf. Section 2.2). These effects describe that not all parts of an experience are equally reflected in retrospective judgments of this experience. With regard to perceived quality, such effects have been investigated mainly for macroscopic fluctuations in one stimulus (cf. Section 2.3). For interactive situations, such as telephone calls, methods have been developed to suggest a specific user behavior and thus limit the impact of varying user behavior on the perceived quality. For the investigation of macroscopic fluctuations of telephony, most prominent is here the method of *simulated conversations* (Weiss et al., 2009). For the assessment of multi-episodic perceived quality, Möller et al. (2011) presented the defined-use method, i. e., multiple participants are exposed to the same multi-episodic condition by defining the performance as well as when, how, and for what a service should be used, i. e., defining all usage episodes. This enables to investigate the formation process of multi-episodic perceived quality by deriving a MOS, as the same multi-episodic condition can be presented to multiple participants. This complements prior work by Duncanson (1969), which allowed free-use of a service, focusing on the perceived quality of a telephone call with *average performance*. Here, no knowledge about the prior experiences of a user with the service under investigation was available. Therefore, the reasons for the judgments could not be derived.

In this thesis, I investigated the formation process of multi-episodic perceived quality using the defined-use method. Here, I pursued two goals with regard to multi-episodic perceived quality. First, I wanted to understand the impact of potential factors (i. e., varying performance, usage situation, and usage period) affecting the quality formation process. Second, I wanted to implement a model using episodic judgments for the prediction of multi-episodic judgments; both in terms of MOS.

Two different usage periods have been explored: multiple usage episodes in one session (up to 45 min) and individual usage episodes distributed over several days. These two usage periods were evaluated, as it was so far unknown if the time between episodes affects the formation process of multi-episodic perceived quality. In fact, the investigation for one session allowed to conduct experiments in a controlled laboratory environment. This reduced technical complexity, effort, and limited environmental influence factors.

For the investigation of multi-episodic perceived quality, seven experiments were conducted. For one session, I investigated the number, position, duration, and strength of degraded episodes (E1, E2a, and E3). This was complemented by the investigation of the impact of a second service (E2b). E6 was designed based on the one-session experiments and the practical findings of the 14 days experiments (E4 and E5). In this experiment, a usage period of 6 days was investigated.

Extending initial work of Möller et al. (2011), the defined-use method was applied to investigate multi-episodic perceived quality in terms of MOS. In the conducted experiments, the performance was kept near constant within each usage episode. Macroscopic fluctuations were not investigated, as the impact on episodic judgments and multi-episodic judgments is not yet fully understood. For the experiments, a severe but unrealistic reduction in performance was used, i. e., applying LPC-10. This degradation was selected to achieve measurable effects on multi-episodic judgments due to the presentation of degraded episodes. In extension to Möller et al. (2011), a training was conducted, presenting typical service-related degradations before the multi-episodic part of these experiments. Furthermore, in all conducted experiments, *new* services were used, i. e., services were specifically created for the experiments. Thus, participants could not have prior experiences with these services but only with such service types in general. These two adaptations were expected to reduce the impact in terms of unexplainable variance of the required between-subject design, as all participants had a common basis for the assessment of the multi-episodic conditions.

The results of the conducted experiments showed several effects, indicating characteristics of the formation process of multi-episodic perceived quality. The largest effect on multi-episodic judgments was observed for an increasing number of degraded episodes. Here, the final multi-episodic judgment decreased until two consecutive degraded episodes/days were presented. Then, no further decrease was observed although the multi-judgment remained well above the episodic judgments of degraded episodes, i. e., a saturation effect was observed. This effect was observed for one session and also in a usage period of 6 days. Furthermore, the occurrence of a recency effect, which was observed in prior work, was investigated. Such an effect could be observed in the one-session experiments. A difference could

be observed between E1 and E2a. Here, a recency effect could be observed for one degraded episode only in case of two-party conversation (E1), whereas it was not observed for third-party listening (E2a). In both experiments, an impact of the position of degraded episodes was observed for two degraded episodes. This is most probably due to the usage situation, i. e., *passive* versus *active* use. A recency effect was only indicated (non-significant) for a usage period of 6 days (E6). In addition, the impact of presenting degraded episodes consecutively and non-consecutively was investigated. However, no clear effect could be observed. The small, potential difference can also be explained by a recency effect. For one session, also the occurrence of a duration neglect for episodic judgments and the final multi-episodic judgment was investigated. Even doubling the duration of one degraded episode did not yield differences of these judgments. This indicates that the actual duration of one degraded episode is not considered for the episodic judgment and the following multi-episodic judgment if no macroscopic fluctuations occurred within this episode. For one session, also a peak effect was investigated. However, the results were inconclusive.

The large number of conditions in E1, E2a, and E6 enabled to evaluate the accuracy of potential prediction models (Chapter 6). Here, it was desired to predict the multi-episodic MOS using prior episodic judgments in terms of MOS. Based on Möller et al. (2011), who evaluated the prediction accuracy of the average of all prior episodic judgments, the *weighted average* was applied. As weight functions, I evaluated a window function and a linear function. The evaluation shows that both weight functions achieve a better prediction accuracy than the unweighted average of all prior episodic judgments. With regard to prediction accuracy as well as robustness for parameter selection, the linear function performs better than the window function. For the one-session experiments (E1 and E2a), a $w = 2$ and for multiple days (E6) a $w = 4$ provided the best prediction accuracy. In addition, the observed saturation was accounted for by adjusting the episodic judgments rather than the weight functions (Section 6.3.3). If three consecutive *similar* degraded episodes/days occur (C6), then the episodic judgment of the first degraded episode/day is set to the average episodic judgments of all prior non-degraded episodes. This adjustment improves the overall prediction accuracy, as it shifts the optimum w of this condition to the overall optimal w considering all conditions. It could thus be concluded that the multi-episodic MOS can predicted successfully using the episodic MOS by applying the weighted average using the proposed weight functions. Nevertheless, it is important to note that the modeling was conducted without cross-validation, as no suitable data sets were available. Thus, the derived models are valid in terms of curve fitting for the underlaying data sets and might lack generalizability.

The derived knowledge on the formation process of multi-episodic perceived quality can be used as input for models on service quality, such as Parasuraman et al. (1985). In fact, knowledge about the business impact of performance fluctuations for repeated-use is highly desired by providers of telecommunication services.

7.1 DISCUSSION

The results of the conducted experiments showed that multi-episodic perceived quality can be assessed in one session as well as over multiple days using the defined-use method. This method enables to present the same multi-episodic condition to several participants and thus derive a MOS, as nearly the same condition can be presented. The conducted investigation was limited on purpose to severe degradations, so potential effects were likely being observable. Especially, the required between-subject design and the complexity of the experiments make this method unsuitable for the precise investigation of non-severe degradations. Although the use of LPC-10 is unconventional for speech telephony and thus limits generalizability of the results, it provided a severe degradation while allowing successful task fulfillment. Moreover, the defined-use method has some inherent disadvantages, which might affect the formation process of multi-episodic perceived quality. First, this method forces participants to use a service in a specific manner. Here, it is defined when, how, and for what a service has to be used. This might affect the formation process of multi-episodic perceived quality, because participants are not free to use a service to fulfill their own needs, i. e., tasks are not necessarily meaningful and important to them. Second, the formation process of multi-episodic perceived quality might be affected by the assessment of episodic perceived quality. It is thus possible that taking episodic judgments affects the memorization process of the experiences. Taking episodic judgments might increase the ability to remember specific information about an episode, which might affect following episodic judgments. In fact, even the judgment processes and their results might be remembered and recallable. It is not yet known if this affects the formation process of multi-episodic perceived quality. This was not investigated in the conducted experiments, as the episodic judgments were necessary for the verification of the experiments and also for the prediction of the multi-episodic judgments. Third, the application of the defined-use method if applied for multiple days requires that participants can embed the episodes into their daily life. This might be complicated and frustrating and thus might affect multi-episodic judgments. Finally, the results of the experiments are limited to speech-only telecommunication services, as only these were investigated in detail. It is likely that the observed effects can be generalized to other telecommunication ser-

vices, such as video consumption, Internet-based gaming, and web browsing. However, differences might be observed due to the very different usage situations, expectations, and also types of degradations.

