ILMENAU UNIVERSITY OF TECHNOLOGY

Open Profiling of Quality: A Mixed Methods Research Approach for Audiovisual Quality Evaluations

Dissertation

zur Erlangung des akademischen Grades
Doktoringenieur (Dr.-Ing.)

vorgelegt der Fakultät für Elektrotechnik und Informationstechnik
Technische Universität Ilmenau

von Dipl.-Ing. Dominik Strohmeier
geboren am 27. Oktober 1980 in Memmingen

Gutachter Prof. Dr.-Ing. Karlheinz Brandenburg
Prof. Dr. Søren Bech
Prof. Dr. Patrick Le Callet

Tag der Einreichung: 12.07.2011
Tag der wissenschaftlichen Aussprache: 14.11.2011

urn:nbn:de:gbv:ilm1-2011000368

Abstract

To meet the requirements of consumers and to provide them with a greater quality of experience than existing systems do is a key issue for the success of modern multimedia systems. However, the question about an optimized quality of experience becomes more and more complex as technological systems are evolving and several systems are merged into new ones, e.g. systems for mobile 3D television and video. To be able to optimize critical components of a system under development with as little perceptual errors as possible, user studies are conducted throughout the whole process. A variety of research methods for different purposes have been provided by standardization bodies since the 1970s. These methods allow researchers to evaluate the hedonic excellence of a set of test stimuli. However, a broader view to quality has been taken recently to be able to evaluate quality beyond its hedonic excellence to obtain a greater knowledge about perceived quality and its subjective quality factors that impact on the user.

The goal of this thesis is twofold. The primary goal is the development of a validated mixed-methods research approach for audiovisual quality evaluations. The method shall allow collecting quantitative and descriptive data during the experiment to combine evaluation of hedonic excellence and the elicitation of underlying subjective quality factors. The second goal is the application of the developed method within a series of studies in the domain of mobile 3D video and television to show its applicability.

Open Profiling of Quality (OPQ) is a mixed-methods research approach which combines a quantitative, psychoperceptual evaluation of hedonic excellence and a descriptive sensory analysis of underlying quality factors based on naïve participants' individual vocabulary. This combination allows defining the excellence of overall quality, understanding the characteristics of quality perception, and, eventually, constructing a link between preferences and quality attributes. The method was developed under constructive research with respect to validity and reliability of test results. A series of quality evaluation studies with more than 300 test participants was conducted along different critical components of a system for optimized mobile 3DTV content delivery over DVB-H. The results complemented each other, and, even more importantly, quantitative quality preferences were explained by sensory descriptions in all studies.

Beyond the development of OPQ, the thesis proposes further research approaches, e.g. a conventional profiling in which OPQ's individual vacobulary is substituted by a fixed set of Quality of Experience components or Descriptive Sorted Napping which combines a sorting task and a short post-task interview. All approaches are compared to Open Profiling of Quality at the end of the thesis. To be able to holistically contrast strengths and weaknesses of each method, a comparison model for audiovisual evaluation methods was developed and a first conceptual operationalization of the model was applied in the comparison.

Kurzfassung

Den Anforderungen der Konsumenten gerecht zu werden und ihnen eine immer besser werdende Quality of Experience zu bieten, ist eine der großen Herausforderungen jeder Neuentwicklung im Bereich der Multimediasysteme. Doch proportional zur technischen Komplexität neuer Systeme, in denen Komponenten unterschiedlicher Technologien zu neuen System wie zum Beispiel mobilem 3D-Fernsehen verschmolzen werden, steigt auch die Frage, wie eine optimierte Quality of Experience eigentlich zu erreichen ist. Daher werden seit langer Zeit Nutzertests zur subjektiven Qualitätsbewertung durchgeführt. Deren Ziel über den gesamten Entwcklungsprozesses ist es, die kritischen Komponenten des Systems mit so wenig wie möglich wahrnehmbarem Einfluss auf die wahrgenommene Qualität des Nutzers zu optimieren. Bereits seit den 1970er Jahren werden hierfür Leitfäden verschiedener Standardisierungsgremien zur Verfügung gestellt, in denen unterschiedliche Evaluationsmethoden definiert sind, um die wahrgenommene Gesamtqualität des Systems mit Hilfe von Skalen quantitativ evaluieren zu können. Aktuelle Ansätze erweitern diese klassische Methoden um Sichtweise, die über die klassische Evaluation hedonistischer Gesamtqualität hinausgehen, um das Wissen über individuell zugrundeliegende Qualitätsfaktoren zu erweitern.

Die vorliegende Dissertation verfolgt dabei zwei Ziele. Zum einen soll eine audiovisuelle Evaluationsmethode entwickelt werden, die eine kombinierte Analyse quantitativer und qualitativer Daten ermöglicht, um eine Verknüpfung hedonistischer Qualität und zugrundeliegender Qualitätsfaktoren zu ermöglichen. Weiter soll diese Methode innerhalb des Gebiets der mobiler 3DTV-Systeme erprobt und validiert werden.

Open Profiling of Quality (OPQ) als Evaluationsmethode kombiniert quantitative Evaluation wahrgenommener Gesamtqualität und deskriptive, sensorische Analyse zur Erhebung individueller Qualitätsfaktoren. Die Methode ist für Erhebungen mit naïven Probanden geeignet. OPQ wurde unter besonderer Beachtung von Validität und Reliabilität in einem konstruktivem Ansatz entwickelt und in einer Folge von Studien während der Entwicklung eines mobilem 3DTV-Systems mit über 300 Probanden angewendet. Die Ergebnisse dieser Studien unterstreichen die sich ergänzenden Ergebnisse quantitativer und sensorischer Analysen.

Neben der Entwicklung von OPQ werden in der vorliegenden Arbeit weitere Ansätze sensorischer Analyse präsentiert und miteinander verglichen. Gerade dieser Vergleich ist ein wichtiger Bestandteil der Validierung der OPQ-Methode. Um die Stärken und Schwächen jeder Methode ganzheitlich erfassen und vergleichen zu können, wurde hierfür ein Methodenvergleichsmodell entwickelt und operationalisiert, das den methodischen Beitrag der Arbeit vervollständigt.

Contents

1. Introduction

1.1. Background

To become successful, new multimedia systems and services need to meet the user requirements, offer pleasurable experiences, and provide higher quality than the existing systems. At the same time, audiovisual systems are becoming more and more complex as technological progress provides new possibilities of presenting content. For example, audiovisual 3D on portable devices requires a high level of optimization of technical resources to handle huge amounts of data, with possible limitations due to transmission channel and device constraints like display size or calculation power. This variation can result in perceivable heterogeneous impairments in the production chain from content capturing to display techniques, ultimately impacting the user's perception of quality. To assess the experienced quality of these novel systems and services, subjective audiovisual quality evaluation experiments are conducted. Subjective quality evaluation is based on human judgments of various aspects of experienced material based on perceptual processes. These quality perceptions encompass both low-level sensorial and high-level cognitive processing, including knowledge, emotions, attitudes, and expectations. Since the 1970s, recommendations for video quality evaluations have offered a strong basis for assessing one dimension of quality: its hedonic excellence. Recently, a broader view of quality has been taken by covering other aspects of active perception in the evaluations, including knowledge, different levels of human information processing, and even contextual behavior. Although these evaluations have made a significant contribution for understanding quality, they are still limited to the investigation of quantitative quality preferences. Subjective impressions, interpretations, and experiences as factors to explain and understand the results (constructed in the evaluations of different system factors) beyond the excellence are rarely considered, in part because of a lack of reliable explorative instruments for tackling the descriptive characteristics of quality or, even more ambitiously, relating quality preferences and descriptions. A few previous attempts have been suggested to those concerned with multimedia quality, but they have constraints in terms of accuracy, complexity, required type of assessors, unimodal evaluations, or their emphasis on qualitative methods only.

1.2. Objectives and scope

The main research problem of this thesis concerns methodological nature and is related to the development of a mixed methods research approach for audiovisual quality assessment. The underlying research question was formulated as follows: **"How can quantitative and descriptive data collected in audiovisual quality assessment be combined into a mixed-methods research**

approach applicable for quality evaluations with naïve assessors?" From that main research question, two supplementary research questions arose: 1) *How can individual quality attributes be generalized to general components of Quality of Experience?* 2) *How can audiovisual mixed-methods research approaches be compared systematically to determine the strengths and weaknesses of each method?* A third supplementary research question relates to the exploratory application of the developed research method in a constructive research approach to study the critical components of a mobile 3D television system: *"Which are the critical quality factors in mobile 3D video and television, and how do they impact the overall Quality of Experience of the system?"*

Scope The scope of the thesis is multidisciplinary. The development of Open Profiling of Quality belongs mainly to the wide field of sensory evaluations that originally stemmed from the food sciences. The methods of descriptive quality evaluations in which verbal descriptors are applied to evaluate individual perceived quality are therefore regarded to be among the most sophisticated research approaches. They have been widely used in the domain of food sciences and were later adapted to other methods of research, for example, audio evaluations. Thus, the scope of this thesis lies in the identification of potential research methods, including aspects of data collection methods of analysis and the adaptation of a suitable method to the field of audiovisual quality evaluation. The secondary scope of the thesis applies to the field of multimedia engineering in which subjective quality evaluations play a crucial role for the optimization of systems during the development process. The practical work during the thesis was done for the development of MOBILE3DTV, a system for user-optimized transmission of stereoscopic videos over DVB-H. The studies were planned in a multidisciplinary team who collaborated along the production chain of the system, ranging from capturing, coding, and transmission to the development of a prototype end-to-end system.

Research method Open Profiling of Quality was developed in constructive research. Kasanen et al. [1] describe this constructive approach to a research problem by six subsequent steps of research:

1. "Find a practically relevant problem which also has research potential.
2. Obtain a general and comprehensive understanding of the topic.
3. Innovate, i.e., construct a solution idea.
4. Demonstrate that the solution works.
5. Show the theoretical connections and the research contribution of the solution concept.
6. Examine the scope of applicability of the solution."

The constructive research approach chosen for this thesis follows these steps which are also represented within the structure of the work. First, the motivation to extend existing research methods towards understanding of underlying quality rationales defines the relevant problem for the approach. The literature review identified shortcomings in the currently available methods for audiovisual quality evaluations and defines methods and applications of sensory evaluation in the field of food sciences. The results of the literature review lead to the development of Open Profiling of Quality as the central innovation of the constructive approach. The method was then tested within a series of studies into system optimization tasks during the system development of MOBILE3DTV, a system for optimized stereoscopic content delivery over DVB-H [209, 210]. Each study targeted a methodological validation problem of the OPQ method as well as an application-related research question within the system development process. All studies were conducted along the production chain of MOBILE3DTV which assured variation of quality parameters and related impact by noticable impairments on quality in the level of coding, transmission, and display. During the examination of the scope of applicability of OPQ with respect to aspects of reliability supplementary methodology-related research questions were identified. These supplementary questions led to the development of the Extended-OPQ approach and the comparison model for audiovisual quality evaluation methods.

1.3. Contribution of the author

The author's contribution in relation to the thesis is twofold. First, the author introduced Open Profiling of Quality as a mixed-methods research approach to the field of audiovisual quality evaluation. Second, the application of OPQ in the research area of mobile 3D television and media has deepened the understanding of critical components for a Mobile3DTV system.

Open Profiling of Quality (OPQ) and its extensions and adaptations are the contributions in the methodological part of the thesis. The author proposes a well-validated mixed research method which in constructive research within six studies with more than 300 test participants. Throughout the development process, different approaches from data collection to data analysis were studied and compared to each other for a careful validation of the method and reliability of the test results. Thus, OPQ is presented as a mixed research method that extends the common approaches of subjective quality evaluation with a descriptive analysis approach based on sensory evaluation. The method allows for a quantitative analysis of quality preferences, a descriptive analysis for evaluation of underlying quality factors, and a possibility of linking preferences and quality factors in a combined analysis. The method is designed to be applicable for naïve test participants. Eventually, it completes the current research approach of user-centered quality of experience evaluation in which the development was embedded. Open Profiling of Quality has been compared to related research methods, and the final method was proposed for standardization to ITU-T SG12 as part of the UC-QoE framework. Beyond the development of the research method, crucial shortcomings were identified, leading to additional contributions in terms of research methods. First, the candidate proposes an Extended-OPQ approach that allows deriving general components of quality of experience from the individuals' sensory data collected in a series of studies. As a first example of operationalization of the developed terminology, another adaptation of OPQ was developed and evaluated in which the developed components are used as fixed vocabulary for sensory evaluation. Second, the need for a holistic comparison of related research methods led to the development of a holistic comparison model for subjective quality evaluation methods. Within this thesis, this model is presented, and a first operationalization towards a tool for method comparison is introduced.

The application of Open Profiling of Quality in the development of a mobile 3D video broadcasting system represents a systematic analysis of the system's critical components along the production chain of mobile 3D television. This application fulfills the demand for sensory evaluations to be applied on a holistic evaluation of a research problem studying the problem from several perspectives. The results of the OPQ studies have deepened the understanding about subjective quality of mobile 3D video and television. Specifically, the author's contribution has guided the development towards a quality-optimized Mobile3DTV system development. In general, the results of the studies identified artifact-free video perception as the key requirement for good mobile 3D video quality. The author showed that depth is only perceived as a positive quality factor if the video is artifact-free. Beyond that, the author identified different processing patterns of audiovisual perception and underscores the importance for better characterization of test samples beyond measurements of visual abilities and hearing levels.

1.4. Related publications by the author

The following original publications of the author are the core publications on which this thesis is based. For each publication, a short overview of the contribution of the author to the manuscript is given in appendix A. A complete list of publications including all supplementary publications can be found in the Own References on page 156.

Peer-reviewed journal publications

P1 *D. Strohmeier*, S. Jumisko-Pyykkö, and K. Kunze, "Open Profiling of Quality: A Mixed Method Approach to Understanding Multimodal Quality Perception," Advances in Multimedia, vol. 2010, Article ID 658980, 28 pages, 2010.

P2 *D. Strohmeier*, S. Jumisko-Pyykkö, and K. Kunze, M. O. Bici, "The Extended-OPQ method for User-centered Quality of Experience evaluation: A study for mobile 3D video broadcasting over DVB-H," special issue "Quality of Multimedia Experience", EURASIP Journal on Image and Video Processing, vol. 2011, Article ID 538294, 24 pages, 2011.

P3 A. Gotchev, G. B. Akar, T. Capin, *D. Strohmeier*, A. Boev, "Three-Dimensional Media for Mobile Devices", Proceedings of the IEEE, Vol. 99, No. 4, pp. 708-741, April 2011.

Conference publications

P4 *D. Strohmeier*, S. Jumisko-Pyykkö, K. Eulenberg, "Open Profiling of Quality: Probing the Method in the Context of Use," Proc. of the International Workshop on Quality of Multimedia Experience (QoMEX 2011), Mechelen, Belgium, Sept. 2011

P5 K. Kunze, *D. Strohmeier*, S. Jumisko-Pyykkö "Comparison of two Mixed Methods Approaches for Multimodal Quality Evaluations: Open Profiling of Quality and Conventional Profiling," Proc. of the International Workshop on Quality of Multimedia Experience (QoMEX 2011), Mechelen, Belgium, Sept. 2011

P6 *D. Strohmeier*, S. Jumisko-Pyykkö, U. Reiter, "Profiling experienced quality factors of audiovisual 3D perception," Proc. of the International Workshop on Quality of Multimedia Experience (QoMEX 2010), Trondheim, Norway, June 2010

P7 *D. Strohmeier*, G. Tech "Sharp, bright, three-dimensional: open profiling of quality for mobile 3DTV coding methods," in Proc. "Multimedia on Mobile Devices" as part of the SPIE Electronic Imaging Conf. 2010, Multimedia on Mobile Devices at Electronic Imaging 2010, San Jose, California, USA, Jan. 2010

P8 S. Jumisko-Pyykkö, *D. Strohmeier*, T. Utriainen, K. Kunze, "Descriptive Quality of Experience for Mobile 3D Video", in Proc. of the 6th Nordic Conference on Human-Computer Interaction (nordiCHI), Reykjavik, Iceland, 2010

1.5. Structure of the thesis

The thesis consists of seven chapters structured in three main parts. The structure is as follows. The first part (chapter 2, chapter 3, chapter 4) introduces Open Profiling of Quality as a well-validated mixed-methods research approach for audiovisual quality evaluations. Chapter 2 reviews the existing research methods for subjective quality evaluations. The author introduces the User-Centered Quality of Experience evaluation framework in which the development of the Open Profiling of Quality has been embedded. Further, this section presents subjective quality assessment methods ranging from standardized, quantitative methods to modern approaches of descriptive and mixed-methods research. Chapter 3 includes the methodological development and description of Open Profiling of Quality as a new tool for audiovisual mixed-methods research. The detailed methodological descriptions consider all steps of an OPQ study and introduce different approaches to conducting the study and analyzing the results. This section is important in establishing the validity of the OPQ. Chapter 4 finalizes the first part of the thesis. It includes four fully reported OPQ studies conducted on subjective quality evaluations of mobile 3D television and video. The application of OPQ to different research questions affirms the reliability and validity of the tool and finalizes the presentation of OPQ as a research tool for mixed-methods quality evaluations.

The second part in chapter 5 of the thesis presents the Extended-OPQ method. This extension of the OPQ method allows for developing general components of Quality of Experience from the individual vocabulary collected in a series of OPQ studies. The methodological presentation of the component model is followed by the presentation of a study in which OPQ was compared to the results of conventional profiling, in which these general components were used instead of individual vocabulary.

As a last part of the thesis, chapter 6 compares Open Profiling of Quality to related methods of descriptive, mixed-methods quality evaluations. The methods under comparison have been chosen and adapted from methods identified in the state-of-the-art review in part 1 of this work. The comparison of methods is based on a comparison model that allows for a holistic comparison of related research methods going beyond juxtaposition of results. From the study presented in part, recommendations for the application of the different related methods are drawn. The whole thesis is then summarized and discussed in chapter 7.

2. Related Research

This section reviews the state-of-the-art and related research approaches. Starting from general considerations about the principles of good experimental research, existing quality evaluation approaches are reviewed. I introduce the User-Centered Quality of Experience evaluation framework as the framework in which the development of Open Profiling of Quality has been embedded. Parts of this sections have been published in Jumisko-Pyykkö and Strohmeier, "Report on research methodologies for the experiments," Tech. Rep. Project MOBILE3DTV, 2008 [211] and Strohmeier, "Wahrnehmungsuntersuchung von 2D vs. 3D Displays in A/V-Applikationen mittels einer kombinierten Analysemethodik," Diploma thesis, Ilmenau University of Technology, Ilmenau, Germany, 2007 [212].

2.1. Principles of good experimental research

Studies in audiovisual quality evaluations, as a part of experimental research, are generally characterized by being both a process and a product [2]. Both aspects of a study must be credible to other researchers and comparable among studies to make comparisons and draw joint conclusions. This requirement leads to strict demands on research methods, described in the principles of good research. Haslam and McGarty [2] define five criteria that every research method has to fulfill: reliability, validity, cumulativity, parsimony, and public replication. Among these five criteria, reliability and validity are the most crucial and need to be met for every existing research method and in every method development process.

Reliability is defined as the "confidence that a given empirical finding can be reproduced" [2]. More in detail, we can differentiate reliability by internal, interrater, and external reliability [3, 4]. Internal reliability refers to the consistency of a test within itself. The ability of test participants to use a test method consistently over the time of their test session is a key aspect of internal reliability. Internal reliability can be assessed by consistent scores of a test participant across time and consistent scores between the original test and retests. In addition, the use of hidden anchors, test items that are obviously of very good (or unimpaired) or very bad quality, is possible. Interrater reliability refers to a group of test participants being able to use the test method in a similar way. Within a research method, a standardized test description and common introduction for all participants assures that the evaluation task and the use of scales are conducted similarly by participants. In addition, training and anchoring tests should be conducted before the evaluation to provide for practicing the evaluation task [5]. Mathematical approaches for measuring internal reliability are the split-half method, which calculates the correlation between scores on two equal parts of the test participants, or Cronbach's Alpha, which measures the scale reliability by using the variance of scores per item in relation to the overall variance of the scale [4]. However, not only the reliability within a study but also aspects between studies need to be assured in quality research. External reliability refers to the stability of test results over time. It describes the ability of researchers to replicate a study and to obtain similar results and conclusions. An interesting aspect related to development of new test methods is the question of whether a researcher other than the developer is able to apply the test method in similar way to the developer. [4]

Validity is related "to our confidence that a given finding shows what it purports to show" [2]. A valid finding has been logically and correctly interpreted. As with reliability, validity also has different aspects. A central aspect to quality evaluation methods is internal validity, which describes the cause-effect relationship and checks whether an effect found in the analysis of a study can be related to the independent variables, thus enabling a conclusion concerning causal impact [4, 6]. External validity or generalizability refers to the extent to which results of research can be generalized beyond the experimental context across samples, settings, and time [4, 6]. Each experimental evaluation represents only a snapshot of a complex system being tested although test parameters and samples are carefully chosen. However, its goals are that the findings are generalizable from the chosen sample to other people, from the test device to other systems, and from laboratory settings to the field. Mediating between internal and external validity, construct validity describes the theoretical accuracy of the research. It is the extent to which the results encompass the intended theoretical construct and asks whether research has arrived at the correct explanation for any cause-and-effect relationship that was found in a study. Eventually, it assures that test methods really measure what they are supposed to measure. In contrast to reliability, validity is hard to measure, but many aspects are solvable within the limits of the logic of statistics. Nevertheless, several threats to validity need to be avoided in valid research methods and designs [4, 6–8] (Table 2.1).

One more principle, parsimony, becomes important in method development processes. Parsimony means that research intends to "explain the largest number of facts in terms of the smallest amount of [theoretical] principles" [2]. Originally, it asserted that the best theory within a domain of research is the one that can provide the most economical or simplest explanation of evidence. Parsimony has also been discussed in relation to the growing amount of data in experimental research caused by a continuous increase in available research methods. The best research method is the one that can build a valid and reliable result with as few data as possible. Conversely, new or extended research methods are needed as soon as the common ones fail to provide a full explanation in accordance to the research question.

The following section will present the related work and the state-of-the-art in audiovisual quality evaluations from a methodological point of view. Starting with the User-Centered Quality of Experience evaluation framework, it will introduce the basic concepts of quality, quality perception, and mixed-methods research before it reviews different research methods ranging from traditional psychoperceptual evaluations to current developments in user-centered quality assessment methods.