The conducted experiments allowed to successfully investigate the formation process of multi-episodic perceived quality. It could be shown that the formation process of multi-episodic perceived quality is affected by several effects. Here, similar effects could be observed that are known to affect retrospective judgments, i. e., recency effect and duration neglect. In addition, a saturation effect was observed for the two investigated usage periods. This effect has so far not been observed for retrospective judgments of perceived quality. In fact, the underlying reason(s) for the observed effects could not be deduced from the conducted experiments.

7.2 FUTURE WORK

Although the experiments showed consistent results, the findings are necessarily limited to the evaluated settings. The here presented results form a useful basis for further investigation of multi-episodic perceived quality. It seems important to investigate if the observed effects also occur for other types of telecommunication services or are specific to the investigated speech-based service types. In fact, the results show that the usage situation seems to affect the multi-episodic judgments. Further investigations are also necessary to evaluate the impact of applied tasks as well as their importance to participants. The perceived quality of an *important* usage episode might have a higher impact on multi-episodic judgments than less important episodes. Another so far not investigated aspect is the inability to fulfill a task due to reduced performance. In fact, the resulting frustration might result in a higher importance for a multi-episodic judgment than a successful episode. A conceptual approach towards integrating the inability to fulfill a task into QoE is presented by Leon-Garcia and Zucherman (2014). This framework might serve as a starting point for multi-episodic perceived quality. Also, the impact of macroscopic fluctuations on multi-episodic judgments has not been investigated so far. Here, it is of interest if episodic judgments are sufficient for the prediction of multi-episodic judgments or if more information about each usage episode is necessary. Furthermore, it is also not known how multi-episodic judgments of service that can be used on multiple, different devices, e. g., a mobile device and a stationary device, are formed. Here, it is not (yet) known if the multi-episodic perceived quality is integrated, for example, by device or rather by service. In addition to knowledge about the quality formation process of multi-episodic perceived quality, also further research on assessment methods is required. For example, the defined-use method allows to de-

rive a MOS. However, this necessarily ignores individual differences between participants, i. e., assumes a similar formation process with similar characteristics for all participants. In fact, the formation process might also be affected by characteristics of subgroups or even individual participants. Thus, the knowledge about the MOS and its prediction should be complemented by investigating the impact of potential individual differences. In addition, the findings should be verified with *real* users that use a service on their own, i. e., free-use, as this allows to verify the findings under ecological valid settings.

Brain: Are you pondering what I'm pondering?

Pinky: Whoof, oh, I'd have to say the odds of that are terribly slim, Brain.

Brain: True.

— Pinky & Brain

APPENDIX

I

EXPERIMENTAL SETUPS

The systems that were used in the conducted experiments are described in the following, focusing on the media processing and required details to reproduce the technical conditions. It must be noted that the systems for experiments E1, E2a, E2b, and E3 were designed especially for use in a laboratory setting and thus focused on a precise presentation of the performance levels. The systems implemented for experiments E4, E5, and E6 were designed for use in field experiments, and thus ease of setup, reliability, and robustness was required.

I.1 EXPERIMENT E1

For E1, two systems were required. First, a listening-only training was conducted, presenting typical speech telephony degradations. For the multi-episodic assessment, a two-party speech telephony system was required.

I.1.1 *Listening-only Training*

The listening-only training was conducted with a tablet computer (*Fujitsu Stylistic ST6012*). For diotic representation, a pair of *AKG K-271* headphones was connected to the internal sound card of the tablet computer. The system was calibrated using a *HEAD acoustics* head and torso simulator *HSM II.3* (sound pressure level of 75 dB20µPa; babble noise). The stimuli were generated with the ITU-T STL2009 tools (ITU-T Recommendation G.191, 2010) and the audio-processing tool SOX[24]. The STL2009 were used for coding G.711 and G.722 (Mode 1), and inserting packet loss for G.722 (Mode 1, PLC Mode 0). SOX was used for coding GSM-FR and LPC-10 as well as for filtering wideband, narrowband, and white noise.

24 The SOX project is hosted under *http://sox.sourceforge.net*.

I.I.II *Two-party Speech Telephony*

For the multi-episodic assessment of E1, a speech telephony system for two-party conversations was required.

NETWORK-BASED SETUP

In the first part of E1, a VoIP-based system consisting of three computers was used. These three computers were connected via Ethernet (CAT-5). One computer (*Lenovo X61*) acted as a Server running the open-source telephony software *Asterisk 11*[25]. Two *Fujitsu Lifebook S761* were running each a customized client based on *PJSIP 2.1*[26], which connected via *Session Initiation Protocol (SIP)* to the server. The clients presented only a minimal user interface, consisting of one button for call initiation and hangup, one button to set the current presence status, and a presence status indicator for the remote client. Call initiation was only possible if both clients set their presence status to available. The two clients could not communicate directly with each other as the Asterisk server acted as a proxy. Transmission of the speech signal between Asterisk and the clients (both directions) was lossless by using the L16 codec (sampling rate: 16 kHz). Asterisk applied the desired performance level, i. e., the selected codec, by compressing the signal and immediately decompressing it before relaying the signal.[27] This system achieved a one-way end-to-end delay of 120 ms.

On each of the two client computers, one *Beyerdynamic DT 790 Pro* headset connected to one *Edirol UA-25EX* sound card was used for recording and diotic reproduction. Before starting the experiment, the output on both clients was once calibrated to a comfortable listening level by the experimental supervisor.

SOUND CARD SETUP

The network-based setup was replaced later by a more elaborate system, which provided easier setup and reduced the complexity for verification. For this system, only one computer without an actual network was used. Rather than transmitting the signals via Ethernet, an analogue transmission via audio cables was used. Here, also the *Beyerdynamic DT 790 Pro* headsets were used. The two headsets were connected to the processing computer (*Lenovo X61*) with one *Edirol UA-25EX* sound card. The signal of each microphone was am-

25 The Asterisk project is hosted under *http://www.asterisk.org*.

26 PJSIP is an open-source library for SIP-based VoIP (*http://www.pjsip.org*).

27 The performance levels were set by using the transcoding capability of Asterisk. On an incoming call coded with L16, Asterisk initiated a call to himself with the desired codec (i. e., G.722, or LPC-10). This second call triggered then an outgoing call to the callee encoded in L16.

plified with a *RME QuadMic II* microphone preamplifier to counter potential signal loss due to the cable length of 10 m.