Threat to validity	Explanation	Example	Precautions
Internal validity			
Sampling bias	Wrong conclusions about the cause-effect relationship are drawn because of an unbalanced distribution of test participants	non-randomized allocation of test participants to different conditions in the test or an insufficient size of the test sample	Thorough randomized allocation of test participants to conditions in the test design; identification of necessary participants per condition for fulfillment of requirements
Instrumentation effect	Changes in the calibration of the test device over time or changes in research personnel may result in differences that masquerade as treatment effects	In quality evaluations, this problem may occur in multi-session design or in test-retest conditions	Definition of exact calibration settings for each study for devices and circumstances; detailed reporting of preparation of test items; detailed preparation of research plan
Incorrect applied statistics	The wrong method can lead to a misinterpreted cause-effect relationship at the end of a test although the test was carefully planned and conducted before	Research method and the characteristics of collected data determine the choice of applicable statistics. Requirements for applicable methods of analysis need to be checked (e.g., type of data (categorical, numerical) or normal distribution).	Analysis of test data with related methods of analysis; thorough check of requirements for specific methods
External validity			
Interaction effect Sample – Test design	The selected samples do not allow generalizing the results to other groups of samples	Experts may emphasize the perception of results which does not allow for generalization towards naïve users	Identification of potential user groups and selection of test samples in accordance; screening of visual and auditive abilities before a test
Interaction effect Settings – Test design	The selected test settings in terms of physical and temporal circumstances do not allow generalizing the results	Evaluation of quality assessments only in the laboratory may emphasize results that will not be critical in other contexts	Identification of user requirements and contexts of use and inclusion of both in the test design
Construct validity			
Researcher Expectancies	The behavior of the person conducting the test has an impact on the data	Researchers indicate during introduction to the test which items are expected to have which quality level	Dividing personnel between planning and conducting a study; training person conducting the test so that his behavior is the same in each evaluation
Mono-Method Bias	A single method does not provide evidence about what is really measured and can lead to wrong conclusions	Quantitative test results do not indicate whether the test participants really perceived and rated impairments by the varied technical parameters	Implementation of multiple measures; multimethodological evaluations within research frameworks; conduction of pilot studies to demonstrate valid measures.

Table 2.1. – Examples of threats to internal, external, and construct validity. A detailed list is provided by Cook and Campbell [7]

2.2. User-Centered Quality of Experience evaluation framework

2.2.1. General considerations of the framework

The User-Centered Quality of Experience (UC-QoE) evaluation framework is a collection of independent methods and factors that relate quality evaluation to the potential use of a system or service. The framework takes into account "1) potential users as quality evaluators, 2) necessary system or service characteristics included in its potential content and critical system components, 3) potential context of use resulting in evaluation quasi-experimental settings and the controlled surroundings, 4) that evaluation tasks are connected to expected usage, and/or they aim also to understand the interpretation of quality parallel to excellence evaluation and can include supplementary ergonomic measures." [9] It represents the methodological part of the UC-QoE model which has been introduced by Jumisko-Pyykkö [9]. Jumisko-Pyykkö's model builds upon five principles [9]:

1. Multimodal quality perception is an active process which encompasses different levels of human information processing and combines information from various modalities.
2. Critical system components need to be holistically optimized by reflecting the factors of external validity in terms of users, systems and services, and contexts of use.
3. Optimization of novel multimedia systems which combine several modalities and multiple parameters requires an overall quality assessment approach and a connection to user requirements.
4. Quality evaluations need to go beyond measures of detectable artifacts and their impact on the user.
5. Quality evaluation experiments can be understood as a part of the user-centered design process. Early-phase prototypes can offer a possibility for quality evaluations to verify user requirements before the high-fidelity prototype is finished.

Jumisko-Pyykkö's approach tackles the existing system-centric paradigms of subjective quality evaluations. It stresses the importance of an increased level of realism by improving the external validity of multimedia quality evaluations in terms of potential users, inclusion of user requirements, and the contexts of use [9]. This approach demands new research methods that extend the ability of existing quality evaluation approaches beyond quantitative ratings of hedonic preferences. The identification of the shortcomings of existing evaluation approaches allows the introduction of new methods with respect to the principle of parsimony. The methodological UC-QoE framework combines a multimethodological approach and extends standardized quality evaluation methods with methods for conducting evaluations in the context of use and a goal to understand and interpret overall quality beyond the measures of excellence in accordance with the dualistic nature of quality (see section 2.2.2).

2.2.2. Understanding multimedia quality perception

2.2.2.1. Multimedia quality as a comparison of produced and perceived quality

Quality is the basic concept of the present thesis. Because my work on Open Profiling of Quality has been integrated into the development of the UC-QoE evaluation framework, my definition of quality is closely related to the definition given in the UC-QoE approach [9][209, 210, 213, 214].

In general, quality relates to the degree of excellence of a product or service [10]. By definition, it can be regarded as the "degree to which a set of inherent characteristics fulfills requirements" [11] or, from a consumers' point of view, as users' "perception of the degree to which [their] requirements have been fulfilled" [12]. However, quality can also be used more specifically to describe only "a distinctive attribute or characteristic" [10] possessed by a product. In contrast to overall excellence, this descriptive understanding relates quality to specific factors, for example, the compression quality of a video codec making it subjectively good or bad. I refer to this juxtaposition of excellence and relationship to attributes as the dualistic nature of quality.

This general definition of quality needs to be specified for the field of multimedia. In the domain of multimedia as "the seamless integration of two or more media" [13], quality is characterized by the relationship between produced and perceived quality [9, 14]. The model of produced and perceived quality is extensively described by Jumisko-Pyykkö [9] and forms the basis upon which the motivation for methodological work in my thesis builds. Produced quality refers to the quality that a technical system is able to provide to its users. Technical constraints for produced quality are given in all abstraction levels of multimedia systems: content, media, and network [15, 16]. In case of mobile 3D television and video, these constraints can result in the juxtaposition of a huge amount of multimedia data to be transmitted over limited bandwidth, a vulnerable transmission channel, and limitations of the receiving devices and the stereoscopic display [209, 210].

While produced quality describes multimedia quality from the viewpoint of the system, perceived (also called subjective or experienced) quality describes the users' or consumers' views on multimedia quality. It refers to the quality perceived and interpreted by the individual user in his active perceptual processes [9, 17]. These perceptual processes characterize perceived quality. Low-level sensory processes are data-driven bottom-up processes that extract relevant information from all incoming sensory sensations [9, 17]. Exemplary relevant features of audiovisual 3D multimedia quality perception are brightness, color, and stereoscopic cues for visual sensation or loudness, pitch, and timbre for auditory sensation [17, 18][212]. After the low-level processing, the processed information is interpreted in high-level cognitive perception. In this stage, stimuli are interpreted according to individual meanings and their relevance to human goal-oriented actions. These top-down processes involve individual emotions, knowledge, expectations, and schemas representing reality that can weight or modify the importance of each sensory attribute, enabling contextual behavior and active quality interpretation [9, 17, 19–21][212]. Neisser's perceptual cycle (see Figure 2.1) is a simplified model of human perception that can be used to explain the mechanisms of perceived quality.

Neisser's perceptual cycle involves schemata, perceptual exploration, and the stimulus environment (Figure 2.1). While the latter is the sum of all sensory stimuli available to human senses, schemata are high-level patterns that represent one's knowledge about the stimulus environment. Schemata have been derived and refined from former similar experiences. They are stored in the human long-term memory and can be understood as individual expectations about the stimuli available in the environment. These structures are connected to emotions and feelings. In perceptual processes, schemata direct one's attention towards a selection of available information from the stimulus environment. Neisser calls this occurrence the exploration. During exploration, humans collect and identify relevant features, so called environmental stimuli, from the whole sensory information available. According to Neisser's model, perceptual exploration leads to a sampling of the environment, and sampled stimuli are merged into objects. While being transferred into our short-term memory, the perceived environmental samples or objects are recognized by the brain and assigned a meaning. The last step is the interpretation of the perceived objects. One tries to match them to the available schemata of the long-term memory. If matching fails for some stimuli, the schemata are modified according to the new perceived object. The modified schemata then drive the environmental exploration anew. Although the model was criticized by Neisser himself [22] as being too simple and too general, it is still useful to sketch the interactions between low-level and high-level cognitive processes, between quality and knowledge and experience, and between quality and contexts and has been used in related publications of audiovisual quality perception [9, 18][212].

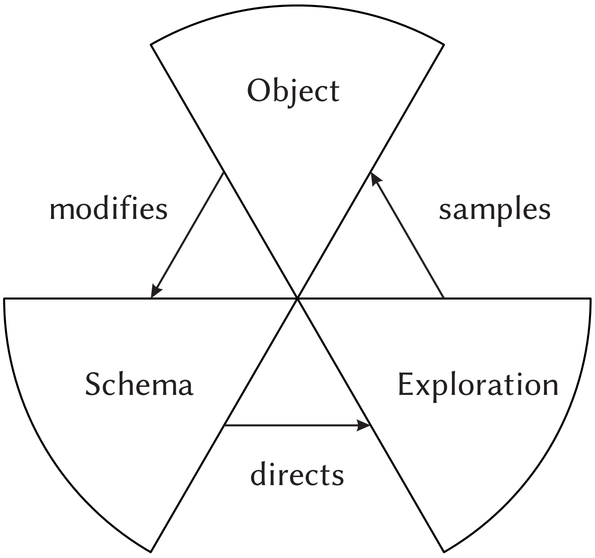


Figure 2.1. – The perceptual cycle by Neisser as a simplified model to understand multimedia quality perception [19]

The model of Neisser shows that human perception is an active exploration towards determining factors that drive perception. This is in accordance with the understanding of the dualistic nature of quality and connects an overall perception of quality to specific attributes. Interestingly, Neisser includes all available sensory information in his exploration and the derivation of schemata. Related to multimodal perception, multimodality cannot be regarded as a separate processing of sensory information of different channels. One sensory channel can complement and modify the perception derived from another channel [23]. The work of Jumisko-Pyykkö [9] shows that these concepts of perception eventually hold true for multimedia quality perception so that these conceptual psychological constructs can be applied in studying and understanding users' perception of multimodal quality.

2.2.2.2. Multimodal quality perception

Multimodal perception is more than just the sum of quality sensation in two or more independent streams [24]. A classical example of audiovisual interactions in perception is the McGurk effect in which auditory and visual information are integrated into a new, different audiovisual perception [25]. The McGurk effect has shown that parallel processing of auditory and visual information does not occur independently. Interaction between the modalities is not just a combinatory mechanism on top of cognitive processing. Coen [26] stresses that the sensory input is shared across all levels of perceptual processing. The dominance of one modality over another strongly depends on multiple factors like intensity, time, and duration of the stimuli [27, 28]. Guski [27] defines three theoretical concepts for dominance of one modality:

> **Hypothesis of the accuracy of modality**: If there is a conflict of information, then the most precise modality dominates all others.
> **Hypothesis of attentional direction**: If there is conflict of information, then the domination of one modality depends on the information the observer pays more attention to.
> **Hypothesis of modality function**: Information conflicts are solved using the modality that offers the best developed function. The visual system seems to be useful for spatial tasks while the auditory system handles temporal problems.

In related audiovisual quality research, cross-modal dominance was found for video-dependent as well as audio-dependent multimodal quality. For good reviews of related studies, the author refers to Soto-Faraco and Kingstone [24] and Jumisko-Pyykkö [9]. Recently, Peregudov et al. [29] presented an audiovisual quality model for mobile multimedia applications, underscoring the importance of interactions between auditory and visual channels for the perception of audiovisual quality.

In summary, the concept of multimedia quality perception describes a complex dependency on technical characteristics and constraints of the system according to individual differences of its users. The relationship between perceived and produced quality in end-to-end systems is described in terms of Quality of Experience (QoE) . QoE is defined as "The overall acceptability of an application or service, as perceived subjectively by the end-user" [30]. More broadly, Wu et al. [31] have summarized it "as a multidimensional construct of user perceptions and behaviors." The goal of modern Quality of Experience evaluation is the optimization of quality factors produced under strict technical constraints or resources with as little negative perceptual effects as possible. Recent multimodal quality evaluation studies have started to stress the importance of high-level cognitive processes for quality perception. With this thesis, the author continues the work along these lines and addresses the challenge of developing an explorative tool to understand the underlying attributes and common rationales of perceived quality.

2.2.3. Theory of mixed-methods research

The UC-QoE evaluation framework combines different research approaches from quantitative and qualitative experimental research into new methods. Its background lies in the theory of mixed-methods research. Becoming more and more popular since the end of the 20^{th} century, it originates from pragmatic philosophy and represents the third wave of research methods [32]. In general, mixed-methods research is defined as a "type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration." [33] This definition provided by Johnson et al. [33] is summarized in various existing definitions of 'mixed-methods research' to comprise a holistic understanding. Its core is the combination of quantitative and qualitative data sets and related methods of analysis to attain broader understanding of research problems than using only one approach. [32, 34]

Quantitative (QUAN) research has traditionally put a focus on deduction, confirmation, theory/hypothesis testing, explanation, prediction, standardized data collection, and statistical analysis. In contrast, traditional qualitative (QUAL) research techniques are induction, discovery, exploration, theory/hypothesis generation, the researcher as the primary 'instrument' of data collection, and qualitative analysis. Eventually merging two data sets into one common result, mixed methods combine these two research traditions to provide complementary viewpoints, to provide a complete picture of phenomena, to expand understand the phenomena, and to compensate for the weaknesses of one method. [32, 34, 35]

Among different design patterns for mixed-methods research (Table 2.2), triangulation is the most commonly used [34]. In triangulation, data collection and analysis are carried out independently for QUAN and QUAL methods with no preference, and the final inference aims at creating a broad picture of the phenomenon [34]. Three possible outcomes can be expected in these studies [36]:

1. the convergence of results in which both results lead to the same conclusions,
2. the complement of results in which the different results highlight different aspects of the same phenomenon, or
3. the results are divergent or contradictory.

The ideas of triangulation and other mixed-methods designs (Table 2.2) have already been used in quality evaluation research although researchers have not explicitly expressed the relationship

Mixed-method, design	Design pattern	Purpose
Triangulation design	Independent collection of QUAN and QUAL data. Interpretation based on both data sets	Comparison of QUAN and QUAL results for a broad interpretation of the results
Embedded design	One data set is used in a supplemental role in studies primarily based on the other data set	Additional qualitative expressions about quantitative results (e.g., supporting decisions about further studies or tasks)
Explanatory design	Two-step design. First collection of QUAN, then QUAL	QUAL data may be needed to explain unexpected results or to detect errors in the QUAN research design
Exploratory design	Two-step design. First collection of QUAL, then QUAN.	QUAL data may be needed to explain unexpected results or to detect errors in the QUAN research design

Table 2.2. – The four mixed-methods design approaches according to [34]

to this methodological approach [14, 37, 38]. In light of this introduction of the basic concepts of this thesis, the next section will now present different approaches that currently exist in audiovisual quality evaluations with respect to standardized methods as well as new approaches in user-centered quality evaluations and mixed-methods research.

2.3. Research methods for perceived quality evaluation

In general, the goal of subjective quality evaluation is the optimization of critical components of a system with as little perceptual effect as possible. To attain this goal, various research methods are available representing different approaches from quantitative evaluations to descriptive analysis.

2.3.1. Standardized quantitative quality evaluation methods

Psychoperceptual quality evaluation aims at examining the relation between physical stimuli and sensorial experience following the methods of experimental research. It has been adapted from classical psychophysics of the 19th century, and derived evaluation methods later used both univariate and multimodal quality assessment [5, 39–41]. Currently, widely applied quality evaluation methods for assessment of audiovisual multimedia systems are standardized in technical recommendations by the International Telecommunication Union (ITU) or the European Broadcasting Union (EBU) [5, 39, 42–44] (see Table 2.3). The goal of these methods is to analyze quantitatively the perceived overall quality of a set of test items in a test laboratory. The resulting degree-of-liking per test item is expressed as Mean Opinion Scores (MOS) or Mean Satisfaction Scores (MSS) . Although a multitude of standardized evaluation methods exists and appropriate selection depends on the specific research question, psychoperceptual quality evaluation studies are generally characterized by a high level of control over the test variables, test laboratory settings, and test participants (also called assessors, viewers, or observers) [5, 39, 45]. Multimedia quality evaluation recommendations advise against using experts as test participants and recommend inviting non-experts who are not “directly involved in picture quality evaluation as part of their work and should not be experienced assessors" [5]. In this thesis, I will refer to non-experts as naïve test participants or naïve assessors in accordance with ISO EN 8586-2 [45, 46]. If a test participant had taken part in a quality evaluation study previously but did not have a technical background, we call him an experienced test participant or experienced assessor as defined by ITU Recommendation ITU-T P.831 [47].

Recommendation	Title	Research methods included
ITU-R BT.500-1	Methodology for the Subjective Assessment of the Quality of Television Pictures	*Double-stimulus methods*: Double-Stimulus Impairment Scale DSIS Double-Stimulus Continuous Quality-Scale DSCQS Simultaneous Double Stimulus for Continuous Evaluation SDSCE *Single-stimulus methods*: Single Stimulus Continuous Quality Evaluation SSCQE Single Stimulus with Multiple Repetition SSMR
ITU-R BT.1438	Subjective Assessment of Stereoscopic Television Pictures	Refers to the assessment methods of ITU Recommendation ITU-R BT.500-11 [39]
ITU-T P.910	Subjective Video Quality Assessment Methods for Multimedia Applications	Absolute Category Rating ACR Degradation Category Rating DCR Pair Comparison method PC
ITU-T P.911	Subjective Audiovisual Quality Assessment Methods for Multimedia Applications	ACR, DCR, PC, SSCQE
EBU BPN 056	SAMVIQ – Subjective Assessment Methodology for Video Quality	SAMVIQ

Table 2.3. – Relevant recommendations for audiovisual quality evaluations [5, 39, 43, 44, 48]

In psychoperceptual studies, the range of quality being tested and the research question define the applicable method. A wrong selection of methods as a threat to validity can lead to invalid results and wrong conclusions (see section 2.1). Basically, two different sets of evaluation methods exist (see Table 2.4). Single stimulus methods are applicable in tests with a large quality range and detectable differences between stimuli. In contrast, pairwise or multiple stimuli methods are powerful for the evaluation of small detectable differences among the test stimuli. A review of existing standardized research methods can be found in Jumisko-Pyykkö and Strohmeier [215]. In audiovisual quality assessment, Absolute Category Rating (ACR) [39] and Subjective Assessment Methodology for Video Quality (SAMVIQ) [42, 49] are popular candidates from single stimulus and multi-stimulus methods, respectively.

Absolute Category Rating (ACR)

Absolute Category Rating (ACR) is standardized in the ITU recommendations ITU-T P.910 and ITU-T P.911 [5, 44]. ACR is a test method that is easy and fast to implement, and the presentation of the stimuli is similar to that of the common use of the systems. Test stimuli are presented consecutively and rated independently retrospectively (Figure 2.2). For quality judgment, the recommendations propose a five-level quality scale. However, they also stress that more detailed scales (9- or 11-point quality scales) can be used if higher discrimination power is needed [5].

	SSCQE	ACR	DSIS	DSCQS	SAMVIQ
Reference	No reference	No reference	Explicit reference	Hidden reference	Explicit and hidden reference
Comparison	Single stimulus	Single stimulus	Double stimulus	Double stimulus	Multiple stimuli
Rating	Continuous	Retrospective	Retrospective	Retrospective	Retrospective, but rating can be adapted several times
Scale	5-point continuous scale	5-point continuous scale (or higher if required)	5-grade impairment scale	5-point continuous scale	5-point continuous scale
Stimuli	from 60 seconds up to 20 minutes	10 seconds	10 seconds	10 seconds	max. 15 seconds

Table 2.4. – Overview of the differences in the implementation of psychoperceptual quality evaluation methods [5, 39, 48]

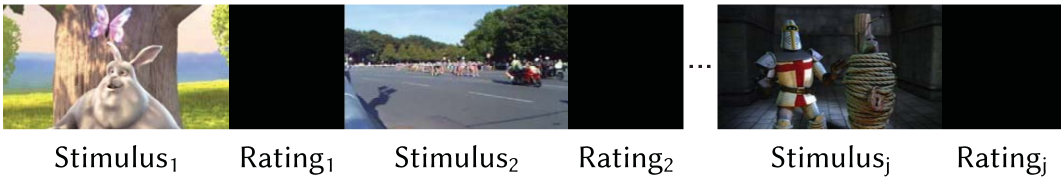


Figure 2.2. – Presentation structure of Absolute Category Rating for j items tested [5]. Test stimuli are presented consecutively and rated retrospectively.

Subjective Assessment Methodology for Video Quality (SAMVIQ)

The Subjective Assessment Methodology for Video Quality (SAMVIQ) (also known as the EBU method) [42, 49] was derived from ITU's DSCQS method [39] to offer a test methodology for multimedia. Blin [49] describes SAMVIQ as an efficient method of assessment of a large range of image quality because it provides reliable discrimination at both high and low quality levels. SAMVIQ uses hidden and explicit references in a multi-stimulus test environment. In contrast to DSCQS, the test participant has the possibility of accessing more than two stimuli at the same time. The direct comparison of multiple stimuli makes SAMVIQ able "to discriminate low qualities as well as high qualities" [49]. All stimuli are evaluated one after the other on a continuous scale from 0 to 100 with five explicit quality levels (excellent, good, fair, poor, and bad). Each stimulus is thus compared to an explicit reference to determine the best quality that can be achieved in the test. During the test, all stimuli are available at all and they can be repeated and reevaluated as often as needed. In addition to the explicit reference, there is a hidden reference of the same quality level as the explicit one, but it is not indicated in the test description. The hidden one acts as an anchor to check the performance of the test participants. With respect to the characteristics of test stimuli, Blin [49] and Kozamernik et al. [42] explain that stimuli of a maximum length of 15 s are sufficient to obtain a stabilized and reliable quality score. Blin [49] tested the performance of the SAMVIQ method in terms of reliability and stability by comparing the standard errors of a SAMVIQ study to those of a DSCQS-based study with the same stimuli in the test conditions. The results show that SAMVIQ renders better results with lower standard deviations. A second test was conducted to assess the stability of the test results based on results of two independent samples, which indicated stability by showing very high correlation between groups [49].

Comparison of ACR and SAMVIQ

Absolute Category Rating and SAMVIQ were compared in a few comparison studies [50, 51]. The studies show that ACR and SAMVIQ can produce comparable results. SAMVIQ results tend to have greater accuracy and better differentiate stimuli with a number of about 25 test participants. However, ACR results can be improved by increasing the number of test participants [50, 51]. Within the comparisons, ACR showed excellent inter-laboratory and between-group reliability [50]. Referring to validity of test results, inter-laboratory comparability of the results is very important to research methods. Beyond comparison of results, ACR was found to be easier to implement and allowed for faster evaluation and a higher number of test stimuli. Brotherton et al. [50] suggest that "ACR tests could present for assessment [at least] twice as many test sequences as the SAMVIQ tests." In addition, SAMVIQ is regarded as being artificial compared to ACR because SAMVIQ allows participants to replay test sequences and adapt the ratings as often as they want. It may lead to a more artificial test method compared to ACR because in real viewing situations, observers do not normally review content [50]. Although this systematic comparison of research methods is important for the purpose of research-question-related method selection, all the psychoperceptual evaluation methods leave other valuable questions unanswered. Because they limit their evaluation to a one-dimensional understanding of quality in terms of Mean Opinion Scores, the presented methods do not conform to the multifaceted understanding of quality in terms of Quality of Experience (see section 2.2) [9].

2.3.2. User-oriented evaluation methods

Recently, conventional psychoperceptual methods have been extended from one-dimensional hedonistic assessments to include more use- and goal-oriented actions (Table 2.5). Quality of Perception (QoP) measures quality as a multidimensional construct of cognitive information assimilation and satisfaction, constructed from enjoyment and subjective, but content-independent perceived quality [52–55]. The method introduces an approach that allows assessing users' satisfaction with the presented quality and their ability to analyze, synthesize, and assimilate information from the content. QoP has been slightly adapted during its development in constructive research [54, 55]. In recent use, Quality of Perception is defined as the sum of the level of information assimilation, QoP-IA, and satisfaction, QoP-S [55]. QoP-IA is measured with the help of questions asking about information seen in the content. QoP-IA is finally expressed as the proportion of correct answers to all questions asked. QoP-S is divided into two measures. Test participants are asked to rate the overall quality of a stimulus on a 5-point scale. In addition, test participants express their enjoyment of the content on a second 5-point scale. QoP studies have shown that an extension of existing methods is needed to obtain deeper understanding of subjective quality and its impact on the user going beyond pure hedonistic judgments.

	Quality of Perception	Method of Limits	Acceptance Threshold	Evaluation in the context of use
Purpose	To open up understanding of quality from a one-dimensional hedonistic measure to a combination of quality satisfaction and information assimilation [55]	To identify the threshold at which quality becomes unacceptable [56]	To identify the threshold of minimum acceptable quality and relationship to satisfaction scores on overall quality [57]	To extend the external validity of results by evaluations in the actual context of use [58]
Methodology	Information assimilation is collected from questions about the content; satisfaction is measured on two 5-point scale for dimensions of enjoyment and quality	Quality is increased and decreased continuously by variation of different quality parameters and test participants indicate if they find the current quality acceptable or unacceptable	Bidimensional research method which evaluates Acceptance of quality on binary yesno scale and Satisfaction with quality on a 11-point unlabelled scale (ACR)	Hybrid methodological framework which complements quantitative evaluations with a set of tools for planning, data collection and data analysis that to identify surrounding contextual factors
Analysis	Information assimilation is measured as the percentage of correctly answered questions. The final QoP is the sum of information assimilation and satisfaction measures	Final quality measure is the ratio between acceptable and unacceptable quality per test sequence	Either analysis based on frequencies of acceptance per independent variable or identification of values ranges for acceptable and unacceptable overall satisfaction ratings	Separate analysis of the different data sets and integration in a final step to achieve complementation and check for convergence of results

Table 2.5. – Overview of user-oriented evaluation methods for audiovisual quality assessment

Other user-oriented evaluation methods focus on the evaluation of the user's quality acceptance as an indicator of service-dependent minimum quality. Acceptance of overall quality is an important factor in the success of modern multimedia systems. McCarthy et al. [56] evaluated quality acceptance based on the classic Fechner psychophysical method of limit [40]. In their approach, McCarthy et al. gradually vary the quality of a stimulus in a continuous rating task. Test participants indicate the points at which the quality changes from acceptable to unacceptable or vice versa in a continuous assessment. Acceptance is finally expressed as perception of total time having acceptable quality. Although the method has shown the importance of measuring acceptance, it is criticized because it cannot be applied to measure quality clearly below or above the threshold [9]. Another approach to measuring quality acceptance is Acceptance Threshold. Jumisko-Pyykkö et al. [57] introduced this approach as an extension of standardized psychoperceptual evaluation methods. In their approach, Jumisko-Pyykkö et al. combine a binary rating of the acceptance of the overall quality (yes-no) and the rating of satisfaction with overall quality on a 11-point, unlabelled ordinal scale. The final goal of the method is “to locate the threshold of minimum acceptable quality that fulfills user quality expectation and needs for certain application or system.” [9] Both ratings are done retrospectively and independently per test item. The results of the binary acceptance rating can be used either to identify the threshold based on the frequencies of acceptance per independent variable or as identification of values ranges for acceptable and unacceptable overall satisfaction ratings [9, 57]. Jumisko-Pyykkö et al.'s Acceptance Threshold has become part of the User-Centered Quality of Experience evaluation framework. Within the development of the framework, the Acceptance Threshold has been applied in several studies in the laboratory and in natural contexts of use [57–60].

2.3.3. Evaluations in the context of use

Psychoperceptual evaluations and the user-oriented quantitative methods lack external validity due to the high level of control over test variables and laboratory environments. The UC-QoE evaluation framework includes research methods for quality assessments in the context [9, 58]. The goal of these evaluations is to relate quality evaluations to the actual contexts of use to make them more generalizable from artificial laboratory settings to the field to gain high external validity and realism within the studies.

Context thereby is a multidimensional construct [61]. As defined by Dey et al. [62], "context is any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves." More detailed concepts for the description of context, especially in the domain of research on mobile devices, differentiate among physical, temporal, social, task, and technical contexts [8, 61]. A description of the different aspects of contents is given in Table 2.6. In general, the need for evaluations in the contexts increases with increased dynamics and heterogeneity of the expected contexts of use for a system under evaluation [58].

The novelty of current research approaches to contextual evaluations is that they complement conventional quantitative evaluation methods with tools to identify impacting factors on the evaluation task in the context [8, 61]. The basic evaluation follows standardized methods and extended approaches like Acceptance Threshold. In addition, contextual researchers use a light-weight mobile usability lab to capture events during the evaluation. Assisted by short semi-structured interviews, this approach allows for detailed description and knowledge about the contextual situation during evaluation. The whole evaluation, which usually takes place in a set of different contexts, is closed using a broader semi-structured interview and targets an elicitation of individual experiences about the contexts and related quality for each test participants. The approaches underscore the importance of developing understanding of participants' experiences and individual quality factors in relation to their individual user requirements in different settings [58]. Jumisko-Pyykkö and Utriainen's "Hybrid Method for Quality Evaluation in the Context of Use" is one of the key methods within the UC-QoE evaluation framework [9].

The contextual evaluation approaches have shown how important it is to collect data beyond quantitative ratings to meet the requirements of modern quality evaluations with respect to the complex concepts of Quality of Experience [58, 60]. These methods require knowledge about in-

Aspect of context	Definition	Example
Physical context	Physical context defines the physical location in which the interaction occurs. It can also include virtual spaces or descriptions about movements between locations.	Typical physical contexts for mobile 3DTV are cafés, waiting rooms or public transports.
Temporal context	Temporal context defines the time and duration in which the interaction occurs. It also reflects the situation before and after the use or synchronized actions between partners (like talking on the phone is synchronized between the two partners on the phone).	Temporal contexts for mobile 3DTV refers for example to short time viewing, or use of the system after work.
Social context	Social context defines the other persons that are present during the interaction with the system occurs. It includes descriptions of the characteristics of the other persons as well as their roles with respect to the interaction.	While mobile 3DTV is often regarded to be single, focused watching of videos, other applications like gaming or videophony can also include shared use of the system between several users, e.g., in a group of young people.
Task context	Task context describes the tasks that are fulfilled by users while the interaction with the system occurs. It relates to multitasking and includes also possible interruptions that are caused by parallel tasks.	While watching mobile 3DTV on a bus ride, the user will also focus on the track so that he does not miss his bus stop.
Technical context	The technical context describes the interaction in relation to other technical devices or networks.	The technical context in mobile 3DTV includes the network over which the video signal is broadcasted or the service from where a video-on-demand is ordered.

Table 2.6. – Different aspects of context with definitions according to the classification by Jumisko-Pyykkö and Vainio [61]. Examples for each aspect are selected in relation to user requirements for mobile 3D television and video systems [216].

terpreted quality and understanding of test participants' quality factors. However, all quantitative approaches lack the possibility of studying the underlying quality rationale of the users' quality perception.

2.3.4. Descriptive quality evaluations: methods and application

The goal of descriptive quality evaluation is "to provide complete sensory descriptions of an array of products, provide the basis for mapping product similarities and differences, and provide a basis for determining those sensory attributes that are important to acceptance." [63] This general definition by Stone and Sidel shows the difference between quantitative evaluations and the descriptive approaches. While psychoperceptual evaluation methods are suitable methods to measure the excellence of a stimulus, descriptive methods target the elicitation of its specific quality attributes. The basic idea of applying descriptive methods in multimedia quality evaluations has been that test participants are asked to describe their quality factors or the reasons for a certain overall quality rating. Those descriptions can be seen as the complement to excellence evaluation to acknowledge the dualistic nature of quality. Common to all approaches is this elicitation of individual quality factors in terms of qualitative data. Descriptive evaluations bring up terms, descriptions, and interpretations of quality, not quantitative ratings. Two general approaches exist in the domain of multimedia quality evaluations: interviews and sensory evaluation, which differ in terms of vocabulary elicitation methods; methods of analysis; and characteristics of participants (Table 2.8).

For descriptive audiovisual evaluations, interviews are currently applied. In the existing interview-based methods, naïve participants describe explicitly the characteristics of stimuli, their degradations or personal quality evaluation criteria under free-description, and if necessary, stimuli-assisted description tasks [37, 58, 64, 65]. The goal of these interviews is to generate terms to describe the quality and to check that the test participants perceive and rate the intended quality aspects. Semistructured interviews are commonly applied. They are especially applicable to relatively unexplored research topics, constructed from main and supporting questions, and, compared to open interviews, they are less sensitive to interviewer effects [4]. The frameworks of data-driven analysis apply for hypothesis-free analysis of qualitative data. The Grounded Theory framework by Strauss and Corbin [66] follows a three-stage model of open, axial, and selective coding. After the interviews are transcribed, open coding is applied in which concepts and related categories in the data are identified and related to the parameters of the study. Axial coding identifies causal relationships between the concepts and seeks to build explicit connections between categories and possible sub-categories. The final selective coding then involves the process of identifying core categories and systematically relates the core categories to other categories. The outcome of the Grounded Theory based analysis is described in terms of the core categories and the most commonly appearing sub-categories [37, 38, 67]. Multidimensionally, these identified components can be analyzed further in correspondence analysis to identify intercategorical relationships. Commonly, the data analysis is conducted by two independent researchers to increase the reliability of the results. Interviews are easy to implement and conduct, but the data analysis requires at least two experienced researchers to obtain reliable results. Another descriptive evaluation approach is offered in methods of sensory evaluation. Section 2.4 will describe these methods and the different existing approaches separately although they can also be classified as descriptive quality evaluation techniques.

Interview-based approaches offer a straight-forward solution for descriptive quality evaluations because they explicitly ask test participants about their individual quality factors. However, they present limitations to measuring the sensation of these quality attributes. Modeling of attributes to understand the dominating factors of the underlying quality rationale is hard to achieve. Solutions for this problem can be found in sensory evaluation, which is a research discipline widely used in the food and odor sciences. In sensory evaluation, test participants' expressions of quality attributes are used to rate test items in a task after the attribute elicitation.

2.4. Sensory evaluations: methods and applications

2.4.1. Sensory evaluation methods

Another descriptive evaluation approach is offered by methods of sensory evaluation. Originating from the food sciences, "sensory evaluation is a scientific discipline used to evoke, measure, analyze and interpret reactions to those characteristics of foods and materials as they are perceived by the senses of sight, smell, taste, touch and hearing"[1] [63] . Sensory evaluation (also called sensory analysis) covers two main classes of evaluation methods. Discrimination tests simply assess whether two products being tested are different. In contrast, descriptive analysis aims at identifying specific product characteristics and uses those characteristics to evaluate them for a set of products on scales of perceived intensity [40, 68].

The assumption upon which descriptive analysis is built is common to all the existing approaches. They assume that perceived quality is the result of a combination of several features and test participants can describe a specific feature by a verbal descriptor [40, 41, 69, 70]. In this thesis, I will use the term attribute to refer to this verbal descriptor. Attributes in sensory evaluation have to fulfill a set of requirements [40, 46, 71, 72] (Table 2.7). All the requirements contribute to the reliability of the attributes.

Requirement	Definition
Differentiation	The attributes focus on describing differences between attributes rather than common characteristics.
Precision	Each attribute describes exactly one quality factor.
Nonredundancy	An attribute must have little or, preferably, no overlap with other terms used.
Identification	In fixed terminologies, a pair of attributes and a related obvious quality factor, like blocking artifacts and blockiness, should be easily identifiable.
Recognition	The meaning of an attribute should be recognizable from its name or at least from a given definition.
Singularity	An attribute should relate to one quality factor rather than being a combination of terms; e.g., video quality is a function of several other attributes like blockiness, blur, and clarity.

Table 2.7. – Overview of the main requirements for attributes of sensory evaluation methods [40, 46, 71, 72].

In descriptive analysis, there is a differentiation between two principle classes of descriptive analysis methods in terms of attribute elicitation. Consensus vocabulary methods use attributes that have been developed as consensus vocabulary for a group of subjects. Individual vocabulary methods apply an individual vocabulary per participant [46].

2.4.1.1. Consensus vocabulary approaches

Consensus vocabulary methods have been developed for different purposes since descriptive analysis became popular in the 1950s. Lawless and Heymann [40] and Stone and Sidel [63] provide good overviews of the different methods like Flavour Profiling or Texture Profiling. The most significant contribution to the field of consensus vocabulary approaches is the Quantitative Descriptive Analysis (QDA) method . Introduced by Stone et al. [73] in 1974[2] and corrected the final methodology later in 1993 [63], QDA provided a full methodology from consensus vocabulary elicitation and evaluation procedure to a defined set of analysis methods. While former descriptive methods were designed to evaluate one specific aspect of the product, for example, flavour or texture, QDA

[1] Anonymous definition from the Institute of Food Technologists as cited by Stone and Sidel [63]

[2] The Quantitative Descriptive Analysis was originally published by Stone et al. [70] in Food Technology.

introduced the holistic approach targeting a complete description of the sensory characteristics of products.

Quantitative Descriptive Analysis starts with the elicitation of a consensus vocabulary and extensive panel training. Consensus about the set of attributes is achieved by group discussions in which experts first develop an extensive list of attributes in a stimulus-assisted task. This list is then reduced to a reasonable number of attributes. On this vocabulary, agreement among the experts is achieved by defining references and definitions that are then presented in several training sessions [40, 46, 63]. Stone and Sidel estimate the time to achieve a consensus vocabulary as roughly 7 - 10 hours split into several sessions [63]. Evaluation of products is then conducted on a line scale. This scale is a blank line labeled with a word anchor on either end. Each attribute is attached to one of these scales on a scorecard, and experts rate the sensation of the attributes for each product specifically. The scores obtained from the evaluation can be analyzed with various methods. QDA methodology proposes a set of analysis methods (e.g., analysis of variances, multivariate data methods, 'spider web' plots) to analyze and visualize QDA results, but also to assess test participants' consistency of ratings or the performance of the whole expert panel [40, 63] (see Stone and Sidel [63] for an overview). Consensus vocabulary methods have been widely applied in sensory evaluations (for an overview see [40]). However, the extensive group discussions until consensus is reached, the need for new vocabulary development when products change, and the need for experts have been criticized and seen as factors limiting the success of the methods [74, 75]. The necessary agreement of the experts in terms of meaning and sensitivity of an attribute is "often very difficult, if not impossible, to obtain" [74]. In addition, QDA and related methods have not been regarded as being useful for consumer research [75].

2.4.1.2. Individual vocabulary approaches

Free-Choice Profiling (FCP) was introduced by Williams and Langron [74] in 1984 and represented a radically different approach to the common consensus vocabulary methods. Instead of highly trained experts, consumers were selected as test participants. Instead of extensive training of a test panel for common consensus about the quality attributes and measures of sensations, Free-Choice Profiling allows test participants to develop their own idiosyncratic attributes. The requirements for these attributes are the same as for consensus vocabulary (Table 2.7). Objective, nonhedonic attributes that the test participants are able to use consistently are needed. However, test participants are free to select and idiosyncratically describe those product characteristics that impact their sensations [40]. In contrast to consensus vocabulary methods, little training is required because only the individual test participant needs to understand his attributes. After attribute elicitation, the evaluation task is similar to QDA except that every participant uses his individual scorecards. However, FCP also required new methods of analysis because standard univariate and multivariate statistics could not handle the individual ratings, also referred to as configurations. Proposing a geometrical scaling of the individual configurations to a group average, Generalized Procrustes Analysis (GPA) was introduced by Gower [76] and adapted for analysis of Free-Choice Profiling. The author will explain GPA in detail in section 3.3.2. Free-Choice Profiling has offered possibilities to bring descriptive evaluations to consumer research because no agreement on attributes is needed. In addition, FCP is faster to conduct and offers a cheaper alternative to consensus vocabulary methods [40]. However, application in different domains of research showed that the elicitation of attributes was difficult for test participants [77]. Supporting tasks like the Repertory Grid method (RGM) were applied. Repertory Grid was originally introduced as an interview technique in personal psychology [78] and was later proposed for attribute elicitation in sensory evaluations by [79–81]. The adapted Repertory Grid technique is a comparison of triads of products. A test participant is asked to separate one product from the triad and describe one attribute in which 1) the chosen product differs from the others and 2) the other two are comparable. By repetition of several triads, a construct of attributes setting the products in relationship is created. The attributes from these constructs are further used as the individual vocabulary. Although the Repertory Grid facilitates the attribute elicitation with its structured approach, systematic comparison between the Repertory Grid-based method and conventional FCP did not show any advantage to the RGM approach [77, 82]. Free-Choice Profiling was used for studies on a wide range of food and beverages (e.g., ham [83], coffee [84], and lager beers [85]). Comparison of Free-Choice Profiling and consensus vocabulary approaches have shown that results obtained from descriptive analysis with consumers and experts are comparable [75, 84, 86–88]. However, there are hints that FCP fails when perceptual differences between the products being tested become smaller and harder to discriminate [89].

2.4.1.3. Sorting-based descriptive approaches

Sorting-based approaches represent the third class of descriptive analysis techniques. While consensus and individual vocabulary approaches develop attributes used to discriminate a set of products, sorting techniques first develop a construct of similarities and dissimilarities between products, and the resulting groups are then described by individual attributes. The goal of these methods is to provide a model of the relationship of a set of products without time-consuming attribute elicitation and training. All sorting methods adapt the ideas of Free-Choice Profiling and allow test participants to use their own attributes [90].

In the Perceptive Free Sorting approach [91–93], test participants sort all products into different groups according to the perceived similarities among them. They are allowed to open up as many groups as needed. After finalizing the sorting, every group is described with test participants' individual words. Projective Mapping [94, 95] or (Sorted) Napping [96, 97] extend the sorting task by introducing similarity as a measure between products. Test participants position the products on a sheet of paper, the 'nappe,' or tablecloth. Products that are perceived similar to each other are thereby placed close to each other. Different perceptions of different products means that they are placed far from each other. Again, test participants are asked to describe each product or groups of products on the tablecloth.

The methods propose different sets of analysis for the sorting or napping data. These methods are either generalized derivates of Multidimensional Scaling (MDS) (e.g., DISTATIS [98]) or Multiple Factor Analysis (MFA) [99]. They all create common similarity maps among test participants, allowing for taking into account individual differences as well as mapping of attributes into the final map. The results of sorting and mapping approaches have been favorably compared to vocabulary-based approaches [91, 100].

2.4.2. Multidisciplinary applications of sensory evaluations to study users' product perceptions

Sensory evaluations have shown different fields of applications in which these techniques delivered useful results. After introducing the different methods of sensory evaluations, the following review will present possibilities of sensory evaluations to study different research problems with respect to methods based on individual vocabulary. Especially, research problems that represent interesting applications for audiovisual quality research were chosen.

2.4.2.1. Applications in the food sciences

Sensory evaluations have their roots in the food sciences, and the methods have been widely used in this domain of research to study different research problems. In general, all sensory evaluation techniques were applied to discriminate a set of products according to users' perception. However, the studies also showed applicability to study differences among user groups and among individual test participants.

Detection and discrimination of products based on sensory differences The main purpose of sensory evaluations has been to achieve discrimination of different products based on participants' sensory perceptions. Within a product development phase, these techniques are useful in understanding sensory problems that occur with a product. The results of the evaluations can be used to monitor the performance of a product in comparison to competitor products and to understand possibilities for user-driven improvements on a sensory level [101]. Within the food sciences, good examples of studies targeting product discrimination can be found concerning beverages [74, 82, 84, 85] and groceries [102, 103]. These studies show that sensory evaluation helps to understand the multidimensional characteristics of taste by identifying and modeling the dominating factors of users' perceptions. Other studies were able to show that the categorization of taste and flavor of products can be dominated by other modalities, like the perception of beer taste being dominated by vision [104] or the perception of flavor by different textures of the product [105].

This application of sensory evaluations is highly relevant for the goals of this thesis because it provides a possibility of modeling perceptual differences among test stimuli with respect to users' quality attributes. This ability addresses the shortcomings of quantitative evaluations and can provide an alternative for descriptive audiovisual quality evaluations by interviews.

Comparison of different user groups While the original studies in the field of sensory evaluations targeted the discrimination of products, current studies applied sensory evaluation tasks to understand the performance of different user groups within the discrimination task. A first comparison of user groups in these studies is related to experts versus naïve test participants [89, 106–109]. These studies were often conducted to evaluate the reliability of Free-Choice Profiling techniques in comparison to the common QDA evaluation. However, the general comparison of quality evaluation of experts and naïve participants is highly relevant to audiovisual quality evaluations [5, 39]. Other comparisons have been conducted to assess the effect of training on the results of sensory evaluation tasks [93, 110]. While expert panels are trained for evaluating specific sensations and dimensions of taste or flavor in the long-term, trained assessors are trained only on the specific test items within a short-term period to familiarize them with the assessment and the products. The results of the studies show that trained and untrained participants produce comparable results, but attributes from trained assessors are more specific and consistently applied. Further reported are applications of comparison panels or user groups and interlaboratory [111] and cross-cultural [112] studies.

Increasing the granularity of differences between assessors, sensory evaluations have started to compare and model individual differences between assessors [113]. These studies do not assume that they test different user groups, but use sensory analysis to see whether differences between the assessors can be found in the results. Especially, these studies target differences in individual weights of quality factors when describing complex sensory sensations, like creaminess of ice cream [103]. These studies also led to the development of new methods of analysis, like the Multiple Factor Analysis [96], that are able to present results of the sensory evaluation under the constraints of individual differences.

Within audiovisual quality evaluations, individual differences are considered to be an important factor in understanding quality rationales [9]. Especially, dominance of different modalities between assessors may lead to different results of multimodal quality perceptions, for example, dominance of visual or auditory mode in audiovisual perception [114].

2.4.2.2. Sensory evaluation approaches in audiovisual research

First adaptations of sensory evaluation methods in the domain of audio and video quality research have shown that these techniques provide useful information for deeper understanding of perceived quality. The “RaPID perceptual image description method” (RaPID) was introduced by Bech et al. [69] in 1996. RaPID is based on a descriptive analysis assuming that image quality is the result of a combination of several attributes and that these attributes can be rated by a trained panel of assessors [41, 69]. RaPID is an adaptation of QDA and, therefore, relies on experts’ consensus vocabulary. During evaluation, trained test participants rate quality based on the vocabulary. A multistep procedure contains 1) extensive group discussions in which panel members first develop a consensus vocabulary of quality attributes for image quality; 2) a refinement discussion in which the panel agrees about the important attributes and the extremes of intensity scale for a specific test according to the test stimuli available; 3) an evaluation task in which each test participant applies each attribute for a set of stimuli in a pair comparison of the test stimulus and a fixed reference. RaPID requires extensive and time-consuming panel training, can be sensitive to context effects, and requires an experienced researcher to conduct the experiments [69]. A comparable methodology is used for audio evaluation in the Audio Descriptive Analysis and Mapping (ADAM) technique [115].

In contrast to consensus vocabulary profiling, Lorho's Individual Vocabulary Profiling Method (IVP) [116–118] is a descriptive quality evaluation for naïve participants. His work was the first approach in multimedia quality assessments to use individual vocabulary from test participants to evaluate quality. The procedure contains four steps. 1) Participants become familiar with describing the attributes of stimuli, and they develop their individual vocabulary in two consecutive tasks. 2) An attribute list is generated in a triad stimulus comparison using an elicitation method called Repertory Grid Technique. 3) The developed attributes are used to generate scales for the evaluation. Each scale consists of an attribute and its minimal and maximal quantities. 4) Test participants train and evaluate quality according to the attributes developed. The data are analyzed through hierarchical clustering to identify underlying groups among all attributes and Generalized Procrustes Analysis [76] to develop perceptual spaces of quality. Compared to the other descriptive methods, the four-step procedure for individual vocabulary training can be time consuming. However, analysis of IVP is relatively easy, and the researcher's interpretive process comes at the very end compared to interview-based methods. The Repertory Grid Methodology was earlier applied by Berg and Rumsey [119] to identify spatial attributes of sound reproduction.