On the processing computer *PureData*[28], an open-source audio processing application, was used to modify the speech signals. As no speech codecs were available in PureData, the application was extended with the speech codecs G.711, G.722 (Mode 1), and LPC-10. It was verified that the PureData setup provides similar characteristics as the network-based setup. The overall system achieved a constant, glitch-free one-way end-to-end delay of 70 ms. Before starting the experiment, the output was once calibrated to a comfortable listening level by the experimental supervisor.

ADDITIONAL PROCESSING

Due to the performance of the *Beyerdynamic DT 790 Pro* headsets, which provide a pair of closed headphones as well as a directional microphone, neither echo cancellation nor denoising algorithms were applied. Both systems were configured to provide neither side tone nor comfort noise if not introduced by the codec. No additional audio processing was applied. The participants were located in two sound-insulated test rooms which met the requirements according to ITU-T Recommendation P.800 (1996).

I.II EXPERIMENTS E2A, E2B, AND E3

For the experiments E2a, E2b, and E3, audio signals were presented with a pair of *Sennheiser HD 25-1* headphones. For E2a and E3, these were connected to the internal sound card of a *Microsoft Surface Pro* tablet computer. For E2b, these were connected to the internal sound card of a *Nexus 7 (2013)*. In all three experiments, the sound pressure level was calibrated to 75 dB20µPa using a *HEAD acoustics* head and torso simulator *HSM II.3*. The experiments were conducted in a sound-proof cabin following ITU-T Recommendation P.800 (1996). In E2b, the VoD service was presented on the display of the Nexus 7, presenting the videos horizontally in full screen. The brightness of this 7 inch display was set to maximum without applying any device-specific color adaptation.

For these three experiments, non-degraded recordings of the two-party conversations conducted in E1 were selected. These recordings were processed with the ITU-T STL2009 tools (ITU-T Recommendation G.191, 2010) for coding G.722 and with SOX for coding LPC-10. For E2b, the video material was cut with *Adobe Premiere 6* and exported with H.264 (1280x720 px, 5 Mbit/s, two-pass) and AAC (448 kbit/s, stereo).

28 The PureData project is hosted under *https://puredata.info/*.

Subsequently, the video material was processed with *FFmpeg*[29] to apply the desired QP factor while leaving the audio unchanged.

I.III EXPERIMENT E4

E4 was the first experiment for this thesis conducted as a field experiment with a usage period of multiple days. For E4, a speech telephony service and a VoD service were implemented, which could be used by participants with their own personal computer and Internet access. For recording and audio reproduction, each participant was equipped with a *Logitech P120* headset.

I.III.I *Speech Telephony Service*

For conducting E4, a publicly reachable VoIP telephony service was set up. The service needed to allow multiple two-party speech conversations at the same time while being able to insert packet loss for individual conversations. For this service, one server was installed in the data center of *Technische Universität Berlin*. As operating system *FreeBSD 9.0* was selected, because it provides a built-in firewall with a traffic shaper, i. e., *Dummynet* (Rizzo, 1997). This traffic shaper allows to add network impairments (packet loss, delay, and jitter) per individual connection, i. e., protocol, port, and address. On this server, the VoIP software *Asterisk 10* was installed, which was configured to act as a SIP registrar as well as RTP-relay for the UDP-based media streams. On call initiation of a client, Asterisk informed Dummynet about the desired packet loss rate for the media streams of this call. Packet loss was not inserted to the SIP connections between clients and server, because this affects the reactivity of the telephony system and might lead to call drops. On each participant's computer, the VoIP client *Jitsi 1.0* was installed.

The speech telephony server was successfully tested, but it was not able to fulfill the desired performance levels in the experiment. Two limitations have been observed. First of all, participants connected to the service using their own Internet access (minimal required bandwidth: 6 MBit/s). The performance of this connection could neither be estimated nor enforced. Thus, for some calls severe, often bursty, packet loss was observed. However, for the media streams no counter measures were taken against packet loss, i. e., neither elaborate PLC (only zero insertion) nor FEC. Second, packet loss was added by Dummynet in addition to packet loss occurring due to the network issues. This lead to potentially higher packet loss than desired.

29 The FFmpeg project is available at *https://www.ffmpeg.org/*.

I.III.II *Video-on-Demand*

For the VoD service, an actual service was simulated offline rather than providing it as an online service. Here, a video player was implemented with *Microsoft Silverlight* that replicated a video streaming website including login procedure and initial buffering. This video player could only be executed in a web browser. The video player was bundled together with the preprocessed videos on a USB flash drive. Participants were instructed that for using the VoD service the USB flash drive needs to be inserted into the computer and remain connected. Storing the content locally rather than downloading it when needed, avoided the setup of such a service as well as potential network-related issues.

In difference to the speech telephony service, this system achieved the desired performance levels in a reliable manner.

I.IV EXPERIMENT E5

Based on the reliability of the simulated VoD service used in E4, an AoD service and VoD service were selected for E5, again simulating both services. For this experiment, ten *Samsung Galaxy SII (GT-I9210)* mobile phones running *Android 2.3* were used. These mobile phones have a 4.5 inch display with a resolution of 480x800 px. A native Android application was implemented with the same functionality as the website used in E4, i. e., storing media content locally while providing login functionality and initial buffering. For media playback, the VoD service presented the video horizontally in full screen. For audio reproduction, each participant received a pair of *Shure 240* headphones.

I.V EXPERIMENT E6

For E6, only an AoD service was used. For this experiment, it was chosen to let participants use their own equipment. The AoD service was implemented as a website. This website used HTML5 multimedia features to download and playback the audio content. Participants were only required to use *Google Chrome/Chromium* as a web browser. This limited the testing effort for the website. As the buffering strategy for media content cannot be controlled by a website using standardized HTML5, preloading was initiated when the website was loaded. Here, a loading animation was presented for 5 s. This avoids issues due to initial buffering while providing a constant waiting time.

This system worked as planed and the desired performance levels could be successfully achieved.

II

EPISODIC AND MULTI-EPISODIC QUESTIONS

Table ii.1: Questions for the episodic judgments and multi-episodic judgments for all conducted experiments.

Experiment	Episodic judgment	Multi-episodic judgment
E1	Wie bewerten Sie die Gesamtqualität des gerade beendeten Telefonates?	Wie bewerten Sie die Gesamtqualität aller bisher geführten Telefonate?
E2a and E2b (Speech)	Wie bewerten Sie die Gesamtqualität des gerade gehörten Telefonates?	Wie bewerten Sie die Gesamtqualität aller bisher gehörten Telefonate?
E2b VoD	Wie bewerten Sie die Gesamtqualität des gerade gesehenen Videos?	Wie bewertest Du die Gesamtqualität aller bisher gesehenen Videos?
E2b (Speech and VoD)	-	Wie bewerten Sie die Gesamtqualität des Systems bezüglich aller Interaktionen?
E3	Wie bewertest Du die Gesamtqualität der gerade gehörten Episode?	Wie bewertest Du die Gesamtqualität aller bisher gehörten Episoden?
E4 (VoD)	Wie beurteilen Sie die Gesamtqualität der gesehenen Folge?	Wie beurteilen Sie die Gesamtqualität des Fernsehdienstes bis zum jetzigen Zeitpunkt?
E4 (Speech)	Wie beurteilen Sie die Gesamtqualität der Telefonverbindung?	Wie beurteilen Sie die Gesamtqualität der Telefonieverbindungen bis zum jetzigen Zeitpunkt?
E5 (AoD)	Wie beurteilen Sie die Gesamtqualität der gerade abgeschlossenen Nutzung?	Wie beurteilen Sie die Gesamtqualität des Audio-on-Demand-Dienstes seit Beginn der Studie?
E5 (VoD)	Wie beurteilen Sie die Gesamtqualität der gerade abgeschlossenen Nutzung?	Wie beurteilen Sie die Gesamtqualität des Video-on-Demand-Dienstes seit Beginn der Studie?
E6	Wie bewertest Du die Gesamtqualität der gerade gehörten Episode?	Wie bewertest Du die Gesamtqualität aller bisher gehörten Episoden?

III

RESULTS

In the following, judgments for the here presented experiments are shown. First, episodic judgments are presented as box plots for each usage episode per condition and experiment. Finally, all multi-episodic judgments per experiment and condition are presented.

III.1 EPISODIC JUDGMENTS

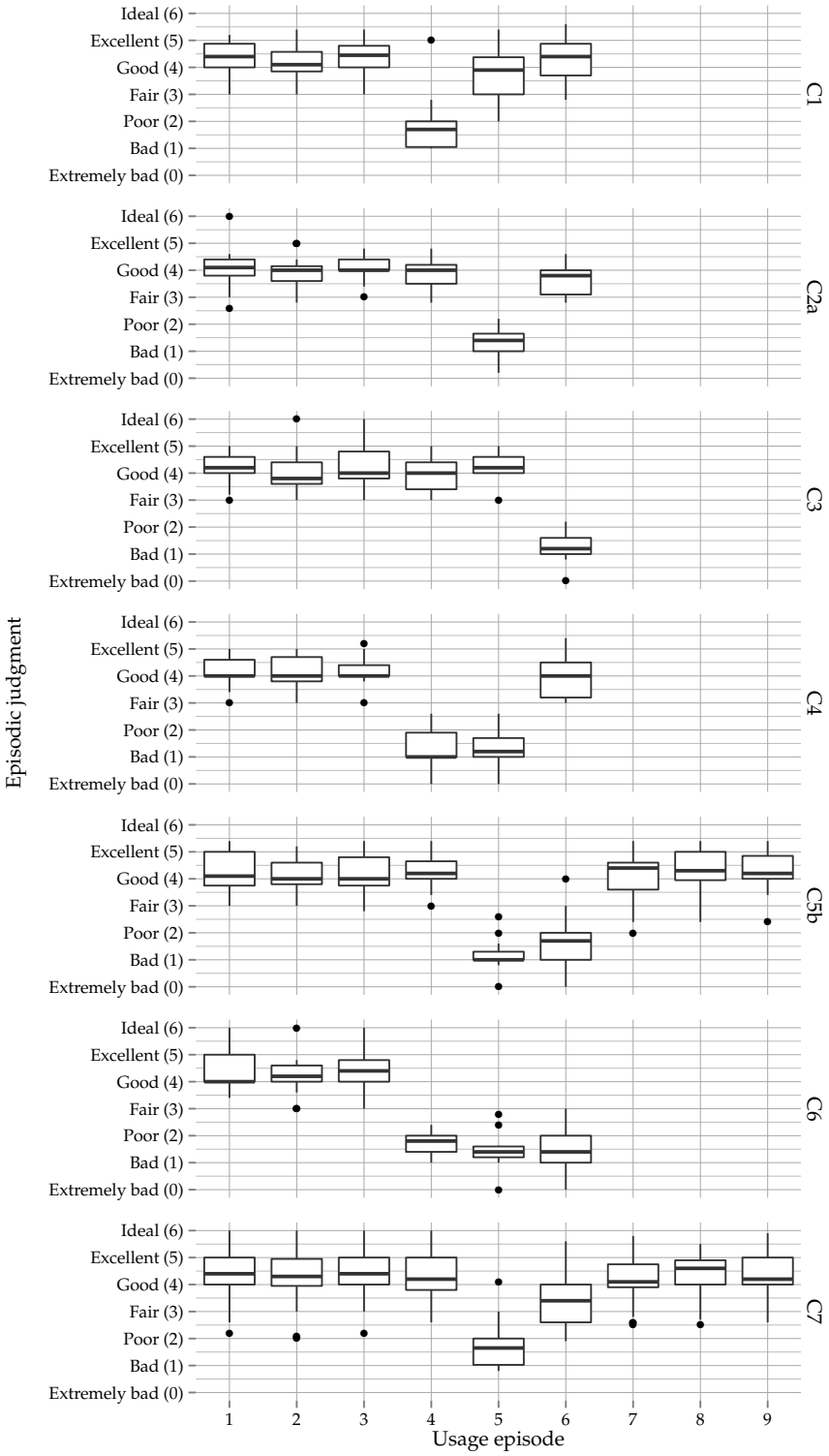

Figure iii.1: One session (E1): box plot of the episodic judgments.