Although the review of applications of sensory evaluations shows that there are various methods for studying perceived multimedia quality quantitatively and qualitatively, the methods do not combine both approaches (Table 2.8). The author sees a challenge of modern evaluation methods also in the combination of quantitative and descriptive data sets in accordance with the theory of mixed-methods research.

2.5. Mixed methods in audiovisual quality evaluation

In multimedia quality evaluation methods, triangulation is the commonly applied mixed-methods design. Jumisko-Pyykkö et al. [37] have introduced an approach of combined quantitative psychoperceptual evaluation and post-task interviews to explore experienced quality factors for audiovisual quality with naïve test participants. The psychoperceptual evaluation thereby follows the ITU recommendations for subjective quality evaluations [5, 39]. The experienced quality factors were collected using a semistructured interview in which test participants described freely their evaluation criteria used during the quantitative evaluation. Data-driven analysis, following the framework of Grounded Theory [66], was used in the interview analysis. The analysis and interpretation of both data sets were first carried out independently and then integrated to support each other's conclusions. The complementing results have indicated that experienced quality is constructed from

	Interview-based approach	Consensus vocabulary profiling approach	Individual vocabulary profiling approach
Methods using this approach	Interpretation-Based Quality evaluation [38], Experienced quality factors [37]	Flavor Profile Method [40], Texture Profile Method [40], Quantitative Descriptive Analysis [40], RaPID [69], ADAM [115]	Free-Choice Profiling [75], Flash Profiling [120]
Vocabulary Elicitation	Interview	Group discussions and consensus attribute list	Individual attribute list, supporting task like Repertory Grid Method applicable
(Statistical) Analysis	Coding (e.g., Grounded Theory) and Interpretation	ANOVA, MANOVA, PCA, MFA	GPA, (H)MFA
Participants	15 or more naïve test participants	Around 10 highly trained participants	Around 15 naïve test participants
Used in Mixed-Methods approaches	Yes	No	No
Applied in audiovisual quality research	IBQ [38, 121, 122], Experienced Quality Factors [37], Contextual quality evaluations [58, 60]	RaPID [69], ADAM [115]	IVP [116, 117]

Table 2.8. – Descriptive methods

the impressions of low-level features of stimuli (e.g., audio, video, audiovisual impairments), high-level factors (e.g., relationship of quality to use, content), and the most varied variable representing the peaks or extremes of quality [14, 37]. This method may suffer from inaccuracy because the descriptions are related to a set of stimuli instead of a single stimulus. However, the descriptive task can be conducted quickly and can be easily adapted to the quality evaluations in challenging circumstances (e.g., field) [60].

Triangulation is also applied in the method called Interpretation Based Quality (IBQ) [38, 67, 121], adapted from [91, 92]. Comparable to Jumisko-Pyykkö et al.'s approach, IBQ also follows a two-step evaluation procedure with naïve participants: 1) a classification task using perceptive free-sorting combined with an interview-based description task for quality attribute elicitation and 2) the psychoperceptual evaluation based on one quality attribute for quantitative evaluation. In the perceptive free-sorting task, test participants form groups of similar items and describe the characteristics of each group. The free-sorting task with naïve participants produces comparable results to consensus vocabulary approach with expert participants in terms of describing the same sensations and the related wording of the attributes [91, 100]. However, the costs of free-sorting are lower because of naïve test participants, missing training, and fast assessment of a large test set [100]. Extending the idea of a free-sorting task, IBQ allows combining preference and description data in a mixed analysis to better understand preferences and the underlying quality factors at the level of a single stimulus [38]. However, the analysis of interview-based methods for large data sets is time consuming because it requires a multistep procedure and interrater reliability estimations. In contrast to the original definition of the method [38, 67], the term IBQ has been inconsistently used in later studies and has referred to only monomethodological designs (complementing psychoperceptual evaluation) and variable procedures of descriptive tasks, especially in 3DTV-related research [121–123]. In this paper, I will discuss IBQ as it was originally presented by Radun et al. [38] when referring to it as mixed-methods research approach.

Summarizing the review of related work, audiovisual quality evaluation research has slowly started to extend its approach from quantitative excellence evaluation to descriptive and mixed methods to create a broader understanding of quality rationales. The work is integrated into efforts of user-centered evaluation of Quality of Experience, emphasizing the importance of high-level human perception integrating affective, knowledge-based, or usage-based dimensions into evaluations. However, the existing approaches of descriptive evaluations do not offer possibilities of modeling individual quality factors in common models. The methods of sensory evaluation and their applications in the food sciences show applicability of these methods to study relevant questions in audiovisual quality evaluation. The methods of individual vocabulary profiling have shown to be applicable to naïve assessors. First attempts to integrate these methods are found in unimodal audio assessments but still lack the possibility to link these descriptive models to quantitative preferences. The main goal of this thesis is to present Open Profiling of Quality (OPQ) as a new quality evaluation method, which follows the methodological considerations of mixed-methods research approach to create a deeper understanding of multimodal quality being applicable to naïve participants.

3. Open Profiling of Quality

This section introduces Open Profiling of Quality as a mixed-methods research approach for audiovisual quality evaluations. OPQ combines psychoperceptual and descriptive evaluation into a combined evaluation approach. Beside theoretical considerations about planning, conducting, and analyzing an OPQ study, I also introduce the whole method according to the standard reporting scheme for audiovisual quality evaluation studies. This is an important aspect in introducing a new research method with respect to reliability and validity of the results. Parts of this sections have been published in Strohmeier et al., "Open Profiling of Quality: A Mixed Method Approach to Understanding Multimodal Quality Perception," Advances in Multimedia, vol. 2010, Article ID 658980, 28 pages, 2010. doi:10.1155/2010/658980 [214] and Strohmeier et al., "The Extended-OPQ method for User-centered Quality of Experience evaluation: A study for mobile 3D video broadcasting over DVB-H," EURASIP Journal on Image and Video Processing, special issue on Quality of Multimedia Experience, vol. 2011, Article ID 538294, 24 pages, 2011, doi:10.1155/2011/538294 [217].

3.1. General considerations

Open Profiling of Quality (OPQ) is a mixed-methods approach that combines the evaluation of quality preferences and the elicitation of idiosyncratic experienced quality factors. It uses quantitative psychoperceptual evaluation and, subsequently, an adaptation of Free-Choice Profiling. OPQ is 'open' in terms of being "free from limitations, boundaries, or restrictions" [124] and "accessible to new ideas" [125] to understand the participants' construct of overall quality without restricting or constraining their descriptions. The extension of quantitative methods with sensory evaluations encompassed evaluation of individually perceived quality with all the participants' senses [63]. Hence, sensory evaluation approaches guaranteeing that the system parameters are assessed holistically in terms of overall quality. The term 'profile' refers to the representation of "the outline [of something]" [125], targeting some kind of identity, characteristics, descriptions, and structure for the phenomenon under study. OPQ conceptualizes test participants' individual quality factors in common spaces and allows linking these models with quantitative preferences of quality. The idea of descriptive evaluation goes beyond identification of which parameter is superior or whether parameters in the implementation of a new system should be changed. Although we cannot assume that every verbal descriptor directly relates to a specific quality parameter, the mix of quantitative and descriptive evaluation deepens the understanding of underlying quality rationales beyond specific research questions [63, 71]. The goals of an Open Profiling of Quality study are as follows:

1. to define the excellence of overall quality for different stimuli using quantitative psychoperceptual evaluation methods;
2. to understand the characteristics of quality perception by collecting individual quality attributes using qualitative sensory profiling methods;
3. to combine quantitative excellence and qualitative sensory profiling data to construct a link between preferences and quality attributes;
4. to provide a test methodology that is applicable to use with naïve test participants.

3.1.1. Research procedure

The original Open Profiling of Quality approach [214] as a research method consists of three subsequent parts (see Figure 3.1): psychoperceptual evaluation, sensory profiling, and external preference mapping. The studies with the first two methods are independently conducted in subsequent tasks. Their data can be combined in the latter part. During the development of OPQ in a constructive research approach, the procedure has been extended with a last step of data analysis. The component model aims at developing a general model of QoE components from the individual attributes. For an OPQ study, this step is not mandatory.

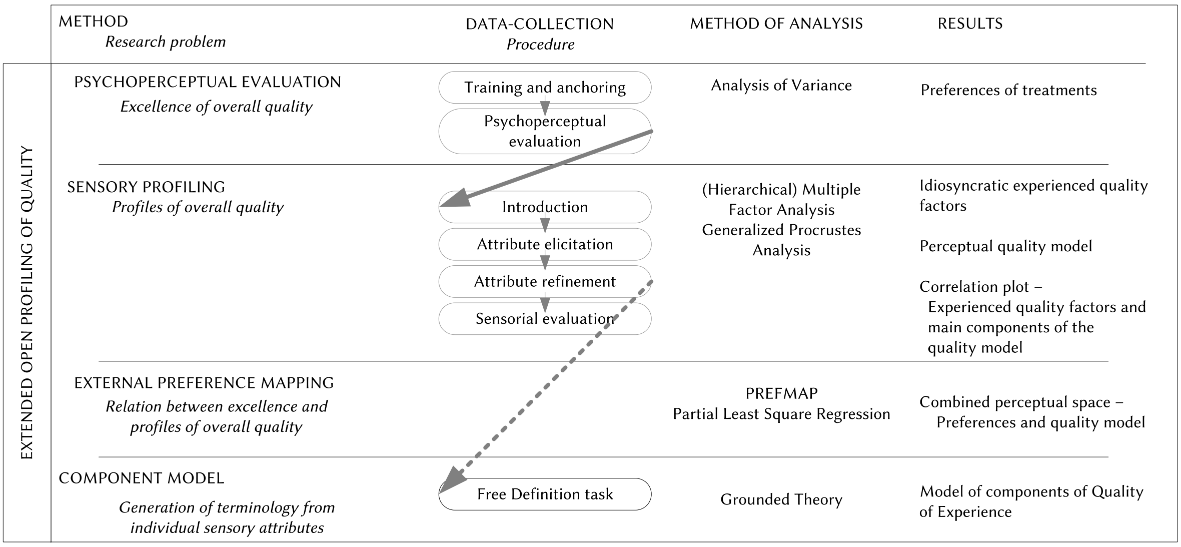

METHOD
Research problem
DATA-COLLECTION
Procedure
METHOD OF ANALYSIS
RESULTS
EXTENDED OPEN PROFILING OF QUALITY
PSYCHOPERCEPTUAL EVALUATION
Excellence of overall quality
Training and anchoring
Psychoperceptual evaluation
Analysis of Variance
Preferences of treatments
SENSORY PROFILING
Profiles of overall quality
Introduction
Attribute elicitation
Attribute refinement
Sensorial evaluation
(Hierarchical) Multiple Factor Analysis
Generalized Procrustes Analysis
Idiosyncratic experienced quality factors
Perceptual quality model
Correlation plot – Experienced quality factors and main components of the quality model
EXTERNAL PREFERENCE MAPPING
Relation between excellence and profiles of overall quality
PREFMAP
Partial Least Square Regression
Combined perceptual space – Preferences and quality model
COMPONENT MODEL
Generation of terminology from individual sensory attributes
Free Definition task
Grounded Theory
Model of components of Quality of Experience

Figure 3.1. – Overview of the subsequent parts of an Open Profiling of Quality study including their respective research questions.

3.1.2. Test participants

OPQ is designed to be applicable for naïve test participants with predefined sensory acuity criteria. In my understanding of 'naïve', the author follows the definition of ISO EN 8586-2 [45]. Naïve is defined as not meeting any particular selection criterion for assessment tests, neither having experience in the research domain nor in the evaluation task [45, 69]. Naïve participants are expected to give holistic quality evaluations and produce unbiased results due to lack of knowledge about the test stimuli and their production while expert assessors are trained for accurate, detailed, and domain-specific evaluation tasks (e.g., visual artifacts) [75, 126]. However, a common sensory acuity level is required for all test participants to make sure that the results are not biased by sensory inaccuracy. These tests encompass screening of visual acuity tests for myopia and hyperopia (Snellen index: 20/40), and color vision according to Ishihara [39], hearing threshold with respect to ISO 7029 [127], or, in given cases, 3D vision using the Randot Stereo Test (>60 arcsec). The common sensory level is an important requirement in establishing the validity of OPQ results. The sample selection contributes to the external validity of the study and defines how well the results from the sample tested generalize to some broader population of interests [6]. The recommended number of participants according to ITU recommendations is at least 15 [5, 39]. However, the author recommends 25-30 participants for the psychoperceptual evaluation to provide good statistical validity in within-subject designs [128]. For sensory profiling and the external preference mapping, a minimum of 12-20 participants is needed. This number of test participants assures that sensory analysis and external preference mapping have a sufficiently large data set [129] for valid results. However, it has been reported that a further increase of test participants may lead to an increase in error and noise in the analysis rather than useful additional information [82].

3.1.3. Scheduling the experiments

In OPQ studies, the psychoperceptual evaluation task is conducted prior to the sensory profiling. Although the order of the tasks may not have an impact on the outcome, as discussed by Faye et al. [100], it is recommended to begin with the psychoperceptual evaluation as assessors are "clear of influence" [100]. In addition, the following profiling task can be accomplished more precisely because of the already-existing comprehension of the stimuli in the test. Due to the duration of each individual study, OPQ experiments are usually divided into several sessions. A meaningful separation into sessions may depend on the number of test items. Depending on their number and the length of time needed for each item, as well as the final specific design of each part, psychoperceptual evaluation and sensory profiling will take 90-120 minutes. Commonly, the length of each part forces the researcher to conduct OPQ in two or three sessions. Variations in session division have been used in the constructive development of OPQ (see section 4).

3.2. Psychoperceptual evaluation

Research goal within Open Profiling of Quality: The goal of psychoperceptual evaluation is to assess the degree of excellence of the perceived overall quality for multimedia.

3.2.1. Data collection

Psychoperceptual evaluation of OPQ is based on the standardized quantitative methodological recommendations [5, 39, 44]. The selection of the appropriate method needs to be based on the goal of the study and the perceptual differences between stimuli. All these methods have been applied widely in the field of audiovisual quality evaluations and have proven their reliability and validity. I chose to follow their guidelines to design and conduct the experiments and the quantitative data analysis. For the evaluation task, assessment of the overall quality of the stimuli was chosen with respect to the general design characteristics of OPQ:

1. It can be used to evaluate heterogeneous stimuli material (e.g., multimedia quality) to develop the global or holistic judgment of quality. Doing so is controversial in assessment of a certain quality attribute, such as brightness, pitch, or synchrony [69].
2. It assumes that both stimuli-driven sensory processing and high-level cognitive processing including knowledge, expectations, emotions, and attitudes are integrated into the final quality perception of stimuli [9, 37, 40].
3. It is a suitable task for consumer- or user-oriented studies of product development conducted with naïve participants [40].

In addition, overall quality evaluations can be complemented with other simple evaluations. Especially for the consumer-oriented studies, the evaluation of an acceptable quality level as an indicator of a minimum useful quality level can be appropriate for quality judgments for novel multimedia services. The UC-QoE evaluation framework therefore provides Acceptance Threshold as additional research method [57].

The test procedure during the data collection includes training and anchoring and the evaluation task. In training and anchoring, participants familiarize themselves with the presented qualities and contents used in the experiment as well as with the data elicitation method in the evaluation task. Often a subset of the actual test set is used, representing the full range of quality in the study. In the following evaluation task, the full set of test stimuli is presented according to the selected research method. The stimuli can be evaluated several times. Presentation order should be changed between repetitions as well as for different test participants to avoid order bias effects.

3.2.2. Method of analysis and results

The quantitative data can be analyzed using the Analysis of Variance (ANOVA) or its comparable nonparametric methods if the presumptions of ANOVA are not fulfilled, especially normal distribution [4]. Coolican [4] suggests a combination of the Friedman test and Wilcoxon test as a nonparametric alternative to ANOVA. Fulfilling the first goal of OPQ, the outdome of the psychoperceptual evaluation is a preference ranking of the excellence of all test stimuli. These results can be translated into preferences of treatments or test parameters under evaluation.

The goal of the quantitative data analysis is to check if the independent variables had an effect on the dependent variable, i.e. the rating of acceptance or satisfaction of the test participants. We assume related data. So firstly, one needs to check if there is a significant difference in the means among the different parameters of the independent variable. [1] The most commonly applied method is a one-way repeated measures ANOVA (Analysis Of Variances) . ANOVA checks if the variability between the different parameters of the independent variable, which is assumed to be an effect due to the independent variable, is larger than the variability within on level, which is assumed to be given by chance [4]. However, ANOVA makes several assumptions that need to be fulfilled. First, ANOVA requires normal distributed data per parameter. Commonly, a Kolmogorov-Smirnov Test or a Shapiro-Wilk used to check for normal distribution [4]. Second, the data must be interval data. Third, but most of the time neglected, the variability needs to be similar among all parameters which can be checked with Levene's test for homogeneity[4]. If the data to be tested violates one of the requirements, especially normal distribution or data type, then one must apply non-parametric methods of analysis. If your data violates the normal distribution assumption, then the non-parametric alternative to the one-way ANOVA is the Friedman test [4]. For binary data, e.g. the ratings of acceptance for determining the acceptance threshold, Cochran's Q can test for significant effects between parameters [130].

If the first test finds any significant effect between parameters, then a second step of analysis are post-hoc tests to check for significant pairwise differences between parameters. Examples of parametric post-hoc tests are Scheffe's test or Tukey's test. The non-parametric alternative is the Wilcoxon Signed-rank test [4]. McNemar test applies for binary data [130].

As for unrelated data so parametric and non-parametric methods of analysis for unrelated data exist. When analyzing data of a between-subject test design, a one-way ANOVA tests if significant differences exist between three or more levels of between-subject parameters under the same constraints as for related data. The non-related data equivalent is a Kruskal-Wallis test. Common post-hoc tests for related data are the Bonferroni t test, the Mann-Whitney U test for unrelated data and a Chi-square test for binary data pairs. [4]

3.3. Sensory evaluation

Research goal within Open Profiling of Quality: The goal of the sensory profiling is to understand the characteristics of quality perception by eliciting individual quality attributes and modeling them in perceptual spaces.

3.3.1. Data collection

OPQ partly follows the method of Free-Choice Profiling (FCP), originally introduced by Williams and Langron in 1984 [74]. It adapts FCP because it allows naïve participants to use their own

[1] This is the H1 hypothesis. The H0 hypothesis for this ANOVA or Friedman test is that there is no significant difference in the means among the different parameters of the independent variable [4].

vocabulary, differing sensitivities, and idiosyncrasies to describe the characteristics of products in a multistep evaluation procedure [40, 74]. FCP is free of time-consuming panel training by reducing training to a short task for each individual participant before sensory evaluation. However, FCP produces comparable results to other methods of descriptive analysis [40, 75]. Furthermore, it is well established in food sciences, acting as a good reference to the multimodal quality evaluation in the other research fields [131]. The sensory profiling task is again divided into subsequent phases as proposed by related methods [117]. The phases allow introducing the method and the different tasks of the test participants in logical order to assure good quality of the sensory profiling data (Figure 3.1): introduction, attribute elicitation, attribute refinement, and evaluation.

3.3.1.1. Introduction

The introduction is focused on training test participants to explicitly describe quality with their idiosyncratic quality attributes. These quality attributes are descriptors (preferably adjectives) for the characteristics of the stimuli in terms of perceived sensory quality [40, 71]. The introduction helps participants to understand the nature of the descriptive evaluation task. The descriptive skills of test participants will limit the attribute elicitation [132]. The ability to express quality is an important requirement for the participants to produce strong quality attributes [77]. In training, I recommend starting with a small task of describing something familiar to participants, such as apples. 'Imagine a basket full of apples. What kind of attributes, properties, or factors can you use to describe similarities and differences of two randomly picked.' The researcher may help the test participant to find attributes, but he never offers his own suggestions. After the introductory task, participants start to describe the audiovisual quality during the attribute elicitation phase.

3.3.1.2. Attribute elicitation

The second phase aims at eliciting individual quality attributes that characterize the participants' quality perception of the different test stimuli. The actual extraction of attributes can be done using different elicitation methods available. In the original Free-Choice Profiling approach, assessors noted attributes without limitations [74]. However, it has been reported that it was difficult for participants to develop their vocabulary, so supporting elicitation techniques should be applied [77, 117] (see section 2.3.4). OPQ applies the original elicitation technique in accordance with Williams and Langron [74] in these studies because no additional benefit in terms of attribute quality has been found for the supporting tasks [77, 82]. Independently of the elicitation method, stimuli can be replicated several times, and people need enough time to watch them and iteratively develop their attributes, as became apparent during the development of OPQ. In general, attribute elicitation is a very important step for successful sensory profiling because only the attributes found in this phase will be taken into account in the later evaluation. Therefore, the author recommends taking the time to be accurate during the introduction and attribute elicitation.

3.3.1.3. Attribute refinement

The attribute refinement aims at separating strong attributes from all developed attributes for each assessor. In FCP, participants may develop unnecessarily many attributes in their elicitation step whereas strong attributes are needed for accurate profiling [82]. From the requirements for valid verbal descriptors in sensory evaluation [40, 46, 71, 72] (see section 2.4.1), especially two rules apply in describing a strong attribute. First, the participants must be able to define the attribute in their own words; that is, they must know very precisely which aspect of quality is covered by the attribute. This is important for the interpretation of the results to understand the individual attributes. Second, the attribute must be unique and nonredundant. Each attribute must describe one aspect of quality. Following these rules, test participants are allowed to modify their list of attributes after the elicitation phase. While I conducted the first OPQ studies without limitation on the number of attributes, it later was deemed useful to limit the maximum number of attributes. A larger set of attributes can add more noise rather than additional information to the sensory data [82]. However, this should be checked in a pilot study for specific needs. At the end of the attribute refinement, test participants write down a definition for each of the final attributes. The attributes are attached to a 10 cm long scale labeled with ’min’ and ‘max’ at its extremes (see Figure 3.2). Doing so results in an individual score card per test participant to be used for the following evaluation.

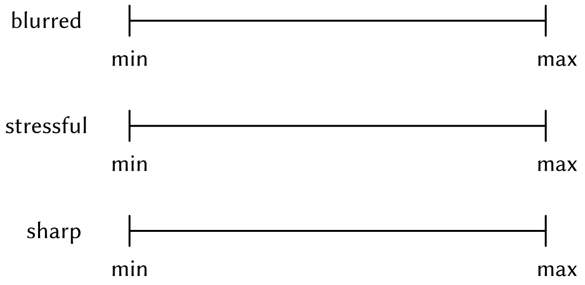


Figure 3.2. – Examples of quality attributes with the related scale on a participant’s individual score card.