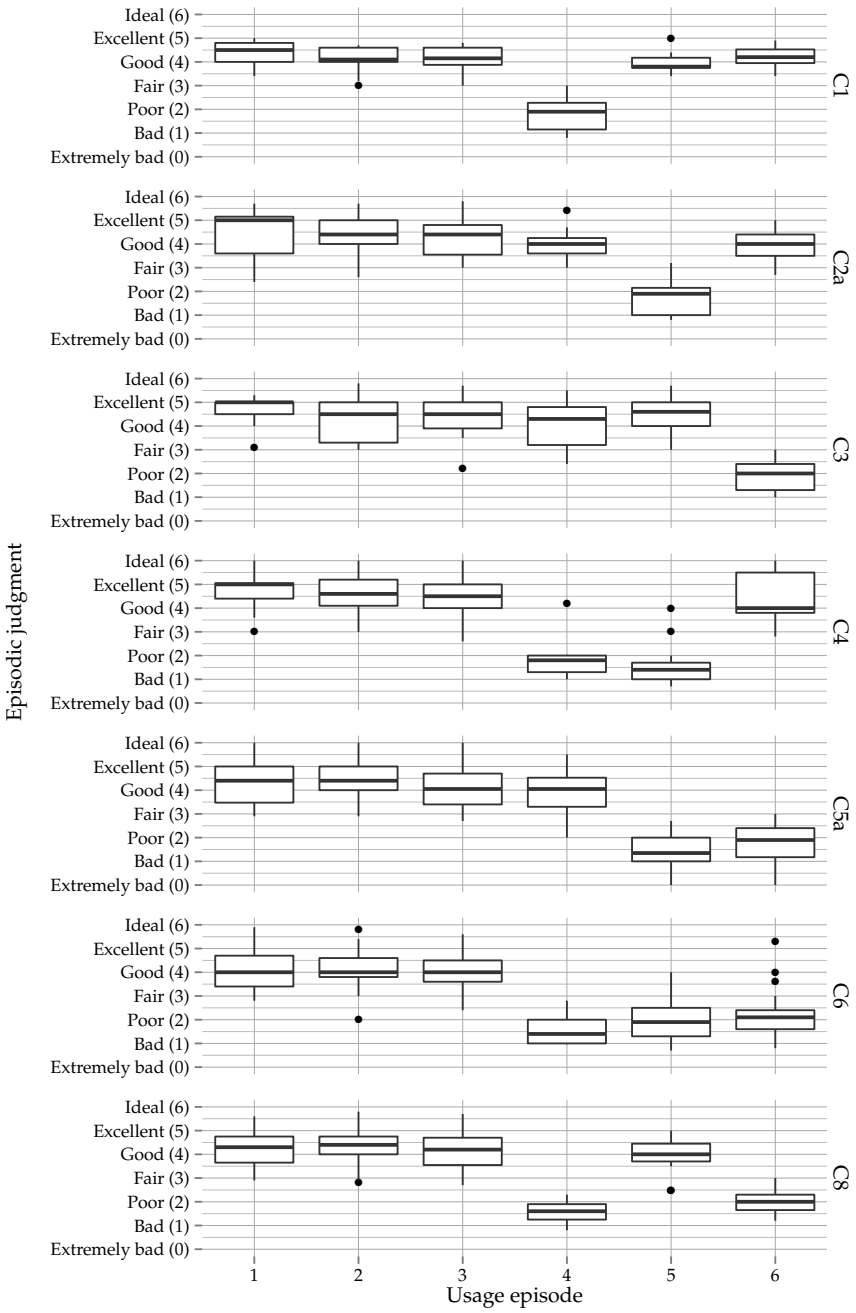

Figure iii.2: One session (E2a): box plot of the episodic judgments.

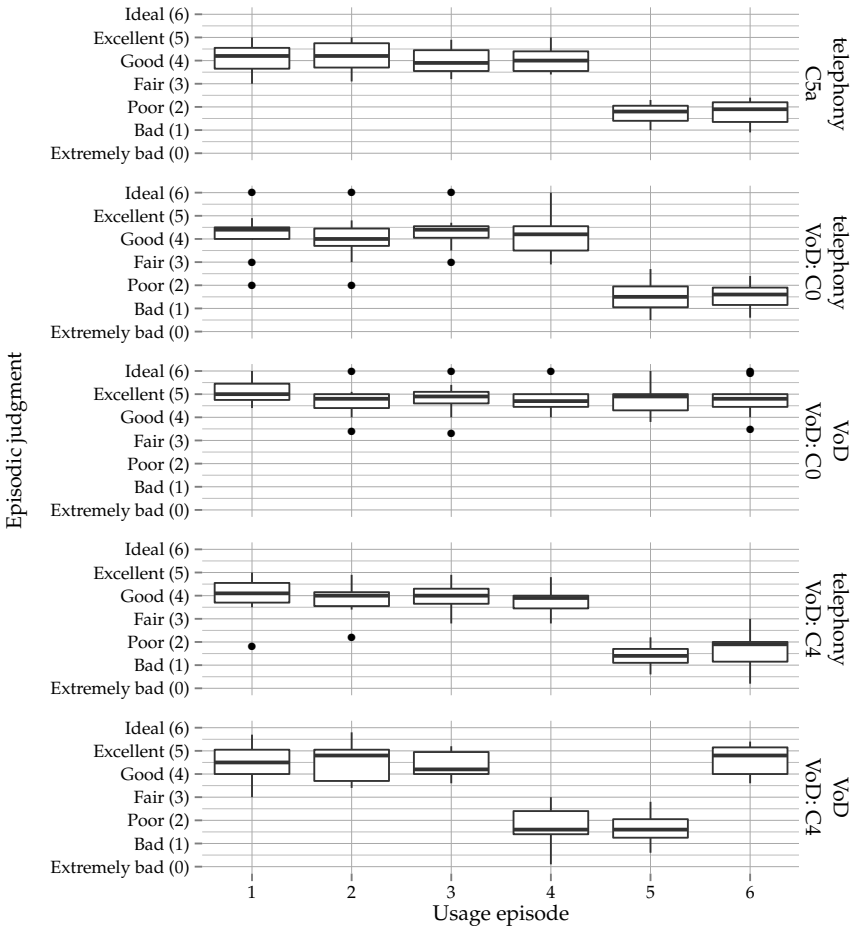

Figure iii.3: One session (E2b): box plot of the episodic judgments.

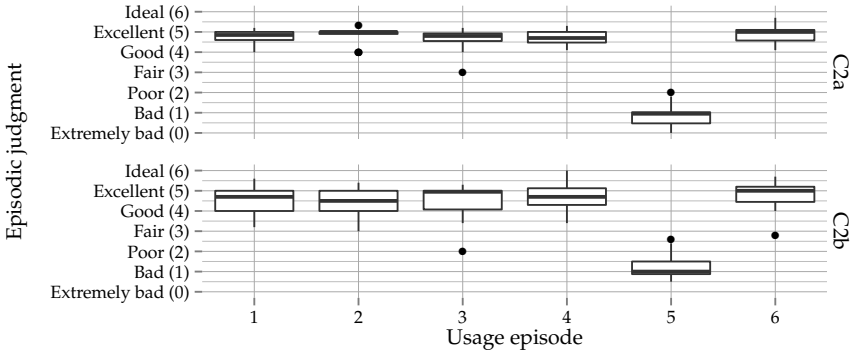

Figure iii.4: One session (E3): box plot of the episodic judgments.

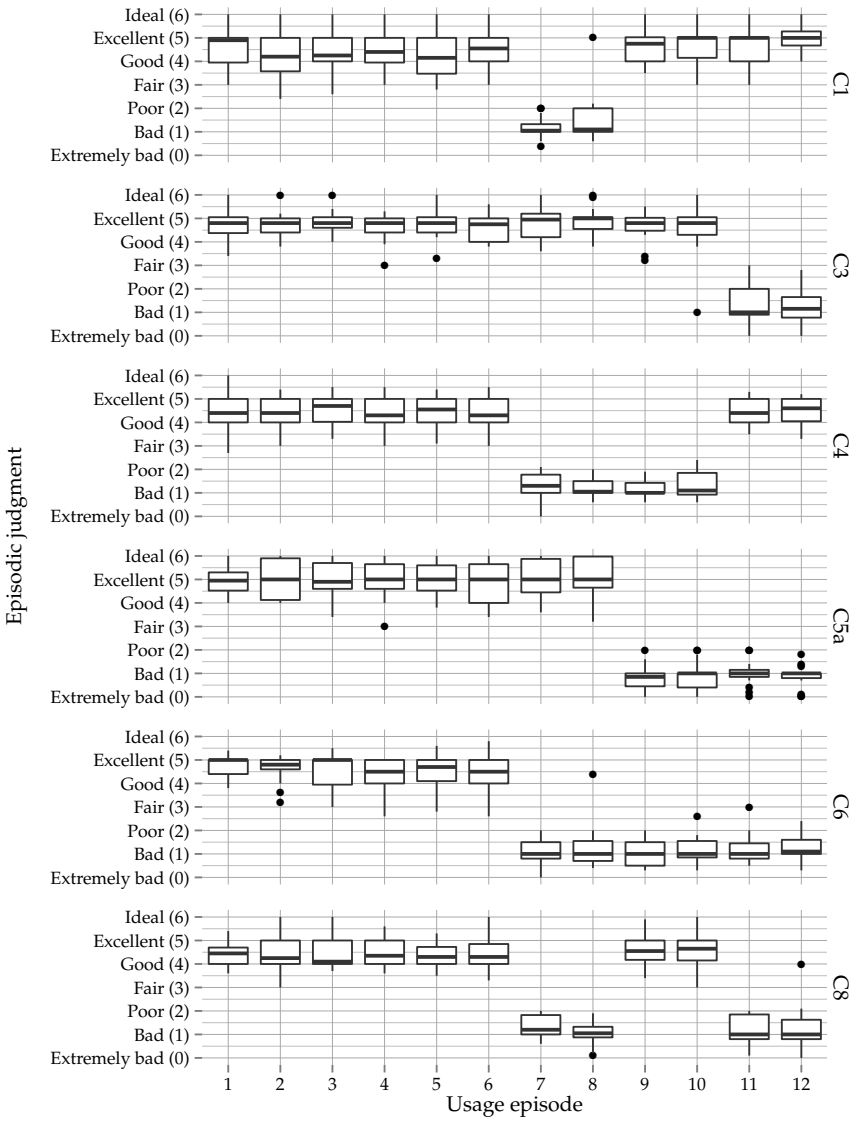

Figure iii.5: Multiple days (E6): box plot of episodic judgments.

III.II MULTI-EPISODIC JUDGMENTS

Table iii.1: Multi-episodic judgments per condition and experiment for E1, E2a, E2b, and E3.

Experiment	Condition	Multi-episodic judgment		
		3rd episode	6th episode	9th episode
E1	C1	4.4 (0.6)	3.7 (0.6)	-
E1	C2a	4.1 (0.4)	3.2 (0.6)	-
E1	C3	4.3 (0.8)	3.1 (0.8)	-
E1	C4	4.2 (0.5)	2.9 (0.5)	-
E1	C5b	4.1 (0.6)	2.3 (0.9)	3.6 (0.6)
E1	C6	4.3 (0.7)	2.1 (0.7)	-
E1	C7	4.4 (0.8)	3.1 (0.7)	4.1 (0.6)
E2a	C1	4.3 (0.6)	3.6 (0.5)	-
E2a	C2a	4.4 (0.7)	3.5 (0.5)	-
E2a	C3	4.3 (0.8)	3.5 (0.5)	-
E2a	C4	4.8 (0.7)	3.0 (0.9)	-
E2a	C5a	4.2 (0.8)	2.4 (0.9)	-
E2a	C6	4.1 (0.8)	2.5 (0.8)	-
E2a	C8	4.3 (0.8)	2.5 (0.6)	-
E2b (telephony)	telephony: C5a; VoD: -	4.2 (0.5)	2.6 (0.8)	-
E2b (telephony)	telephony: C5a; VoD: C0	3.8 (1.3)	2.7 (0.7)	-
E2b (VoD)	telephony: C5a; VoD: C0	4.4 (0.8)	4.7 (0.6)	-
E2b (telephony)	telephony: C5a; VoD: C4	4.0 (0.7)	2.7 (0.8)	-
E2b (VoD)	telephony: C5a; VoD: C0	4.4 (0.9)	3.5 (0.8)	-
E3	C2a	4.8 (0.3)	3.9 (0.6)	-
E3	C2b	4.5 (0.8)	3.7 (0.6)	-

BIBLIOGRAPHY

Ariely, Dan (1998). "Combining experiences over time: The effects of duration, intensity changes and on-line measurements on retrospective pain evaluations." In: *Journal of Behavioral Decision Making* 11.1, pp. 19–45 (cit. on p. 11).

Barsalou, L. W. (1988). "The content and organization of autobiographical memories." In: *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory*. Ed. by U. Neisser and E. Winograd. Vol. 2. Emory symposia in cognition. Cambridge: Cambridge University Press, pp. 193–243. ISBN: 978-0-521-33031-2 (cit. on p. 10).

Belmudez, Benjamin (2015). *Audiovisual Quality Assessment and Prediction for Videotelephony*. Springer International Publishing (cit. on p. 14).

Berger, Jens, Hellenbart, Arpad, Weiss, Benjamin, Sebastian Möller, Gustafsson, Jörgen, and Heikkila, G. (2008). "Estimation of 'quality per call'in modelled telephone conversations." In: *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4809–4812 (cit. on p. 13).

Berry, Leonard L., Zeithaml, Valarie A., and Parasuraman, A. (1985). "Quality counts in services, too." In: *Business Horizons* 28.3, pp. 44–52 (cit. on p. 2).

Black, John B. and Bower, Gordon H. (1979). "Episodes as chunks in narrative memory." In: *Journal of Verbal Learning and Verbal Behavior* 18.3, pp. 309–318 (cit. on p. 10).

Blauert, Jens (1996). *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press. ISBN: 0-262-02413-6 (cit. on pp. 5, 6).

Borowiak, Adam and Reiter, Ulrich (2013). "Long duration audiovisual content: impact of content type and impairment appearance on user quality expectations over time." In: *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. Klagenfurt am Wörtersee: IEEE, pp. 200–205 (cit. on p. 13).

Conway, Martin A. and Pleydell-Pearce, Christopher W. (2000). "The construction of autobiographical memories in the self-memory system." In: *Psychological Review* 107.2, pp. 261–288 (cit. on pp. 9, 10).

Duncanson, James P. (1969). "The average telephone call is better than the average telephone call." In: *The Public Opinion Quarterly* 33.1, pp. 112–116. JSTOR: *10.2307/2747626* (cit. on pp. 17, 18, 20, 26, 85).

Egger, Sebastian, Schatz, Raimund, and Scherer, Stefan (2010). "It Takes Two to Tango - Assessing The Impact of Delay on Conversational Interactivity on Perceived Speech Quality." In: *Proceedings of Interspeech 2010*. ISCA, pp. 1321–1324 (cit. on pp. 19, 30).

ETSI (2011). *Speech and multimedia Transmission Quality (STQ); Estimating Speech Quality per Call (V1.4.1)* (cit. on pp. 13, 14).

Ezzyat, Youssef and Davachi, Lila (2011). "What Constitutes an Episode in Episodic Memory?" In: *Psychological Science* 22.2, pp. 243–252. ISSN: 0956-7976 (cit. on p. 10).

Fredrickson, Barbara L. and Kahneman, Daniel (1993). "Duration neglect in retrospective evaluations of affective episodes." In: *Journal of personality and social psychology* 65.1, pp. 45–55 (cit. on p. 11).

Geerts, David, De Moor, Katrien, Ketyko, Istvan, Jacobs, An, Van den Bergh, Jan, Joseph, Wout, Martens, Luc, and De Marez, Lieven (2010). "Linking an integrated framework with appropriate methods for measuring QoE." In: *Second International Workshop on Quality of Multimedia Experience (QoMEX)*. Trondheim: IEEE, pp. 158–163 (cit. on pp. 2, 16).

Gros, Laetitia and Chateau, Noel (2001). "Instantaneous and Overall Judgements for Time-Varying Speech Quality: Assessments and Relationships." In: *Acta Acustica united with Acustica* 87.3, pp. 367–377 (cit. on pp. 12, 14).

Gros, Laetitia, Chateau, Noel, and Busson, Sylvain (2004). "Effects of context on the subjective assessment of time-varying speech quality: Listening/conversation, laboratory/real environment." In: *Acta Acustica united with Acustica* 90.6, pp. 1037–1051 (cit. on p. 14).

Guéguin, Marie, Le Bouquin-Jeannès, Régine, Gautier-Turbin, Valérie, Faucon, Gérard, and Barriac, Vincent (2008). "On the Evaluation of the Conversational Speech Quality in Telecommunications." In: *EURASIP Journal on Advances in Signal Processing* 2008.1, p. 185248. ISSN: 1687-6180 (cit. on p. 32).