3.3.1.4. Evaluation

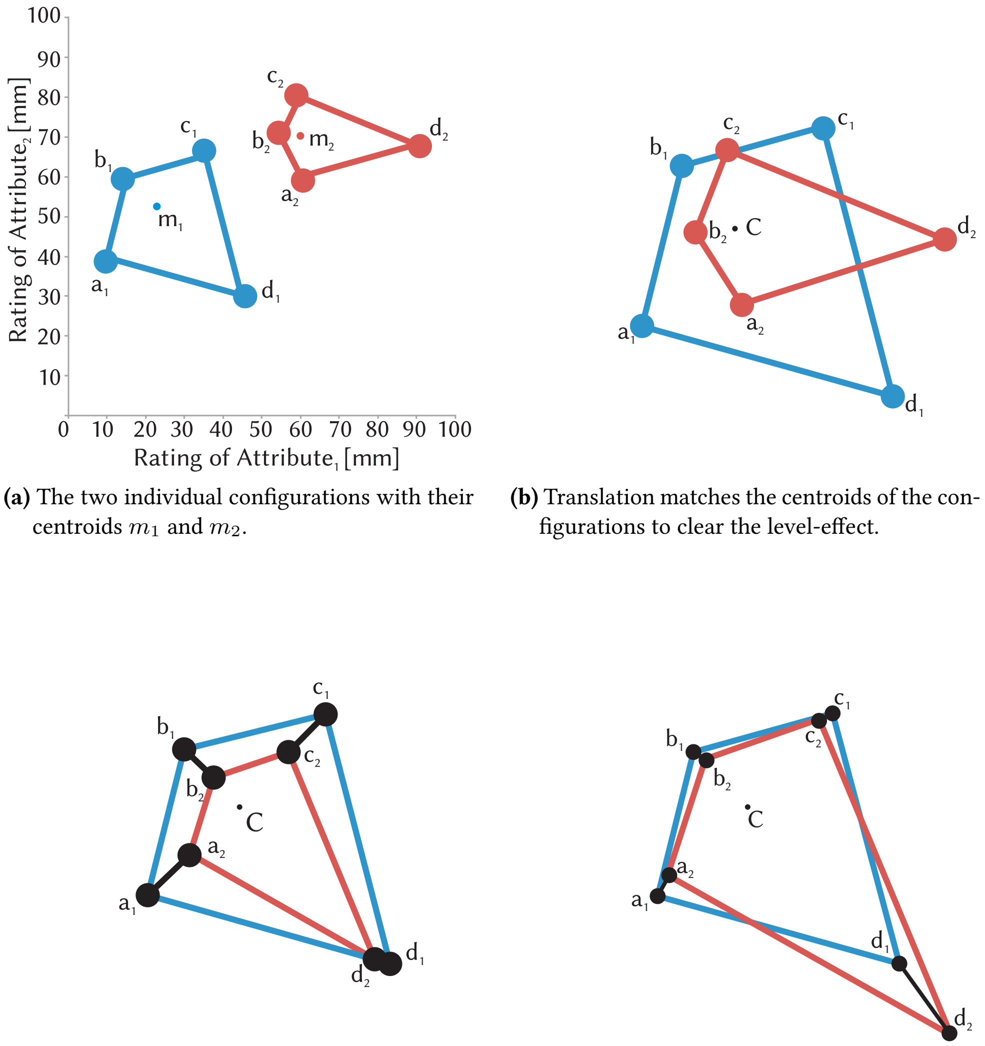
The evaluation is focused on quantifying the strength of sensation of each attribute on the score card per stimulus. The stimuli are presented one by one, and the assessment of each attribute is marked on the scale. ’Min’ means that the attribute is not perceived at all while ’max’ refers to a maximum sensation. The evaluation is the actual phase of collection of descriptive data.

3.3.2. Methods of analysis and results

The goal of the sensory data analysis is to construct perceptual spaces that are low-dimensional representations of the individual data sets. Different methods of analysis are available in the field of perceptual mapping, of which Principal Component Analysis (PCA) is the most common [40]. In general, perceptual mapping techniques transform the attribute ratings on a set of test items into a model that consists of a small number of dimensions while explaining a large amount of the variance of the input data. The problem with FCP data is that there is no agreement about attributes among test participants due to the individual characteristics of each configuration [133]. For analysis of Free Choice Profiling data, individual difference methods must be applied according to Dijksterhuis [133]. Two main approaches applied in sensory evaluation have been considered and applied in the development of OPQ. Generalized Procrustes Analysis (GPA) represents the common method that has been applied for a long time in sensory evaluation [40, 76, 133]. Multiple Factor Analysis (FMA) [99] has become popular in recent evaluations because the method can also be applied to other evaluation methods in descriptive analysis, such as sorting tasks. The results of MFA and GPA have shown to be comparable [134].

3.3.2.1. Generalized Procrustes Analysis

When Free-Choice Profiling was introduced by Williams and Langron [74], they referred to Generalized Procrustes Analysis (GPA) as the method of choice for FCP data sets. By measuring the distance from the beginning of the 10 cm long line to the mark for the rated intensity for each attribute on a score card, the sensory sensation is transformed into quantitative values. Each test participant thus produces one configuration, that is, a $M \times N$-matrix with M rows = 'number of test items' and N columns = 'number of individual attributes' (Figure 3.3a). Generalized Procrustes Analysis was introduced by Gower [76] in 1975 and has been linked closely to the development of FCP [74, 133]. To analyze FCQ data, the individual configurations must be matched according to a common basis (Figure 3.3). Gower [76] called this common basis consensus configuration. Currently, researchers refer to it as the GPA group average because this term is more representative of the common basis [135]. The group average is the mean of the individual configurations. The individual configurations thus pass through a three-step algorithm so that the residuals are minimized (Figures 3.3b-(d)) [40, 133]. Translation (Figure 3.3b) clears the level effect, which can arise among participants due to differing use of the attribute scales. Geometrically, translation refers to matching all configurations' centroids. Then, rotation takes into account that attributes do not have the same meaning due to the idiosyncratic characteristics. In rotation, the points of each configuration are brought to agreement (Figure 3.3c). This results in a residual between each pair of points. Finally, isotropic scaling minimizes the residual between configurations (Figure 3.3d). These scaled configurations can then be analyzed by conducting a PCA. The author refers to Dijksterhuis [133] for a review of the mathematics of GPA and its applications to consensus and individual vocabulary approaches. The result of the GPA is a low-dimensional perceptual model. The results are finally plotted as item maps showing the loadings of each test item on the principal components and word charts (or correlation plots) showing correlation of the individual attributes with the principle components of the low-dimensional model. In contrast to interview-based evaluation methods (see section 2.3.4), no personal data interpretation has been introduced in the analysis. At this stage, the researcher will start to identify the principal components of the perceptual space, the GPA scores of the items, and the attributes' correlation with the components to understand the rationale behind the model. This fulfills the second goal of the OPQ method.



(a) The two individual configurations with their centroids m_1 and m_2.

(b) Translation matches the centroids of the configurations to clear the level-effect.

(c) Rotation brings the points of the configurations to agreement. The resulting residuals are shows by the black lines.

(d) Isotropic scaling minimizes the residual between configurations. The scaled configurations are then analyzed by a PCA.

Figure 3.3. – Illustration of the three-step-algorithm of the Generalized Procrustes Analysis in accordance with Dijksterhuis and Gower [135]. The example is based on two individual configurations that both consist of four test items (A, B, C, and D) and two idiosyncratic attributes.

Kunert and Qannari [136] presented an alternative approach to analyzing sensory profiling data, claiming this approach to be more applicable for FCP data analysis. Kunert and Qannari use isotropic scaling of the individual configurations so that the sum of squares becomes equal for all data sets. The scaling of the data assures a subsequent PCA that each configuration has equal contribution to the PCA model irrespective of its size. This approach does not use the geometrical scaling of configurations as in the GPA but uses the variance of the original data seta for scaling.

3.3.2.2. (Hierarchical) Multiple Factor Analysis

Multiple Factor Analysis (MFA) is a method of multivariate data analysis that studies several groups of variables describing the same test stimuli and has been applied successfully in the analysis of sensory profiling data [99, 111, 137, 138]. Its goal is a superimposed representation of the different groups of variables. This goal is comparable to that of Generalized Procrustes Analysis. It is difference to GPA in terms of scaling the individual configurations. While GPA follows a geometrical approach and scales the data sets to the group average, MFA aims at respecting the individual structure of each individual configuration. MFA starts with conducting a separate PCA for each assessor's configuration. Doing so detects a common structure per configuration. From the PCA result, the first singular value is chosen as a normalization factor for the configuration. The first singular value is the matrix equivalent of the standard deviation and can be calculated as the square root of the first eigenvalue of the PCA [99, 137]. This normalization assures equal contribution of configurations comparable to the approach of Kunert and Qannari. The normalized matrices are then merged into a global data matrix on which a second PCA is computed. The results of the MFA are comparable to those of a GPA. The model can be plotted as item maps (loadings of each test item) and correlation plots. In addition, MFA allows for visualizing the impact of each individual assessor as partial plots by projecting each configuration onto the global analysis [137, 138]. Thus, MFA allows for simultaneous analysis of a structure common to all configurations as well as analysis of specific structures of only some or even individual configurations. A further advantage of the MFA in the analysis of sensory data is its flexibility. In MFA, a Principal Component Analysis is conducted for every group of variables. The data within each of these groups must be of the same kind but can differ across the different groups. This allows taking into account additional data sets. In sensory analysis, these data sets are often objective metrics of the test stimuli that are included in the MFA as supplementary variables, providing deeper insight on the model [139].

The approach of MFA has been extended to Hierarchical Multiple Factor Analysis (HMFA) by Dien and Pagès [109]. HMFA is applicable in comparisons of data sets of similar structure (Figure 3.5) or for data sets organized hierarchically (Figure 4.10). The HMFA is basically an MFA on each hierarchical level of the data set. Examples of the application of HMFA in sensory analysis are the comparison of the results of different sensory research methods, sensory profiles of untrained assessors and experts, and the combination of subjective and objective data [109, 140, 141].

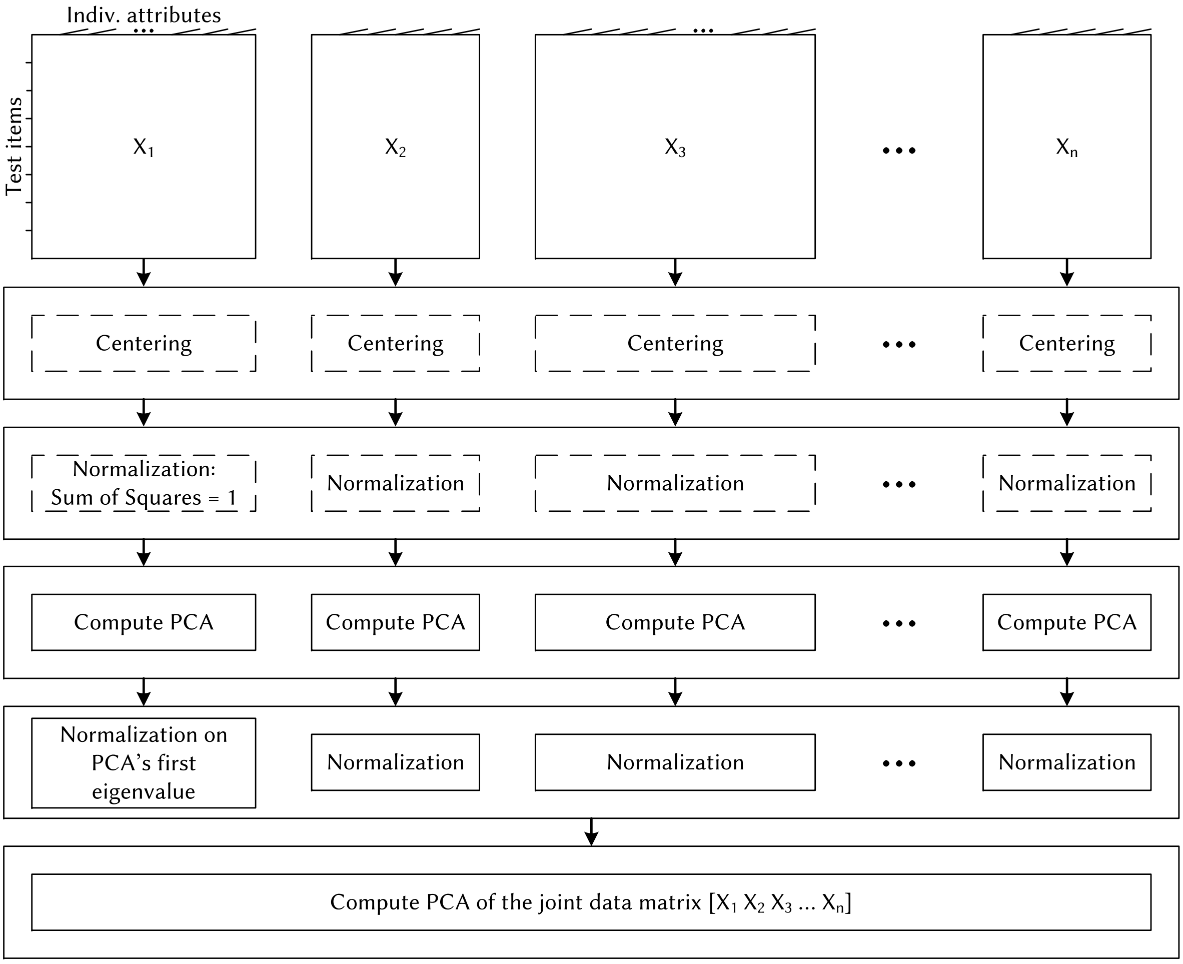


Figure 3.4. – The principle of Multiple Factor Analysis based on a set of individual configurations. The steps in dashed lines are optional and depend on the data set. Usually they must be performed for FCP data.

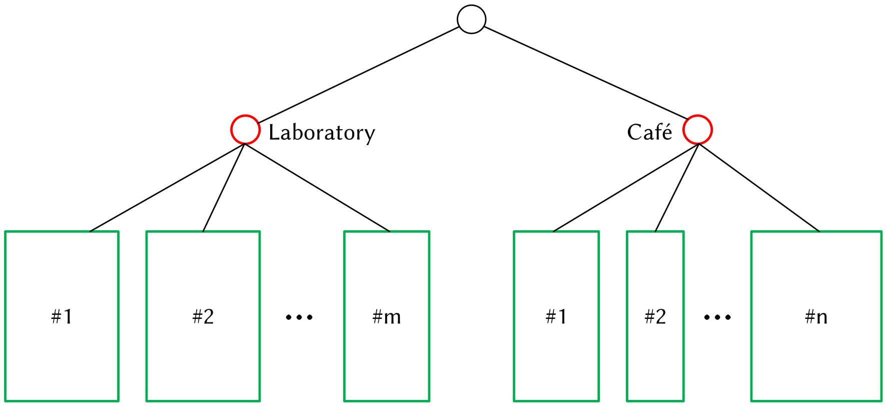


Figure 3.5. – Example for a hierarchical relationship within a data set of sensory evaluations. The hierarchical structure can be analyzed by applying Hierarchical Multiple Factor Analysis. The data set is taken from Study 4 (see section 4.5).

3.3.2.3. Issues of validity in analyzing sensory profiling data

Validity is an important aspect in the analysis of OPQ data due to the individual characteristics of the configurations. Here, the significance of the output model of GPA or MFA must be discussed because this aspect has often been neglected in previous work. In general, the explained variance, the amount of variance of the high-dimensional space that is represented by the GPA/MFA model, is taken as a value of excellence for the GPA or MFA results [40]. Although the explained variance is an important aspect of the validity of the analysis, it does not contain information about the validity of the GPA group average or the MFA model that is applied for normalizing and scaling the data sets [142–144]. Wakeling et al. [143] introduced a significance test for the GPA group average, which the author adapted for MFA within this thesis. In their approach, Wakeling et al. test whether the obtained structure of the GPA group average or the MFA model is derived from structures in the input data set or given by chance. They apply a permutation test in which the rows of each configuration are permuted randomly and total explained variance for these permuted data sets is calculated. This result represents the chance level of the explained variance. Through a large number of repetitions on different permutations, Wakeling et al. calculate a distribution of the chance level and then calculate its 95% quantile (Figure 3.6). Significance of the GPA group average or MFA result is given if the explained variance of the original data set is higher than the 95% quantile. The test presents an easy way to check the significance of GPA and MFA models obtained within this work. The result for each data set is presented in Table 3.1. For further discussions about permutation tests and their importance in assessing significance in multivariate analysis see Dijksterhuis and Heiser [142].

Study and method of analysis	Explained variance of the GPA/MFA model	95% quantile from the permutation tests (1000 permutations)	Model significant?
Study 1; GPA	81.4%	72%	yes
Study 2; GPA	81%	55,7%	yes
Study 3; MFA	30%	15.2%	yes
Study 4, laboratory; MFA	56.4%	24.7%	yes
Study 4, context; MFA	47.8%	28.7%	yes
Study 5, CP data set; MFA	53.4%	37%	yes
Study 6, Sorted Napping; MFA	62.7%	57.8%	yes

Table 3.1. – Results of the significance test with 1000 permutations according to Wakeling et al. [143] for each sensory data set within this thesis.

3.4. External Preference Mapping

Research goal within Open Profiling of Quality: The goal of the External Preference Mapping (EPM) is to combine quantitative excellence and sensory profiling data to construct a link between preferences and quality construct.

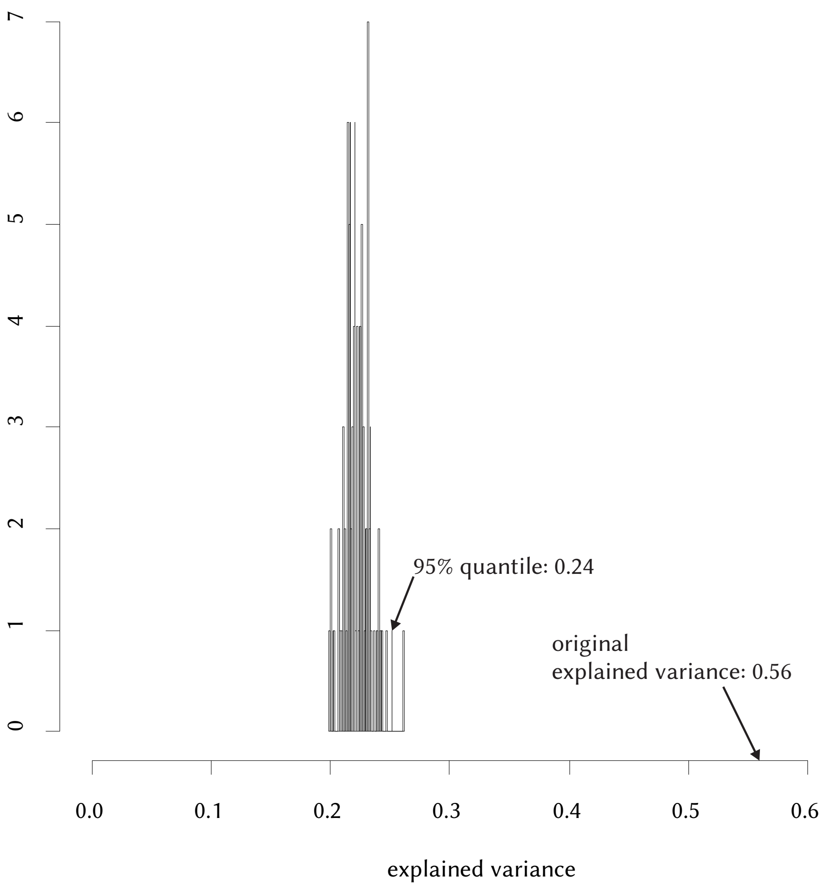


Figure 3.6. – An example of the significance test for MFA results in accordance with Wakeling et al. [143]. The data are based on the data set collected in Study 4 (see section 4.5).

In general, External Preference Mapping maps the participants' preference data into the perceptual space obtained from sensory analysis, thereby enabling the understanding of perceptual preferences by sensory explanations[2] [129]. EPM is carried out using methods of multiple polynomial regressions, for example, Partial Least Square Regression (PLS) [145, 146] or PREFMAP [129, 147]. While PREFMAP is an approach to regress quantitative data from the psychoperceptual evaluation onto the GPA or MFA model, PLS takes into account both data sets in the calculation of a PLS model.

PREFMAP is a regression analysis in which the GPA/MFA model is the independent variable and the preference data constitute the dependent variable [147]. The general regression model can be formulated as $Y = \alpha + \sum_i a_i X_i + \sum_i b_i X_i^2$, with Y being the preference data and X being the sensory dimensions (i = 1,...,number of participants). For the regression, in general, two different models are assumed. The linear regression (or vector model) refers to cases in which preferences refer to maximum sensation of sensory attributes ($b_i = 0$). McEwan [129] refers to it as the "'the more, the better' type acceptance behaviour". Preferences are mapped as vectors into the sensory

[2] In contrast to External Preference Mapping, also Internal Preference Mapping (IPM) exists . However, IPM does not allow connecting sensory and psychoperceptual data. It only contains a PCA of the preference data to be able to understand different preference patterns, e.g. from different user groups [129].

model, and the length of the vector is a measure for the degree of acceptance of the respective participant; that is, the preference increases with the length of the vector. Problems with this model occur in cases for which the preference does not correspond to a maximum (or minimum) sensation of an attribute. The ideal point model (or circular model) models preferences for the "'some amount is ideal' type acceptance behaviour" [129]. It calculates an optimum (maximum or minimum) preference point on the perceptual space, which is influenced by all sensory dimensions resulting in circular contours around the ideal plot [129, 147]. PREFMAP algorithms used within this thesis provide possibilities for automatically determining the best solution for the regression and calculating the output as either a linear or a circular model.

Partial Least Square Regression [145, 148] (also known as projection on latent structures [146]) is a multivariate regression analysis that tries to analyze a set of dependent variables from a set of independent predictors. In sensory analysis, PLS is used as a method for the External Preference Mapping with the goal of predicting the preference (or hedonic) ratings of the test participants from the sensory characteristics of the test items, obtained during the sensory evaluation of OPQ [148]. It addresses the shortcomings of PREFMAP in which the space chosen for the regression does not represent the variability of the preference data. PREFMAP performs a regression of the quantitative data on the space obtained from the analysis of the sensory data set. The advantage of applying PLS is that it looks for components (often referred as latent vectors T) that are derived from a simultaneous decomposition of both data sets. PLS therefore applies an asymmetrical approach to finding the latent structure [146]. The latent structure T of the PLS is a result of the task of predicting the preferences Y from the sensory data X. T would not be the same for a prediction of X from Y. The PLS approach allows taking into account both hedonic and sensory characteristics of the test items simultaneously [145, 146]. As a result of the PLS, a correlation plot can be calculated. This correlation plot presents the correlation of the preference ratings and the correlation of the sensory data with the latent vectors. By applying a dummy variable, even the test items can be added to the correlation plot. This correlation plot refers to the link between hedonic and sensory data that is targeted in External Preference Mapping.

With the results of the External Preference Mapping, the goals of OPQ are achieved. The theoretical description and the presented guidelines in this section contribute to the reliability and validity of the method. To support the applicability of the developed method in multimedia quality research, the next chapter will present four experiments conducted in the field of mobile, audiovisual 3D quality. The first experiment explores experienced audiovisual quality when room acoustics, audio reproduction, and visual presentation mode (2D/3D) on a midsized screen are varied. The second experiment investigates the influence of different 3D video coding methods on experienced quality on small screens. In the third experiment, different transmission parameters for an optimized DVB-H transmission of 3D mobile content are examined. Finally, the fourth study probes the applicability of Open Profiling of Quality in the context of use in contrast to laboratory evaluations. In all experiments, the level of quality can be considered as moderate, containing perceivable impairments in all presentations.