Guse, Dennis and Möller, Sebastian (2013). "Macro-temporal Development of QoE: Impact of Varying Performance on QoE over Multiple Interactions." In: *Proceedings of AIA-DAGA Conference on Acoustics*. Vol. 46. Merano, Italy: Deutsche Gesellschaft für Akustik, pp. 452–455 (cit. on pp. IX, 58).

Guse, Dennis, Weiss, Benjamin, and Möller, Sebastian (2014). "Modelling multi-episodic quality perception for different telecommunication services: first insights." In: *Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*. Singapore. IEEE, pp. 105–110 (cit. on pp. IX, 62, 73).

Hamberg, Roelof and de Ridder, Huib (1999). "Time-varying image quality: Modeling the relation between instantaneous and overall quality." In: *SMPTE journal* 108.11, pp. 802–811 (cit. on pp. 13–15).

Hands, David S. and Avons, Steve. E. (2001). "Recency and duration neglect in subjective assessment of television picture quality." In: *Applied Cognitive Psychology* 15.6, pp. 639–657. ISSN: 1099-0720 (cit. on pp. 12, 14).

Hassenzahl, Marc (2008). "User Experience (UX): Towards an Experiential Perspective on Product Quality." In: *Proceedings of the 20th Conference on L'Interaction Homme-Machine*. IHM '08. New York, NY, USA: ACM, pp. 11–15. ISBN: 978-1-60558-285-6 (cit. on p. 17).

Hogarth, Robin M. and Einhorn, Hillel J. (1992). "Order effects in belief updating: The belief-adjustment model." In: *Cognitive psychology* 24.1, pp. 1–55 (cit. on pp. 3, 14).

Hopper, Robert (1992). *Telephone Conversation*. Indiana University Press. 270 pp. ISBN: 978-0-253-20724-1 (cit. on p. 30).

IEEE Audio and Electroacoustics Group (1969). "IEEE Recommended Practice for Speech Quality Measurements." In: *Audio and Electroacoustics, IEEE Transactions on* 17.3, pp. 225–246. ISSN: 0018-9278 (cit. on p. 2).

ITU (1992). *Handbook of Telephonometry*. International Telecommunication Union. ISBN: 92-61-04911-7 (cit. on p. 30).

ITU-R Recommendation BT.500-13 (2012). *Methodology for the subjective assessment of the quality of television pictures (01/2012)* (cit. on p. 12).

ITU-T Recommendation G.107 (2015). *The E-model: a computational model for use in transmission planning (06/2015)*. International Telecommunication Union (cit. on pp. 8, 42).

ITU-T Recommendation G.191 (2010). *Software tools for speech and audio coding standardization (03/2010)*. International Telecommunication Union (cit. on pp. iii, v).

ITU-T Recommendation G.711 (1988). *Pulse Code Modulation (PCM) of Voice Frequencies (11/1988)*. International Telecommunication Union (cit. on p. 41).

ITU-T Recommendation G.722 (2012). *7 kHz audio-coding within 64 kbit/s (09/2012)*. International Telecommunication Union (cit. on p. 41).

ITU-T Recommendation P.800 (1996). *Methods for subjective determination of transmission quality (08/1996)*. International Telecommunication Union (cit. on p. v).

ITU-T Recommendation P.800.2 (2013). *Mean opinion score and interpretation (05/2013)*. International Telecommunication Union (cit. on p. 8).

ITU-T Recommendation P.805 (2007). *Subjective Evaluation of Conversational Quality (04/2007)*. International Telecommunication Union (cit. on pp. 22, 31, 43).

ITU-T Recommendation P.832 (2000). *Subjective performance evaluation of hands-free terminals (05/2000)*. International Telecommunication Union (cit. on pp. 22, 31).

ITU-T Recommendation P.851 (2003). *Subjective quality evaluation of telephone services based on spoken dialogue systems (11/2003)*. International Telecommunication Union (cit. on p. 22).

ITU-T Recommendation P.863 (2014). *Perceptual objective listening quality assessment (09/2014)*. International Telecommunication Union (cit. on p. 41).

Jekosch, Ute (2005). *Voice and Speech Quality Perception*. Signals and Communication Technology. Berlin: Springer. 207 pp. ISBN: 978-3-540-24095-2 (cit. on pp. 6, 7).

Kahneman, Daniel (2000). "Experienced Utility and Objective Happiness: A Moment-Based Approach." In: *Choices, Value and Frames*. Ed. by Daniel Kahneman and Amos Tversky. Vol. 1. Cambridge University Press, Russel Sage Foundation, pp. 673–708. ISBN: 978-0-521-62749-8 (cit. on p. 16).

Kahneman, Daniel, Fredrickson, Barbara L., Schreiber, Charles A., and Redelmeier, Donald A. (1993). "When more pain is preferred to less: Adding a better end." In: *Psychological Science* 4.6, pp. 401–405 (cit. on pp. 10, 11).

Karapanos, Evangelos, Zimmerman, John, Forlizzi, Jodi, and Martens, Jean-Bernard (2009). "User Experience over Time: An Initial Framework." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. New York, NY, USA: ACM, pp. 729–738. ISBN: 978-1-60558-246-7 (cit. on p. 17).

Köster, Friedemann, Guse, Dennis, Wältermann, Marcel, and Möller, Sebastian (2015). "Comparison between the discrete ACR scale and an extended continuous scale for the quality assessment of transmitted speech." In: *Proceedings of AIA-DAGA conference on Acoustics*. Vol. 48. Nuremburg, Germany: Deutsche Gesellschaft für Akustik, pp. 150–153 (cit. on pp. 8, 22, 41, 42).

Kujala, Sari, Roto, Virpi, Väänänen-Vainio-Mattila, Kaisa, Karapanos, Evangelos, and Sinnelä, Arto (2011). "UX Curve: A method for evaluating long-term user experience." In: *Interacting with Computers* 23.5, pp. 473–483. ISSN: 09535438 (cit. on p. 17).

Kurby, Christopher A. and Zacks, Jeffrey M. (2008). "Segmentation in the perception and memory of events." In: *Trends in Cognitive Sciences* 12.2, pp. 72–79 (cit. on p. 10).

Le Callet, Patrick, Möller, Sebastian, and Perkis, Andrew, eds. (2013). *Qualinet white paper on definitions of quality of experience* (cit. on p. 5).

Leon-Garcia, Alberto and Zucherman, Leon (2014). "Generalizing MOS to assess technical quality for end-to-end Telecom session." In: *2014 IEEE Globecom Workshops (GC Wkshps)*. Institute of Electrical & Electronics Engineers (IEEE), pp. 681–687 (cit. on p. 89).

Lewcio, Blazej (2014). *Management of Speech and Video Telephony Quality in Heterogeneous Wireless Networks*. Springer International Publishing (cit. on p. 14).

Möller, Sebastian (2000). *Assessment and Prediction of Speech Quality in Telecommunications*. Boston, MA: Springer US. ISBN: 978-1-4419-4989-9 (cit. on pp. 1, 31).

Möller, Sebastian (2005). *Quality of Telephone-Based Spoken Dialogue Systems*. Springer Science mathplus Business Media. ISBN: 978-0-387-23190-7 (cit. on p. 1).