4. Application of OPQ in audiovisual quality research for 3D mobile media

This section presents the results of four OPQ studies I conducted in the context of mobile 3D media. The series of studies presents different research questions that have been answered by the application of Open Profiling of Quality. For each study, a methodological contribution and an application-related research problem are defined at the beginning. Overall, the studies contribute to the validity and reliability of results obtained from Open Profiling of Quality. Parts of this chapter have been published in Strohmeier et al., "Open Profiling of Quality: A Mixed Method Approach to Understanding Multimodal Quality Perception," Advances in Multimedia, vol. 2010, Article ID 658980, 28 pages, 2010, doi:10.1155/2010/658980 [214], Strohmeier et al., "The Extended-OPQ method for User-centered Quality of Experience evaluation: A study for mobile 3D video broadcasting over DVB-H," EURASIP Journal on Image and Video Processing, special issue on Quality of Multimedia Experience, vol. 2011, Article ID 538294, 24 pages, 2011, doi:10.1155/2011/538294 [217], and Strohmeier et al., "Open Profiling of Quality: Probing the Method in the Context of Use," Proc. of the Third International Workshop on Quality of Multimedia Experience (QoMEX 2011), Mechelen, Belgium, 2011 [218].

4.1. Quality optimization of a DVB-H broadcasting system for mobile 3DTV

Sensory evaluation is regarded as a global problem. It does not target the evaluation of just one specific research question but should be seen as a tool to study a bigger research problem from different perspectives [63]. Consequently, the development of new research methods must be embedded in a field of research in which sensory evaluation can be applied to different research questions so that the new methodology can reveal its strengths and limitations.

The methodological development of Open Profiling of Quality was embedded in the development process of a system for mobile 3DTV content delivery over an optimized DVB-H system (MOBILE3DTV) [209, 210]. Three-dimensional media are currently emerging in consumer systems and are expected to provide better experiences for users through higher immersion and the stronger feeling of presence [149]. While the general concept of 3D media systems is related to large 3D screens for home entertainment or cinemas, users have reported possible fields of application of 3D media on mobile devices [209, 210, 216]. Technically, challenges with optimizing a mobile 3DTV system exist along the whole production chain from capturing and encoding of content and error-resilient transmission to post-processed visualization on small-sized autostereoscopic screens [150][209, 219]. Each step of the production chain of mobile 3D television and video adds impairments to the content, and impairments and errors then propagate along the production chain. Boev et al. [151] provide an extensive overview of impairments of artifacts in mobile 3D television with respect to each segment of the production chain (Figure 4.1).

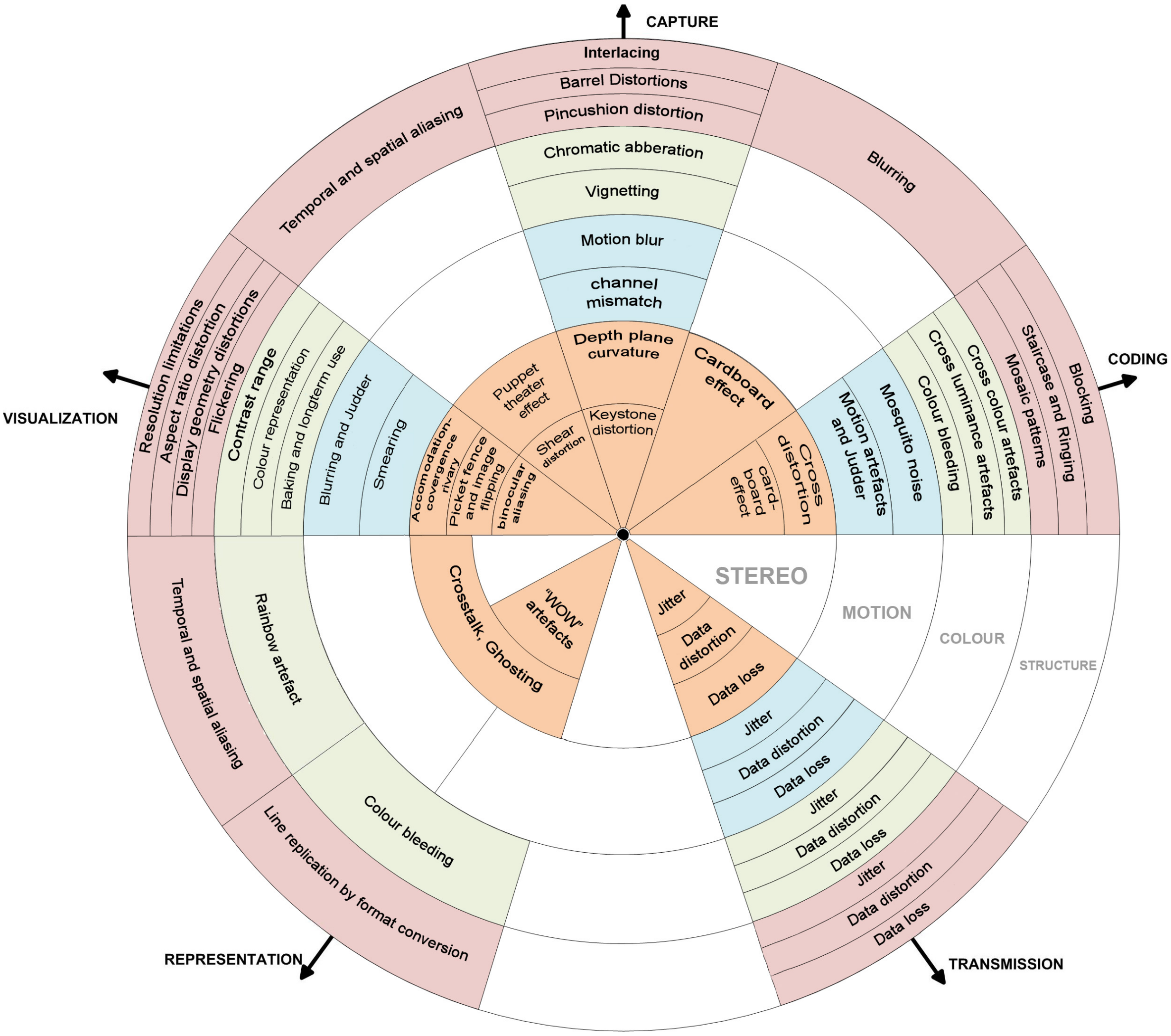


Figure 4.1. – Classification of artifacts of mobile 3D video with respect to the different stages of the production chain and spatial, temporal, and depth domains [151]

From the viewpoint of subjective quality, the quality of 3DTV is, in general, determined by three blocks: video quality, depth perception, and visual comfort. The model of 3D Visual Experience by Seuntïens [152] is an extension of the general image quality circle by Engeldrum [41] and extends Engeldrum's model with depth perception and visual comfort. In the model, visual quality and depth perception are summarized in the concept of naturalness [152, 153]. However, current studies in experienced quality of 3DTV have shown that quality has been studied separately only for different stages of the production chain. Studies in 3DTV research evaluated video quality with respect to coding errors, chroma, depth rendering, and 2D-3D comparisons or display-related artifacts like crosstalk between the channels [154–159]. However, recent studies have begun to go beyond the common quality factors to evaluate the impact of content or the impact of different display sizes with respect to mobile 3D applications [122, 123]. In addition to video quality, the understanding of visual discomfort has played an important role in 3D video evaluations [160–162]. Although studies have addressed different aspects of the production chain of mobile 3D television and video, no systematic evaluation along the production chain has yet been undertaken.

Within my work in quality optimization for mobile 3DTV, a series of studies along the production chain of MOBILE3DTV was conducted [214, 217, 218, 220]. With respect to the goal of sensory evaluation targeting a bigger research problem from different perspectives [63], different research questions were addressed. The basis of all studies was the user requirements for mobile 3D video and television, which were elicited at the beginning of the project [216]. Within the studies, different quality parameters as well as spatial and temporal characteristics of the test stimuli were varied to achieve a broad representation of impact factors within the series. Following, the author will present four studies in which Open Profiling of Quality was applied successfully to attain a broader understanding about experienced quality in comparison to common psychoperceptual evaluations.

4.2. Study 1: Experienced Quality of Audiovisual Depth

Methodological contribution: The study shows that naïve participants are able to describe their individual quality factors and to use these idiosyncratic descriptions to evaluate perceived quality. These attributes are derived from different levels of descriptions, from technical level to affective attributes, and indicate that perceived quality goes beyond perception of technological parameters. Further, the study shows the applicability of OPQ in studying individual differences in sensory evaluations.

Research problem for MOBILE3DTV: The goal of the first experiment is to explore the influence of audiovisual depth on perceived quality. In the previous work, bimodal depth experiences were studied for virtual reality systems with large screen sizes and very high-quality multichannel audio, or only one modality was explored at a time [18, 149, 163][221]. This study investigates multimodal quality perception applying OPQ when depth is varied in visual and auditory modalities. The independent variables are mono- and stereoscopic visualizations on a mid-sized screen and audio-related room acoustic simulations for small and large spaces with multichannel loudspeaker reproduction.

4.2.1. Research method

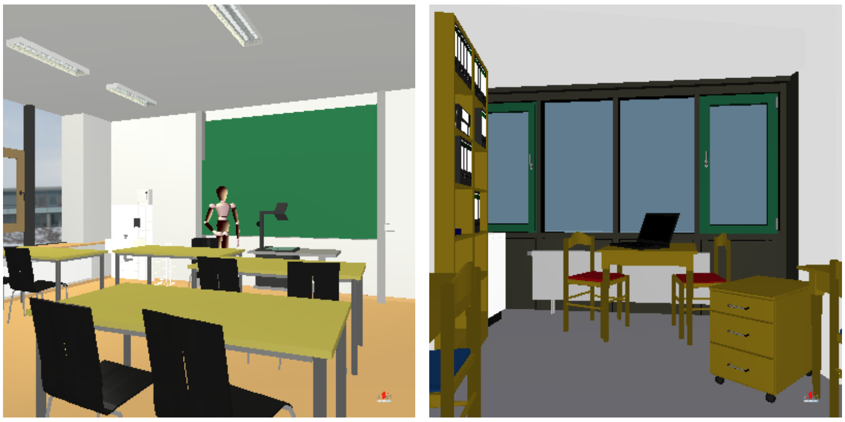
4.2.1.1. Test participants

Twenty-five naïve assessors took part in a psychoperceptual quality evaluation task (gender: 9 females, 16 males; age: 18-27 years) [45, 46, 126]. Sensory profiling was conducted with a subsample of 19 participants. All participants had normal or corrected-to-normal visual acuity and normal audio acuity.

4.2.1.2. Test stimuli

We varied depth in visual presentation mode (2D/3D) and room acoustic simulations (small/large room) in audio. Two different audiovisual contents, rendered from different sized virtual rooms, were used. Visually, a sharp display offers the possibility of physically switching between 2D and 3D presentation of the content. For the audio part, the IAVAS player offers functions to render different room acoustics [164].

In a large room, visualized as a classroom, the audio was the voice of a male speaker, and the sound source was represented by a manikin (see Figure 4.2a). In a small room, visualized as a student's living room, the audio consisted of drums and bass music, and the sound source was represented by a laptop (see Figure 4.2b). The users' movement through the room was automated. It consisted of a straight movement towards the sound source and then a turn to the left and the right. In total, eight 15-second long stimuli were used in the experiment.



(a) The virtual classroom with manikin as avatar

(b) The virtual living room with laptop as avatar.

Figure 4.2. – Snapshots of the content used in the study.

The rooms were designed using Maya software. For playback in the IAVAS I3D player [164], the scenes were exported into Binary Format for Scenes (BIFS) . The audio was included using Advanced Audio BIFS. The audio files were encoded with AAC at a bit rate of 128 kbit/s. The room acoustics were modeled using the perceptual approach provided by the player. For each room, a suitable room acoustic was modeled, taking into account the different sizes and acoustical characteristics of the rooms. To vary depth in audio perception, the room models were exchanged between the rooms.

4.2.1.3. Stimulus presentation

The tests were conducted in the Listening Lab at Ilmenau University of Technology, set according to ITU Recommendation ITU-T P.910 [5]. The videos were presented on a 15" Sharp AL3DU stereoscopic display based on parallax barrier technology. The parallax barrier has a secondary LCD layer that can be switched on and off so that the screen can be used for monoscopic and stereoscopic videos. The viewing distance was 55 cm. The sound was played back on a four-speaker surround setup at 30° and 110° and a distance of 1 m from the assessor [221]. The stimuli were repeated twice in random order for psychoperceptual evaluation.

4.2.1.4. Test procedure

Psychoperceptual Evaluation

Prior to the actual evaluation, training and anchoring were conducted. Participants trained for viewing the scenes (i.e., finding a sweet spot) and conducting the evaluation task were shown all contents and the range of constructed quality, including four stimuli. Absolute Category Rating was applied for the psychoperceptual evaluation for the overall quality, rated with an unlabelled 11-point scale [5]. In addition, the acceptance of overall quality was rated on a binary (yes/no) scale [57]. All stimuli were presented twice in a random order. The simulator sickness questionnaire (SSQ) was filled out prior to and after the psychoperceptual evaluation [165][222].

Sensory Profiling

The Sensory Profiling task was based on Free-Choice Profiling [74] methodology. The procedure had four parts carried out in two sessions over three days. 1) An introduction to the task used the imaginary apple description task. 2) During attribute elicitation, all stimuli were presented three times, one by one. The participants were asked to write down their individual attributes on a blank sheet of paper. Participants were not limited in the number of attributes, nor were they given any limitations in describing sensations. 3) During attribute refinement, participants were given a task to rethink (add, remove, change) their attributes to define their final list of words. The list was transformed into the assessor's individual score card. Finally, four randomly chosen stimuli were presented once, and the assessor practiced the evaluation using a score card. In contrast to the following evaluation task, all ratings were done on one score card. Thus, the test participants were able to compare the different intensities of their attributes. 4) During the evaluation task, the stimulus was presented three times in a row, and the participants rated it on a score card. If necessary, they could ask for a fourth repetition.

4.2.1.5. Methods of Analysis

Psychoperceptual Evaluation

Nonparametric methods of analysis were used (Kolmogorov-Smirnov: P<.05). Friedman's test is applicable to the measurement of differences between several ordinal dependent variables and Wilcoxon's test to their pairwise comparisons [4].

Sensory Profiling

The sensory data were analyzed using Microsoft Excel and the GPA routine of XLSTAT 2.9.0. The data were also analyzed using Kunert and Qannari's method [136]. Because the GPA produced stronger results in terms of explained variance of the model, the GPA model will be used for further analysis.

4.2.2. Results

4.2.2.1. Psychoperceptual Evaluation

Acceptance of Overall Quality

All presented stimuli provided a highly acceptable quality level, reaching an acceptance level of 83% at the minimum. The test parameters did not have an impact on the acceptance of overall quality (Cochran's Q = 0.79, $p > .05$, ns). All items were rated equally (McNemar: all comparisons $p > .05$).

Overall Quality Satisfaction

Visual presentation modes and room acoustic simulations did not have significant influence on the overall quality satisfaction (Friedman, $F_r = 3.341$, $df = 7$, $p > .05$, ns). All stimuli were equally rated (all pairwise comparisons $p > .05$, ns).

4.2.2.2. Sensory Evaluation

The first three components of the GPA model contribute to 81.37% of the explained variance. Considering the elbow criteria and the Heymann and Lawless' rule of interpretability [101], these first three components of the PCA were used for further data interpretation. To understand the perceptual space, the attributes and test stimuli are plotted in the model, resulting in a three-dimensional space. For better interpretation, component 2 and component 3 are always plotted against component 1 to render two-dimensional slices of the perceptual space shown in Figures 4.3 and 4.4.

The item names are substituted for by the corresponding variables. Comparing variables and separation of items in the perceptual space allows for determining the components. Figure 4.3 shows that Dimension number 1 (PC1) relates to content (classroom or student's room). Dimension number 2 (PC2) separates the test items according to the visual Presentation Mode (2D or 3D presentation). PC2 is identified as "video quality". Dimension number 3 (PC3) divides the items according to the room acoustics (simulated small room and simulated large room). It relates to the "audio quality" of the stimuli. Although this interpretation was based on the test items or their related test parameters, we will refer to the quality aspects of content, video representation, and room acoustics in further interpretation. This first finding confirms that test participants derived their individual quality factors from the chosen test parameters.

Correlation of Attributes and the Perceptual Space

The attributes can be classified into two different groups. First, technical descriptions directly describe the characteristics of the test variables (like reverberation or grainy). The second group of attributes is characterized by experiences, subjective impressions, and feelings about the test items (e.g., monotone, lively, or obtrusive). This group is called impression descriptions. Word charts

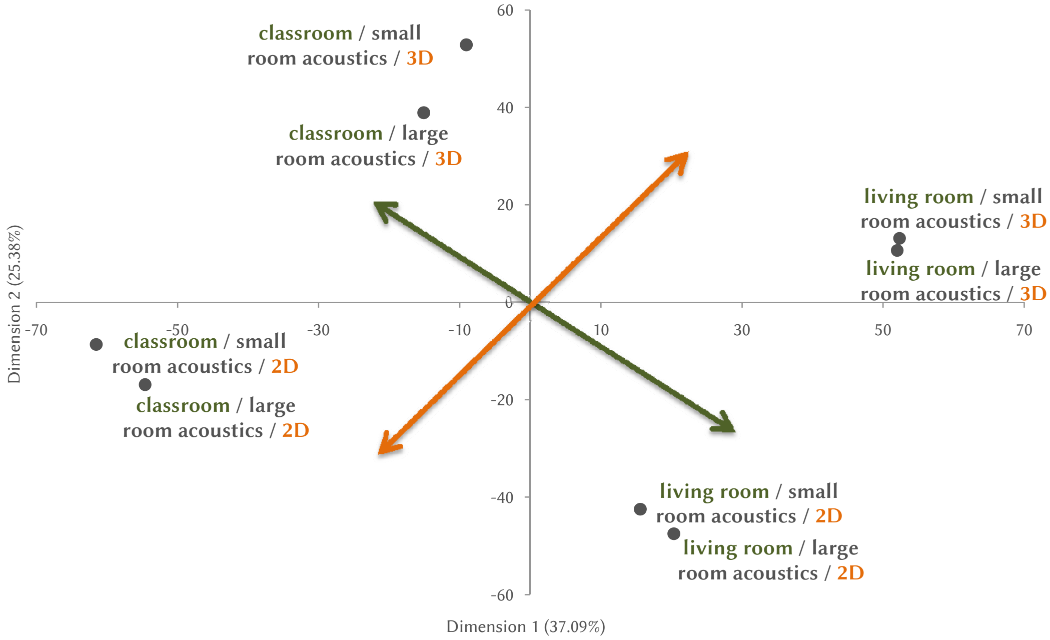


Figure 4.3. – PC1-PC2 slice of the model showing test items plotted into the GPA model. The brown and green arrows indicate the dimensions of content (PC1) and video representation (PC2).

represent the correlation of the individual attributes with the perceptual space (see Figures 4.5 and 4.6). The closer an attribute is placed to one of the dimensions, the more it correlates with this dimension. Attributes placed between two dimensions correlate with both dimensions equally.

The Dimension "Content" (PC1, 37.09% of Explained Variance) I was able to identify the two polarities of this dimension as classroom on the one side and student's room on the other side. But only a few attributes such as 'unpleasant voice', 'comiclike' or 'messy' describe the content or the layout of the room directly. PC1 is more a description of the individual impression of the content. Descriptions like 'lifeless', 'emotional' and 'likeable', or 'monotone' and 'sterile' highly correlate with one of the two polarities, respectively. The high amount of impression descriptions shows that quality perception is formed on an abstract level by the test participants. The assessors were able to find individual attributes that describe quality on a general level among the test items.

The Dimension "Visual Presentation Mode" (PC2, 25.38% of Explained Variance) The polarities agree with varied visual presentation modes (mono (2D), autostereoscopic (3D)). The 2D polarity shows descriptions of 'sharpness' or 'sharp edges', 'high contrasts', 'clear', 'light', or 'colorful'. In contrast, 3D presentation mode is described with a negative description of the visual artifacts, such as 'skewed outline', 'unclear', or 'interlaced lines'. It seems that the artifacts and reduced brightness of 3D results from limitations of the display technique (parallax barrier and viewing angle of the display). However, the results also show the participants' ability to experience visual depth. It is described as 'integration', 'three-dimensional', 'spacious', or 'tangible'.

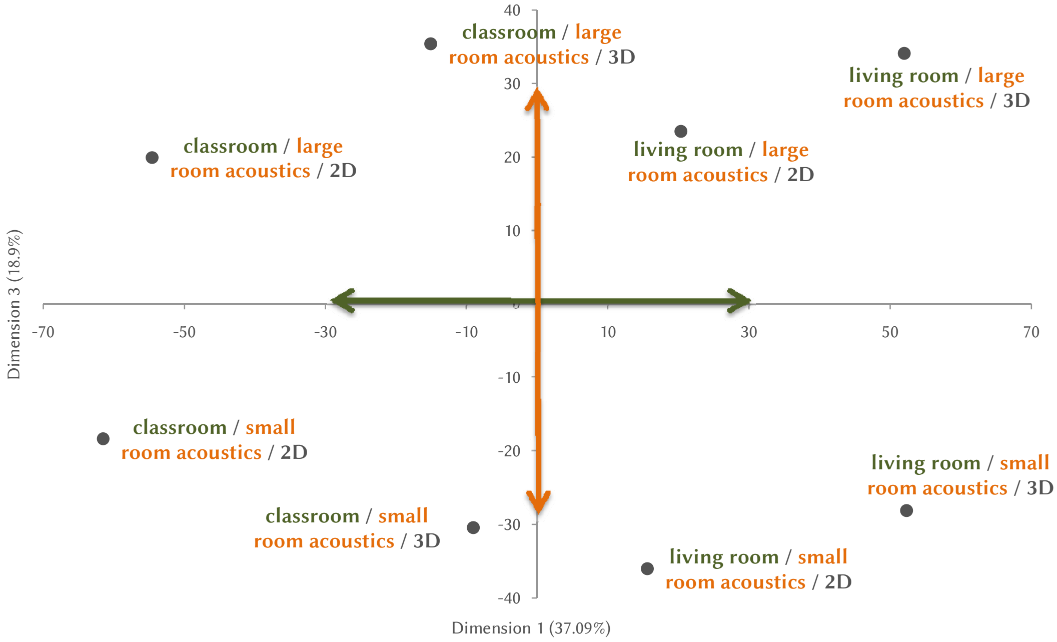


Figure 4.4. – PC1-PC3 slice of the model showing test items plotted into the GPA model. The brown and green arrows indicate the dimensions of content (PC1) and room acoustics (PC3).

The Dimension "Room Acoustic Model" (PC3, 18.9% of Explained Variance) PC3 also corresponds directly to the varying room acoustic models used in the test. The dimension can be considered in terms of the extreme values in the large room and the small room. While the small room acoustics are described as being poor, many quality factors can be found for the large room acoustics. In this dimension, technical descriptions dominate. The large room correlates with a high amount of reverberation, 'full spacious sound', and 'filling the room'. On the level of impression descriptions, PC3 is characterized as 'imaginable', 'insistent', or 'shrill'.

Interpretation of perceptual spaces between assessors Attributes that correlate with more than one dimension can be interesting, especially attributes that correlate with PC2 and PC3 as they describe audiovisual effects. Interdimensional attributes between audio and video dimension are rare (see Figure 4.6). Especially, depth-related attributes, which one could expect to correlate with both dimensions, correlate either with the video (spacious (P3)) or with the audio dimension (spacious (P14)). These results show that depth was perceived or rated independently either in auditory or visual perception. Further investigation of this finding showed that assessors favored either audio or video for deriving their quality attributes. By plotting the assessors' attributes into the perceptual space independently, I was able to find sensory preferences among participants. The usual goal of the GPA is to fit all assessors' configurations to a common consensus [76] with the aim of modeling a common quality rationale for all assessors. Individual differences among the test participants are not taken into account. Figure 4.7a shows the correlation of attributes

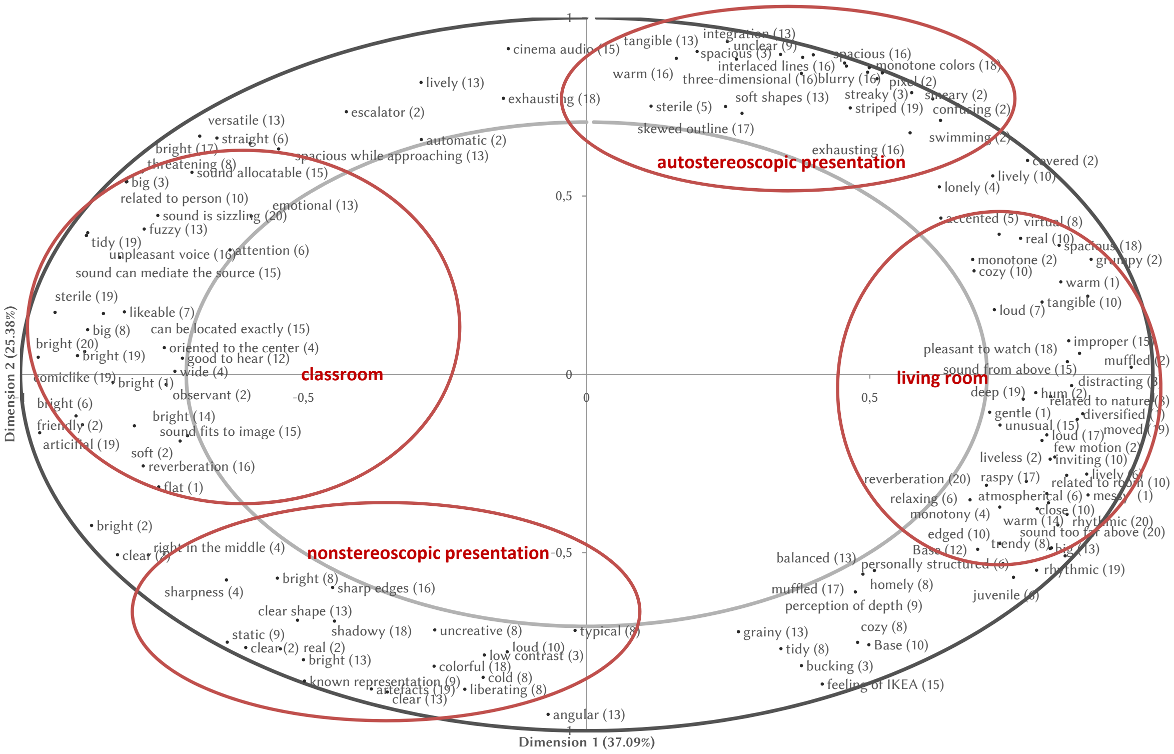


Figure 4.5. – GPA correlation loadings with attributes in the space of PC1 and PC2

with PC1 and PC2. It is remarkable that attributes from Participant #1 (shown as stars) correlate only with PC1 or the content dimension. In contrast, quality factors from Participant #13 (squares) show a high correlation with the dimension of video quality (PC2). Attributes from Participant #14 (dots) rarely are mapped along PC1 and PC2. Instead, they correlate highly with the audio quality component, as can be seen in Figure 4.7b. What can be seen from these plots is that participants use different parameters of the test items to derive their individual quality parameters from. An analogous analysis for other assessors shows that only few of them use two or even three parameters for deriving quality attributes. For example, items from participant #25 correlate with all three PCs of the GPA model (see Figures 4.5 and 4.6).

4.2.3. Discussion and Conclusion

The results of psychoperceptual quality evaluation did not show the influence of audiovisual depth on perceived quality. However, the results of sensory profiling offered further understanding for this finding. First, the nonsignificant difference was not caused by the nondetectable differences between stimuli, as the participants qualitatively differentiated them. Second, the perceived depth was highlighted by both modalities contributing to the overall audiovisual perception. Third, when visual 3D presentation mode was used, it was described as spacious and three-dimensional, but more importantly it was associated with several negative terms of inferiority. These findings confirm that

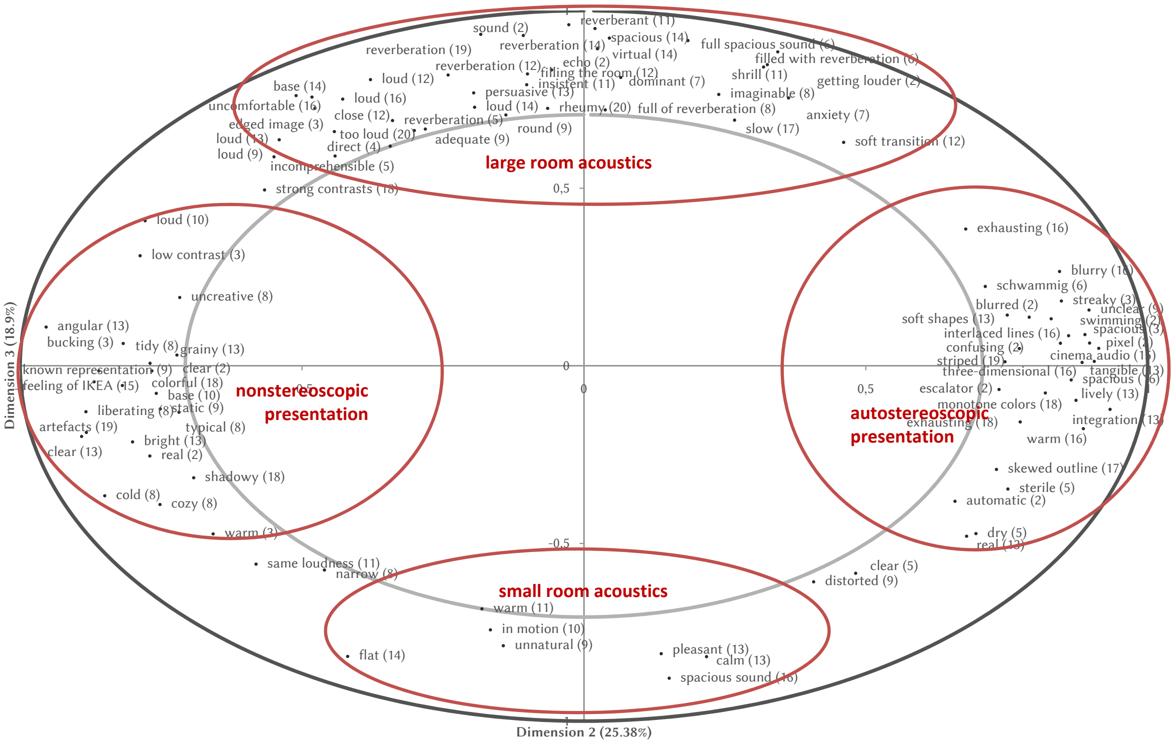


Figure 4.6. – GPA correlation loadings with attributes in the space of PC2 and PC3

the added value of the visual depth perception is acknowledged only if the level of visible artifacts is low enough [152, 166].

The results also showed individual preferences for the quality of one modality. It is known that there are modality-dependent individual differences in human information processing styles. For example, categorizing visual and verbal information processing styles is common [114]. The findings in this study indicate that these different processing styles can also contribute to final multimodal quality judgments. Two suggestions for further study arise from these findings. First, the influence of different processing styles on multimodal quality perception with different quality levels and heterogeneous stimulus material needs to be addressed in detail to confirm the phenomenon. Second, for practitioners of audiovisual quality, a well-validated tool is needed to identify groups of different information processing styles and characterize these groups.

4.3. Study 2: Experienced Quality of Video Coding Methods for Mobile 3D Television

Methodological contribution: The study confirms the methodological goal of using Open Profiling of Quality to link quantitative and descriptive data by External Preference Mapping. External Preference Mapping can provide sensory explanations for existing preference structures of psycho-

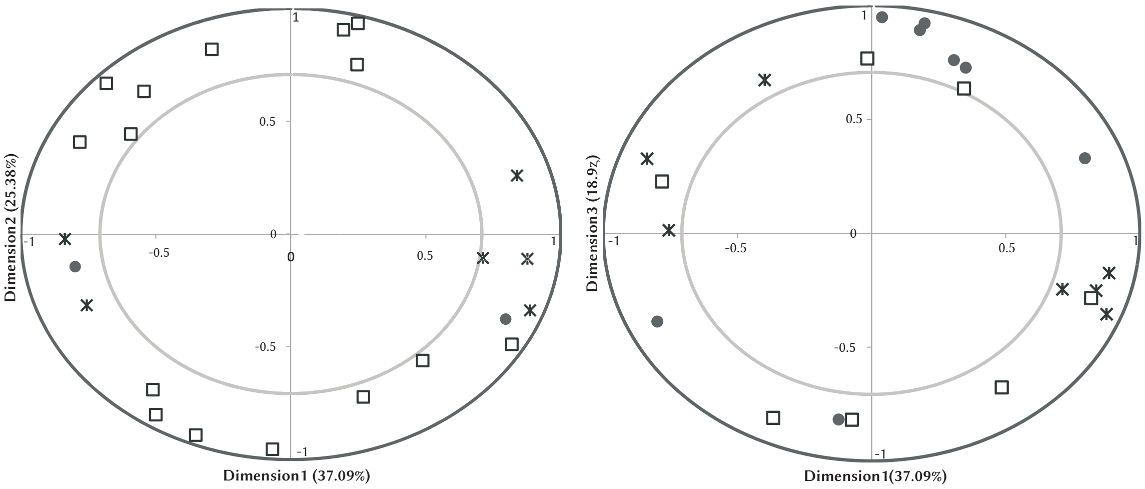


Figure 4.7. – Correlation of attributes from Participant #1 (stars), Participant #13 (squares), and Participant #14 (dots) with the main components of the GPA model; a) Dimensions 1 and 2, b) Dimensions 1 and 3.

perceptual results. Further, the study shows that naïve test participants are able to evaluate a large set of test stimuli with a wide quality range in an OPQ study.

Research problem for MOBILE3DTV: The second study targeted the selection of an optimum stereo video coding method for mobile 3D television and video applications. Different approaches of coding algorithms have been optimized for mobile 3D video [167]. No previous work had evaluated these approaches in a large-scale study. In addition, previous work on stereo coding was mainly conducted on still images [152, 154, 155]. These studies showed that the added value of stereoscopic stimuli given for the uncompressed case is not valid for MPEG2 or JPEG compressed material [152, 154–156, 166]. In these cases, the depth perception did not increase the perceived overall quality of the stimuli. This result indicates that visual quality dominates the overall quality perception. The study targeted an evaluation of this finding for mobile 3D video.

4.3.1. Research method

4.3.1.1. Test participants

Forty-seven naïve assessors (gender: 23 females, 24 males; age: 16-37, mean: 24) took part in the psychoperceptual evaluation task. Fifteen of them were randomly selected from this sample for the sensory profiling task. All assessors passed a screening for visual acuity, color, and 3D vision and were also among potential users of mobile 3D television [216]. Parents' consent was required for the participation of under-aged assessors.

4.3.1.2. Test stimuli

Variables and Their Production

Four coding methods and two quality levels were varied in this study. The coding methods were especially adapted for mobile stereo video [167]. As Video + Video approaches, H.264/AVC Simulcast [168], a straight-forward coding solution; H.264/AVC MVC as an advanced approach [169]; and Mixed Resolution Stereo Coding (MRSC) [170] as a recently introduced coding approach were chosen. In addition, Video + Depth using MPEG-C part 3 [171] as an alternative approach to the Video + Video coding methods was selected. As a coding profile, the Baseline profile, that is, IPPP structure and CAVLC (Context Adaptive Variable Length Coding), was used. The GOP size was set to 1. A low and a high quality level were defined for each test sequence. To guarantee comparable low and high quality for all sequences, individual bit rate points had to be determined for each sequence. For the definition of low quality for all sequences, the quantization parameters (QPs) for simulcast coding were set to 30. The resulting bit rates for each sequence are given in Table 4.1. These bit rates were used as target rates for the other three approaches.

Quality level	Bullinger	Butterfly	Car	Horse	Mountain	Soccer2
Low	74	143	130	160	104	159
High	160	318	378	450	367	452

Table 4.1. – Target bit rates of the final test sequences for Study 2.

Two different codecs were used for video encoding. H.264/AVC Reference Software JM 14.2 was used for the Simulcast, Mixed Resolution, and Video + Depth. MVC was performed using H.264/MVC reference Software JMVC 5.0.5. The test stimulus production for Simulcast and MVC-encoded sequences was straightforward according to the target bit rates in Table 4.1. To achieve these target bit rates, the quantization parameters for the left and the right were both changed. Thus, the left and the right views were of the same quality. The depth for the Video + Depth approach has been estimated from the left and the right view using a Hybrid Recursive Matching algorithm [172]. The view synthesis was performed using Merkle et al.'s algorithm [173]. For the generation of Mixed Resolution sequences, the right view was decimated by a factor of two in both the horizontal and vertical direction. For up- and down-sampling, tools provided with the JSVM reference software for Scalable Video Coding were used. The applied optimization approach is described in [174]. The frame rate of all sequences was set to 15 fps.

Contents

Six different contents were chosen to create the test stimuli (Table 4.2). The selection criteria for the videos were spatial details, temporal resolution, amount of depth, and the user requirements for mobile 3D television and video [216]. None of the contents contained scene cuts.

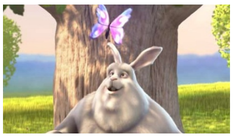
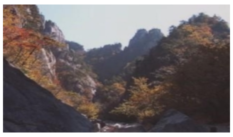
Screenshot	Genre and their audiovisual characteristics
	Videoconference – Bullinger V_{SD}: med, V_{TD}: low, V_{D}: med, V_{DD}: low, Length: 7.7 s
	**Animation – Butterfly** V_{SD}: high, V_{TD}: med, V_{D}: med, V_{DD}: low, Length: 7.7 s
	Action/Movie – Car V_{SD}: high, V_{TD}: high, V_{D}: med, V_{DD}: med, Length: 7.7 s
	Nature/Documentary – Horse V_{SD}: high, V_{TD}: low, V_{D}: high, V_{DD}: low, Length: 7.7 s
	**Nature/Documentary – Mountain** V_{SD}: high, V_{TD}: low, V_{D}: high, V_{DD}: high, Length: 7.7 s
	Sports – Soccer2 V_{SD}: med, V_{TD}: high, V_{D}: high, V_{DD}: high, Length: 7.7 s

Table 4.2. – Snapshots of the six contents under assessment (V_{SD}=visual spatial details, V_{TD}=temporal motion, V_{D}=amount of depth, V_{DD}=depth dynamism)

4.3.1.3. Stimulus presentation

The controlled laboratory conditions were similar to Study 1. A NEC autostereoscopic 3.5" display with a resolution of 428px x 240px was used to present the videos. This prototype of a mobile 3D display provides equal resolution for monoscopic and autostereoscopic presentation. It is based on lenticular sheet technology [175]. The viewing distance was set to 40 cm. The display was connected to a Dell XPS 1330 laptop via DVI. The stimuli were presented in a counterbalanced order in both evaluation tasks. All items were repeated once in the psychoperceptual evaluation task. In the sensory evaluation task, stimuli were repeated only when the participant wanted to see the video again.

4.3.1.4. Test procedure

Psychoperceptual Evaluation

The psychoperceptual evaluation followed the same method as described in Study 1. Test participants evaluated overall quality acceptance and satisfaction with overall quality in this study. The session took about 90 minutes.

Sensory Profiling

Sensory profiling was conducted in a second session, lasting 75 minutes. A Free-Choice Profiling approach was applied with the following subtasks:

1. Introduction – An introduction to the task was carried out using the imaginary apple description task.
2. Attribute elicitation – the test participants watched a subset of 24 randomly chosen test items. While watching, they wrote down their idiosyncratic quality attributes. No limit for the number of attributes was given in this step. During the last clips, the test participants were encouraged to review their attributes by checking whether all quality aspects were covered by their noted attributes.
3. Attribute refinement – at the beginning of the attribute refinement, the assessors were asked to select a maximum of 15 attributes for their score card. After the selection, 12 test items were presented, and the test participants evaluated these on their score cards. Furthermore, the possibility of revising the score card (adding, removing, redefining attributes) was given. The score card was then finalized, and each assessor defined his quality attributes.
4. Evaluation task – in the final evaluation task, all 48 items were rated independently. Each item was shown three times in a row to allow enough time for assessors to apply all attributes. The rating time was not limited.

4.3.1.5. Methods of Analysis

Psychoperceptual Evaluation, Sensory Profiling, and External Preference Mapping

Psychoperceptual evaluation and Sensory Profiling were analyzed exactly as they were in Study 1. External Preference Mapping was applied to map the users' preferences into the perceptual space. Two models can be used to describe the participants' preferences: the vector model and the ideal point model [147]. Within the PREFMAP method in XLSTAT, the most suitable model is chosen automatically.

4.3.2. Results

4.3.2.1. Psychoperceptual Evaluation

Acceptance of Overall Quality
All coding methods provide highly acceptable quality at the high quality level, 80% at the minimum. At the low quality level, MVC and Video + Depth still reached a 60% acceptance level while the acceptance for MRSC and Simulcast was below 40%. The distributions of acceptable and unacceptable ratings on the satisfaction scale differ significantly ($\chi^2(10)$ = 2368, p < .001). The scores for nonaccepted overall quality are found to be between 1.4 and 4.2 (Mean: 2.8, SD: 1.4). Accepted quality was expressed with ratings between 4.5 and 8.5 (Mean: 6.5, SD: 2.0). Thus, the Acceptance Threshold can be determined as being between 4.2 and 4.5.

Overall Quality Satisfaction
At the high quality level, coding methods had an influence on quality satisfaction (F_r = 241.83, df = 3, p < .001; Figure 13). MVC and Video + Depth provided the highest overall quality satisfaction scores when averaging over the content (MVC versus V + D: Z = -.828; p > .05; ns), outperforming MRSC and Simulcast (all pairwise comparisons: P<.001). The results were confirmed for low quality level (F_r = 648.97, df = 3, p < .001), where MVC and Video + Depth outperform MRSC and Simulcast (all pairwise comparisons p < .05). Content-by-content analysis showed that Video + Depth outperformed all other methods at the high and low quality levels (all comparisons p < .01). For Butterfly content, MVC had the best satisfaction scores for both quality levels (all comparisons: p < .01). Coding methods did not have an influence on Bullinger content at the high quality level (F_r = 2.942; df = 3; p > .05; ns).

4.3.2.2. Sensory Evaluation

Fifteen assessors in the sensory profiling session developed 102 individual quality attributes.

Identification of Dimensions and Attributes
Considering Lawless and Heymann's rule of interpretability [101], two dimensions were identified as important for the GPA model. The first two components of the GPA model had 88.36% explained variance, of which PC1 covered the majority (83.32%). Figure 4.8 shows the item plot and Figure 4.9 the correlation plot of the GPA model. The analysis emphasizes attributes explaining more than 50% of the variance. As can be identified from the plots, PC1 is largely determined by video quality. PC2 discriminates the items (Figure 4.8) into items with high amount of motion (soccer) and low amount of motion (Bullinger).

Dimension 1 ("video quality", 83.32% explained variance) PC1 shows a high correlation of its negative polarity with attributes like 'blurry', 'blocky', or 'grainy'. On its positive polarity, it correlates with attributes like 'sharp', 'detailed', and 'resolution'. This component describes the

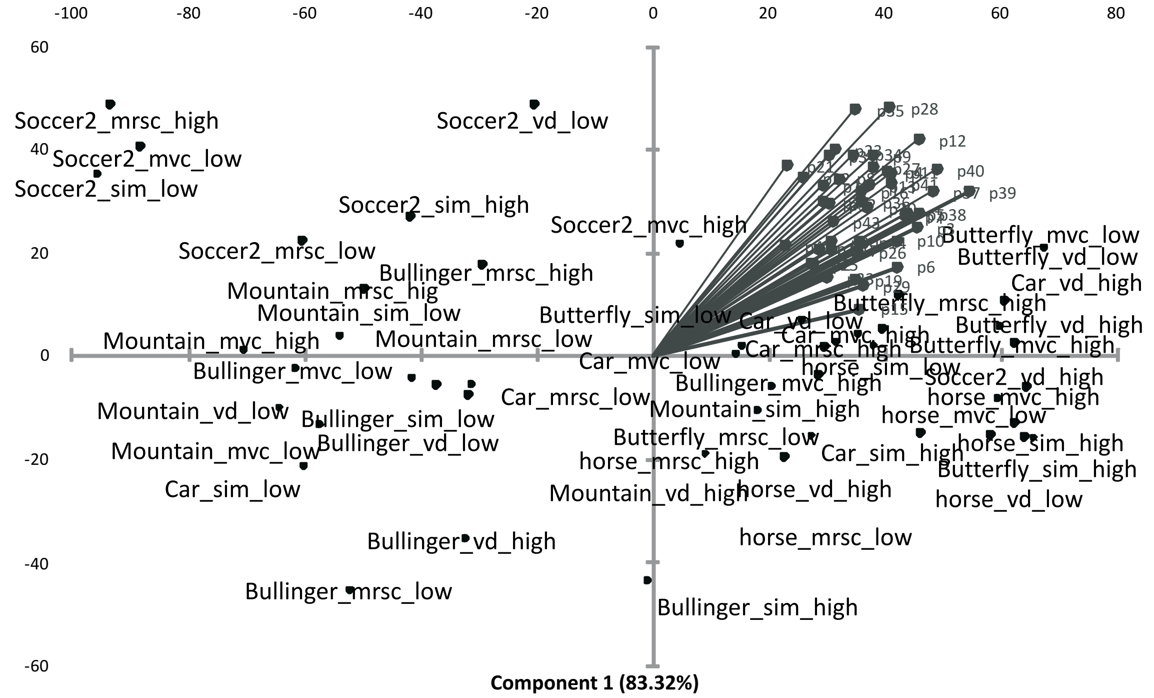


Figure 4.8. – The item plot of the GPA model showing the first two principal components and the test items within the space. Gray arrows mark users' preferences, mapped into the model using PREFMAP

video quality. It separates the model into good and bad quality. The bad quality mainly contains descriptions of artifacts.

Dimension 2 ("amount of motion", 5.03% explained variance) Along PC2, static content (Bullinger, Mountain, Horse) and content containing motion (Butterfly, Soccer2, Car) are separated (Figure 4.8). It is remarkable that the explained variance of PC2 is very small compared to the first dimension. However, it is reasonable that the amount of motion has an impact on perceived quality due to the applied coding methods. No attributes were identified to describe the perception of motion. A separate depth component was not identified in the GPA model. The correlation plot shows that 3D-related attributes like 'spacious', '3D reality', or 'background depth' correlate with the positive polarity of PC 1. The results show that depth descriptions seem to be part of good quality. If video quality is low due to coding artifacts, this quality degradation will exceed the additional value provided by the stereoscopic video presentation. Depth will not be taken into account to describe quality.