Möller, Sebastian, Bang, Chihuy, Tamme, Teele, Vaalgamaa, Markus, and Weiss, Benjamin (2011). "From single-call to multi-call quality: a study on long-term quality integration in audio-visual speech communication." In: *12th Annual Conference of the International Speech Communication Association*. INTERSPEECH. Florence, Italy: ISCA, pp. 1485–1488 (cit. on pp. IX, 18, 22–29, 33, 38, 43, 46, 57–59, 62, 64, 65, 73–76, 85–87).

Murdock Jr., Bennet B. (1962). "The serial position effect of free recall." In: *Journal of Experimental Psychology* 64.5, pp. 482–488 (cit. on p. 10).

Ninassi, Alexandre, Meur, Olivier Le, Callet, Patrick Le, and Barba, Dominique (2009). "Considering Temporal Variations of Spatial Visual Distortions in Video Quality Assessment." In: *Selected Topics in Signal Processing, IEEE Journal of* 3.2, pp. 253–265. ISSN: 1932-4553 (cit. on p. 14).

Parasuraman, Ananthanarayanan, Zeithaml, Valarie A., and Berry, Leonard L. (1985). "A Conceptual Model of Service Quality and Its Implications for Future Research." In: *Journal of Marketing* 49.4, pp. 41–50 (cit. on pp. 2, 22, 88).

Pitrey, Yohann, Engelke, Ulrich, Barkowsky, Marcus, Pépion, Romuald, and Le Callet, Patrick (2011). "Aligning subjective tests using a low cost common set." In: *Euro ITV*. Lisbonne, Portugal, irccyn contribution (cit. on p. 8).

Raake, Alexander (2006a). "Short- and Long-Term Packet Loss Behavior: Towards Speech Quality Prediction for Arbitrary Loss Distributions." In: *IEEE Transactions on Audio, Speech and Language Processing* 14.6, pp. 1957–1968. ISSN: 1558-7916 (cit. on p. 12).

Raake, Alexander (2006b). *Speech quality of VoIP: assessment and prediction*. Chichester, England: Wiley. ISBN: 978-0-470-03060-8 (cit. on p. 9).

Raake, Alexander and Egger, Sebastian (2014). "Quality and Quality of Experience." In: *Quality of Experience*. Ed. by Sebastian Möller and Alexander Raake. Springer International Publishing, pp. 11–33. ISBN: 978-3-319-02681-7 (cit. on pp. 1, 2, 5–7).

Redelmeier, Donald A. and Kahneman, Daniel (1996). "Patients' memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures." In: *Pain* 66.1, pp. 3–8 (cit. on pp. 10, 11).

Reiter, Ulrich, Brunnström, Kjell, De Moor, Katrien, Mohamed-Chaker, Larabi, Pereira, Manuela, Pinheiro, Antonio, You, Jungyong, and

Zgank, Andrej (2014). "Factors Influencing Quality of Experience." In: *Quality of Experience*. Ed. by Sebastian Möller and Alexander Raake. Springer International Publishing, pp. 55–72. ISBN: 978-3-319-02681-7 (cit. on pp. 1, 7).

Rizzo, Luigi (1997). "Dummynet: A Simple Approach to the Evaluation of Network Protocols." In: *SIGCOMM Comput. Commun. Rev.* 27.1, pp. 31–41. ISSN: 0146-4833 (cit. on p. vi).

Robinson, John A. (1992). "First Experience Memories: Contexts and Functions in Personal Histories." In: *Theoretical Perspectives on Autobiographical Memory*. Ed. by Martin A. Conway, David C. Rubin, Hans Spinnler, and Willem A. Wagenaar. Vol. 65. NATO ASI Series. Springer Netherlands, pp. 223–239. ISBN: 978-90-481-4136-4 (cit. on p. 10).

Roeser, Ross (2007). *Audiology*. New York: Thieme. ISBN: 978-1-58890-542-0 (cit. on p. 43).

Rosenbluth, J. H. (1998). "Testing the Quality of Connections Having Time Variant Impairments." In: *ITU-T Delayed Contribution D.064* (cit. on pp. 14, 15).

Roto, Virpi, Law, Effie, Vermeeren, Arnold, and Hoonhout, Jettie, eds. (2011). *User Experience White Paper: Bringing clarity to the concept of user experience* (cit. on p. 17).

Schacter, Daniel L., Chiao, Joan Y., and Mitchell, Jason P. (2003). "The Seven Sins of Memory: Implications for Self." In: *Annals of the New York Academy of Sciences* 1001, pp. 226–239 (cit. on p. 9).

Schatz, Raimund, Egger, Sebastian, and Masuch, Kathrin (2012). "The impact of test duration on user fatigue and reliability of subjective quality ratings." In: *Journal of the Audio Engineering Society* 60.1/2, pp. 63–73 (cit. on p. 30).

Schatz, Raimund, Fiedler, Markus, and Skorin-Kapov, Lea (2014). "QoE-based Management and Application Management." In: *Quality of Experience*. Ed. by Sebastian Möller and Alexander Raake. Springer International Publishing, pp. 411–426. ISBN: 978-3-319-02681-7 (cit. on p. 2).

Schoenenberg, Katrin (2015). "The Quality of Mediated-Conversations under Transmission Delay." Technische Universität Berlin (cit. on p. 19).

Schoenenberg, Katrin, Raake, Alexander, and Koeppe, Judith (2014). "Why are you so slow? – Misattribution of transmission delay to attributes of the conversation partner at the far-end." In: *International Journal of Human-Computer Studies* 72.5, pp. 477–487. ISSN: 10715819 (cit. on p. 30).

Tulving, Endel (1972). "Episodic and Semantic Memory." In: *Organization of Memory*. Ed. by Endel Tulving and Wayne Donaldson. Academic Press, pp. 381–402 (cit. on p. 9).

Wechsung, Ina and De Moor, Katrien (2014). "Quality of Experience versus User Experience." In: *Quality of Experience*. Ed. by Sebas-

tian Möller and Alexander Raake. Springer International Publishing, pp. 35–54. ISBN: 978-3-319-02681-7 (cit. on p. 17).

Weiss, Benjamin, Guse, Dennis, Möller, Sebastian, Raake, Alexander, Borowiak, Adam, and Reiter, Ulrich (2014). "Temporal development of quality of experience." In: *Quality of Experience*. Ed. by Sebastian Möller and Alexander Raake. Springer International Publishing, pp. 133–147. ISBN: 978-3-319-02681-7 (cit. on pp. IX, 8).

Weiss, Benjamin, Möller, Sebastian, Raake, Alexander, Berger, Jens, and Ullmann, Raphael (2009). "Modeling call quality for time-varying transmission characteristics using simulated conversational structures." In: *Acta Acustica united with Acustica* 95.6, pp. 1140–1151 (cit. on pp. 13–15, 85).

Zeithaml, Valarie A., Berry, Leonard L., and Parasuraman, Ananthanarayanan (1996). "The Behavioral Consequences of Service Quality." In: *Journal of Marketing* 60.2, pp. 31–46 (cit. on p. 2).

Zielinski, Slawomir, Rumsey, Francis, and Bech, Søren (2008). "On some biases encountered in modern audio quality listening tests-a review." In: *Journal of the Audio Engineering Society* 56.6, pp. 427–451 (cit. on p. 8).