4.3.2.3. External Preference Mapping

The results show a preference for artifact-free stimuli (Figure 4.8). The content with the highest user preference is identified along PC1. The least preferred items are all Bullinger clips at the opposite side of the marks. It can also be seen that the Bullinger clips correlate with an attribute called

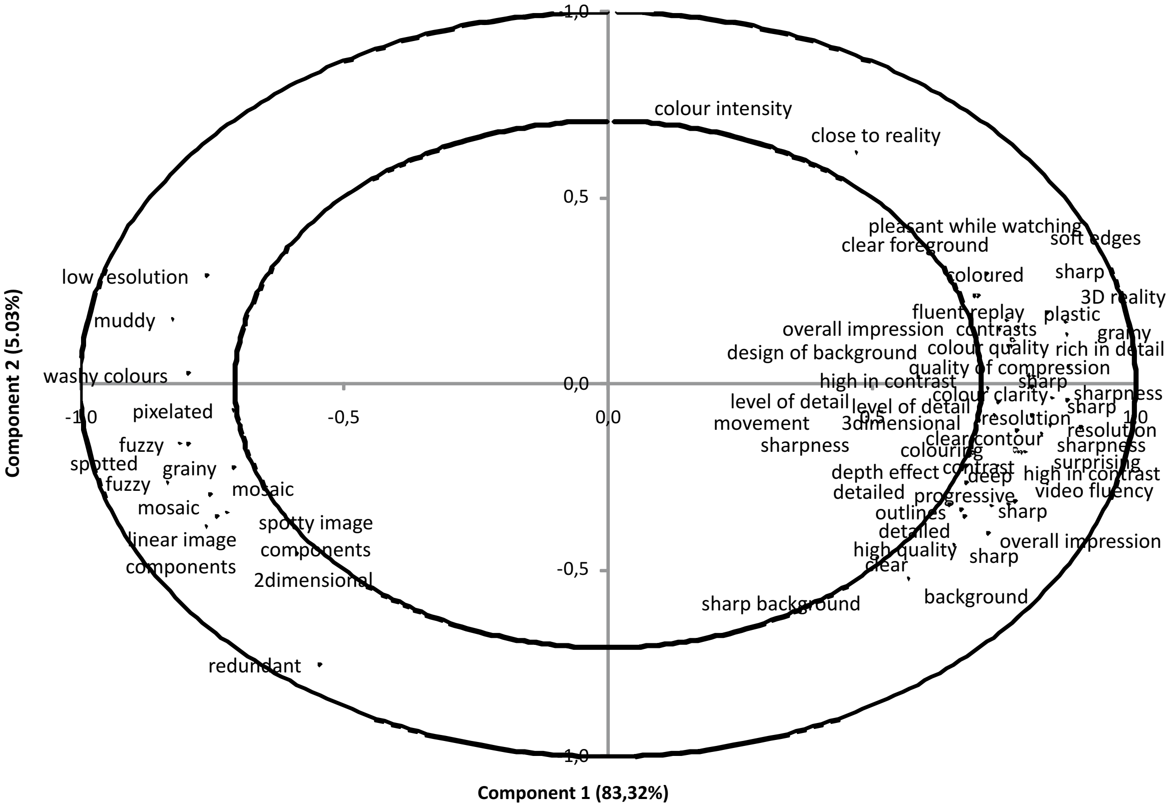


Figure 4.9. – Correlation plot of the experienced quality factors. The figure shows the first two principal components of the GPA model and the correlation of the attributes with these components. Inner and outer circles show 50% and 100% explained variance, respectively.

’redundant’. Although this attribute appears only once, it may explain the quantitative results of Bullinger clips. Quantitative analysis has shown that the differences between coding methods are rather small for Bullinger content. The ’redundancy’ of the Bullinger items may show that the participants evaluated the content on a more affective level, not according to its provided quality.

4.3.3. Discussion and Conclusion

The results of psychoperceptual evaluation showed that Multiview Coding and Video + Depth provide the highest experienced quality among the tested coding methods. They also represent contrary methods in the coding of 3D video. While MVC uses inter- and intra-view dependencies of the two video streams (left and right eye), the Video + Depth approach renders virtual videos from a given view and its depth map [167]. In addition, the provided quality level was highly acceptable compared to other related studies within the MOBILE3DTV project [60].

The results of sensory profiling showed that artifacts are still the determining quality factor for 3D. The expected added value through depth perception was rarely mentioned by the test participants. When mentioned, it was connected to the artifact-free video. These results are in line with previous studies concluding that depth perception and artifacts both determine 3D quality perception [152, 156]. In contrast to Seuntïens' model [152], the profiles in this study show a hierarchical dependency between depth perception and artifacts. When the visibility of artifacts is low, depth perception seems to contribute to the added value of 3D. With respect to stereo video coding methods, it appears that the compression of the depth map in Video + Depth approaches directly impacts depth quality. In contrast, depth is not affected in Video + Video approaches in related coding methods. Further work needs to investigate more deeply the interaction between artifacts and depth to improve coding methods for mobile stereo video.

4.4. Study 3: Experienced Quality of 3D video transmission over DVB-H

Methodological contribution: The methodological validity of OPQ is tested by introducing new methods of analysis (threat of mono-method bias). The study increases the flexibility in data analysis by introducing (Hierarchical) Multiple Factor Analysis to analyze the sensory data. Further, a comparison of PLS and PREFMAP results for External Preference Mapping is presented. The results of the study fix HMFA and PREFMAP as the methods of choice for analyzing sensory data within OPQ.

Research problem for MOBILE3DTV: The study targets an evaluation of optimum transmission settings under the constraints of the selected coding methods for mobile 3D video. Several options for error resilience and error protection are provided in DVB-H transmission that has been optimized for MOBILE3DTV. Different approaches to error protection and error resilience are evaluated at varying levels of channel loss rates.

4.4.1. Research method

4.4.1.1. Test participants

Seventy-seven participants (gender: 31 female, 46 male; age: 16-56, mean = 24 years) took part in the psychoperceptual evaluation. All participants were recruited according to the user requirements for mobile 3D television and system. They were screened for normal or corrected-to-normal visual acuity (myopia and hyperopia, Snellen index: 20/30), color vision using the Ishihara test, and stereo vision using the Randot Stereo Test ($\geq$ 60arcsec). The sample consisted largely of naïve participants who had not had any previous experience in quality assessments. Three participants had taken part in quality evaluations previously, one of them regularly. No participants were professionals in the field of multimedia technology. Simulator Sickness of participants was controlled during the experiment using the Simulator Sickness Questionnaire. The results of the SSQ showed no severe effect of 3D on the condition of the test participants [222]. For the sensory analysis, a subgroup of 17 test participants was selected. During the analysis, one participant was removed from the sensory panel.

4.4.1.2. Test stimuli

In this study, we varied three different coding methods using slice and noslice modes, two error protections, and two different channel loss rates with respect to the Mobile 3DTV system [176]. The Mobile 3DTV transmission system consists of taking stereo left and right views as input and displaying the 3D view on a suitable screen after broadcasting and receiving with necessary processing. The building blocks of the system can be broadly grouped into four blocks: encoding, link layer encapsulation, physical transmission, and receiver. Targeting a large set of impacting parameters on the Quality of Experience in mobile 3D video broadcasting, the different test contents were varied in coding method, protection scheme, error rate, and slice mode.

Coding methods The effect of coding methods on the visual quality in a transmission scenario is twofold. The first is that different artifacts result from encoding methods prior to transmission [223]. The second is that different perceptual qualities of the reconstructed videos after the transmission losses are caused by different error resilience/error concealment characteristics of the methods. We selected three different coding methods representing different approaches in compressing mobile 3D video in line with previous results [223, 224]: Simulcast Coding (Sim) according to H.264/AVC [71], Multiview Video Coding (MVC) [72], and Video + Depth Coding (VD) using MPEG-C Part 3. For all the coding methods, an IPPP prediction structure, a group of pictures (GOP) having 8 each, and a target video rate of 420 kbps total for the left and right views were selected.

Slice mode For all the aforementioned encoding methods, it is possible to introduce error resilience by enabling slice encoding, which generates multiple independently decodable slices corresponding to different spatial areas of a video frame. The aim of testing the slice mode parameter is to observe whether the visual quality is improved subjectively with the provided error resilience.

Error protection To combat higher error rates in mobile scenarios, the Multi-Protocol Encapsulation-Forward Error Correction (MPE-FEC) block in the DVB-H link layer provides additional error protection above the physical layer [177, 178]. It is possible to protect the left and right transmitted streams with the same protection rates (Equal Error Protection, EEP) as well as with different rates (Unequal Error Protection, UEP). The motivation for using unequal protection is that the independent left view is more important than the right or depth view. The right view requires the left view in the decoding process, and the depth view requires the left view to render the right view. However, the left view can be decoded without the right or depth view. In equal error protection (EEP), the left and right (depth) views are protected equally by assigning a 3/4 FEC rate for each burst. Unequal error protection (UEP) is attained by transferring half of the RS columns of the right (depth) view burst to the RS columns of the left view burst. In this way, EEP and UEP streams achieve the same burst duration [176].

Channel Loss Rate Two channel conditions were applied to allow for the characteristics of an erroneous channel: low and high loss rates. As the error rate measure, MPE-Frame Error Rate (MFER), which is defined by the DVB Community, is used to represent the losses in DVB-H transmission system. MFER is calculated as the ratio of the number of erroneous frames after decoding over the total number of frames. MFER 10% and 20% values were chosen as representative of a low rate and a higher rate with the goals of 1) having different perceptual qualities and 2) maintaining acceptable perceptual quality for watching the high error rate condition on a mobile device.

4.4.1.3. Contents

Four different contents were used to create the stimuli for the test. The selection criteria for the videos were spatial details, temporal resolution, amount of depth, and the user requirements for mobile 3D television and video (Table 4.3).

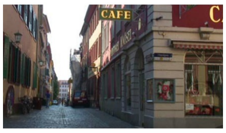
Screenshot	Genre and their audiovisual characteristics
	Animation – Knight's Quest 4D (60 s @ 12.5fps) V_{SD}: high, V_{TD}: high, V_D: med, V_{DD}: high, V_{SC}: high, A: music, effects
	**Documentary – Heidelberg (60 s @ 12.5fps)** V_{SD}: high, V_{TD}: med, V_D: high, V_{DD}: low, V_{SC}: low, A: orchestral music
	Documentary – Rhine Valley (60 s @ 12.5fps) V_{SD}: med, V_{TD}: low, V_D: med, V_{DD}: low, V_{SC}: low, A: orchestral music
	User-generated Content – Rollerblade (60 s @ 12.5fps) V_{SD}: high, V_{TD}: high, V_D: high, V_{DD}: med, V_{SC}: low, A: applause, rollerblade sound

Table 4.3. – Snapshots of the four contents being assessed (V_{SD}=visual spatial details, V_{TD}=temporal motion, V_D=amount of depth, V_{DD}=depth dynamism, V_{SC}=amount of scene cuts, and A: audio characteristics)

4.4.1.4. Production of Test Material and Transmission Simulations

The test sequences were prepared using the parameters shown in Table 4.4. First, each content was encoded with the three coding methods applying slice mode on and off. During the encoding, the QP parameter in the JMVC software was varied to achieve the target video bit rate of 420 kbps. The bit streams were encapsulated into transport streams using EEP and UEP, generating a total of twelve transport streams. The encapsulation was accomplished using the FATCAPS software

[179]. For each transport stream, the same burst duration was assigned for all left and right (depth) views to achieve fair comparisons by allocating the same resources. Finally, low and high loss rate channel conditions were simulated for each stream. The preparation procedure resulted in 24 test sequences. The loss simulation was performed by discarding packets according to an error trace at the TS packet level. Then, the lossy compressed bit streams were generated by decapsulating the lossy TS streams using the decaps software [180]. Finally, the video streams were generated by decoding the lossy bit streams with the JMVC software. For the error concealment, frame/slice copy from the previous frame was employed. For a detailed description of the prepared loss simulations see Strohmeier et al. [217].

Transmission parameter	Value
Modulation	16 QAM
Convolutional Code Rate	2/3
Guard Interval	1/4
Channel Bandwidth	8 MHz
Channel Model	TU6
Carrier Frequency	666 MHz
Doppler Shift	24 Hz

Table 4.4. – Parameters of the transmission used to generate transport streams for Study 3.

4.4.1.5. Stimulus presentation

The presentation setup was the same as described in Study 2 (see section 4.3).

4.4.2. Test procedure

The test procedure followed the same methodology as described in Study 2. During the study, visual discomfort was evaluated using the Simulator Sickness Questionnaire (SSQ) [165]. The results of the SSQ showed effect in oculomotor and disorientation for the first post-task measure. However, the effect quickly decreased within twelve minutes after the test to pretest level [222].

4.4.2.1. Methods of Analysis

Psychoperceptual Evaluation Non-parametric methods of analysis were used (Kolmogorov-Smirnov: $p < .05$) for the acceptance and the preference data. Acceptance ratings were analyzed using Cochran's Q and McNemar's Test. Cochran's Q is applicable to study differences between several related, categorical samples, and McNemar's test is applied to measure differences between two related, categorical data sets [4]. Comparably, to analyze overall quality ratings, a combination of Friedman's test and Wilcoxon's test was applied to study differences between the related, ordinal samples. The unrelated categorical samples were analyzed with the corresponding combination of Kruskal-Wallis H and Mann-Whitney U tests [4].

Sensory Profiling The sensory data were analyzed using R and its FactoMineR package [181, 182]. Multiple Factor Analysis (MFA) was applied to study the underlying perceptual model. Multiple Factor Analysis is applicable when a set of test stimuli is described by several sets of variables. The variables of one set therefore must be of the same kind [99, 138]. Hierarchical Multiple Factor Analysis (HMFA) was applied to study the impact of content on the perceptual space [109]. The structure of this data set is visualized in Figure 4.10.

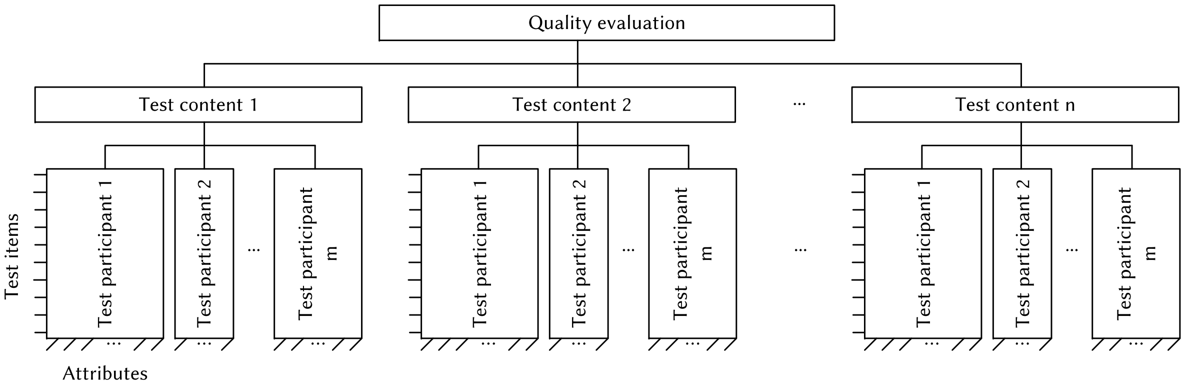


Figure 4.10. – The principle of the hierarchical structure in the test set of Study 3.

External Preference Mapping Partial Least Square Regression was conducted using MATLAB and the PLS script provided by Abdi [146] to link sensory and preference data. To compare the results of the PLS regression in terms of validity and reliability, an additional PREFMAP was conducted using XLSTAT 2010.2.03.

4.4.3. Results

4.4.3.1. Psychoperceptual Evaluation

Acceptance of Overall Quality In general, all MFER10 videos had higher acceptance ratings than the MFER20 videos ($p < .001$) (Figure 4.11). Also the error protection strategy showed significant effect (Cochran Test: Q = 249.978, $df = 7$, $p < .001$). The acceptance rate differed significantly between equal and unequal error protection for both MVC and VD codec (both: $p < .001$). In addition, the error protection strategy had no effect on the MFER20 videos (both: $p > .05$). Comparing the different slice modes, a significant effect can be found only between videos with VD coding and error rate 10% (MFER10) (McNemar's Test: $p < .01$, all other comparisons $p > .05$). Videos with slice mode turned off were preferred in general, except for Video + Depth videos with high error rates, which had higher acceptance in slice mode. Concerning the applied coding method, the results of the acceptance analysis revealed that for MFER10 MVC and VD had higher acceptance ratings than Simulcast ($p < .001$). The MVC coding method had significantly higher acceptance ratings than the other two coding methods for MFER20 ($p < .01$).

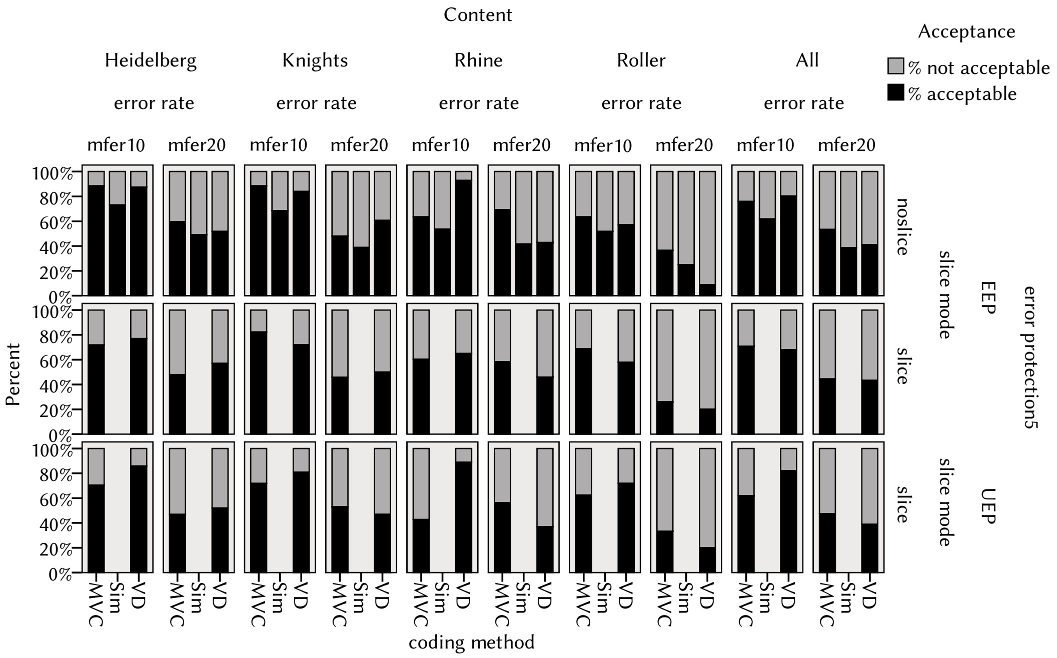


Figure 4.11. – Acceptance ratings in total and content by content for all variables.

To identify the acceptance threshold, we applied the approach proposed by Jumisko-Pyykkö et al. [39] (Figure 4.12). Due to related measures on two scales, the results from one measure can be used to interpret the results of the other measure. The Acceptance Threshold method connects binary acceptance ratings to the overall satisfaction scores. The distributions of acceptable and unacceptable ratings on the satisfaction scale differ significantly ($\chi^2(10)$ = 2117.770, *df* = 10, *p* < .001). The scores for nonaccepted overall quality fell between 1.6 and 4.8 (Mean: 3.2, SD: 1.6). Accepted quality was expressed with ratings between 4.3 and 7.7 (Mean: 6.0, SD: 1.7). So, the Acceptance Threshold was determined to be between 4.3 and 4.8.

Satisfaction with Overall Quality The test variables had significant effect on the overall quality when averaged over the content (F_r = 514.917, *df* = 13, *p* < .001). The results of the satisfaction ratings are shown in Figure 4.13 averaged over contents (all) and content by content. Coding methods showed significant effect on the dependent variable (Kruskal-Wallis: mfer10: H = 266.688, *df* = 2, *p* < .001; mfer20: H = 25.874, *df* = 2, *p* < .001). MVC and VD outperformed Simulcast coding method for MFER10 and MFER20 videos (all comparisons vs. Sim: *p* < .001) (Figure 4.13). For MFER10, Video + Depth outperformed the other coding methods (Mann-Whitney: VD vs. MVC: Z = -11.001.0, *p* < .001). In contrast, MVC received significantly higher satisfaction scores for MFER20 (Mann-Whitney: MVC vs. VD: Z = -2.214.5, *p* < .05).

The error protection strategy had an effect on overall quality ratings (Friedman: F_r = 371.127, *df* = 7, *p* < .001). MFER10 videos with equal error protection were rated better for MVC coding method (Wilcoxon: Z = -6.199, *p* < .001). On the contrary, MFER10 videos using VD coding method were

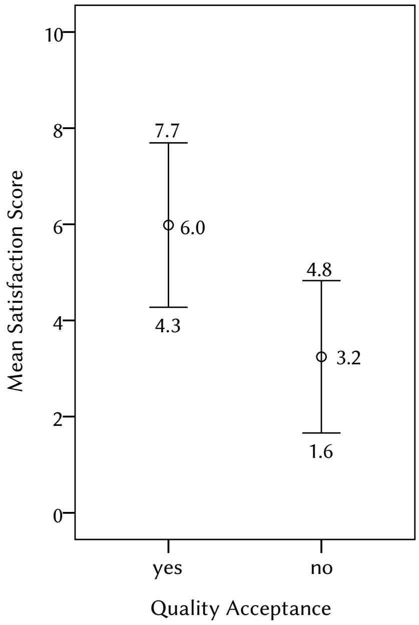


Figure 4.12. – Identification of the Acceptance Threshold. Bars show means and standard deviation.

rated better with unequal error protection (Z = -7.193, *p* < .001). Error protection strategy had no significant effect for MFER20 videos (Figure 4.13) (Z = -1.601, *p* = .109, ns). MFER10 videos with slice mode turned off were rated better for both MVC and VD coding method (all comparisons *p* < .05). Mfer20 videos were rated better when slice mode was turned on, with significant effect for VD coded videos (Z = -2.142, *p* < .05) and no significant effect for videos coded with MVC method (Z = -.776, *p* > .05, ns). In contrast to the general findings, the results for content Roller show that videos with slice mode turned on were rated better for all coding methods and error rates than were videos without slice mode (Figure 4.13).

4.4.3.2. Sensory Evaluation

One hundred sixteen individual attributes were developed during the sensory profiling session. The average number of attributes per participant was 7.25 (min: 4, max: 10). The results of the Multiple Factor Analysis are shown as representations of test items (item plot, Figure 4.14) and attributes (correlation plot, Figure 4.15). The item plot shows the first two dimensions of the MFA. All items of the content Roller are separated from the rest along both dimensions. The other items are separated along dimension 1 in accordance to their error rates. Along dimension 2, the Knight items separate from the rest of the items on the positive polarity.

A better understanding of the underlying quality rationale can be found in the correlation plot. The interpretation of the attributes can help to explain the resulting dimensions of the MFA. The negative polarity of dimension 1 is described with attributes like ’grainy’, ’blocks’, or ’pixel errors’, clearly referring to perceivable block errors in the content. Also, attributes like ’video stumbles’

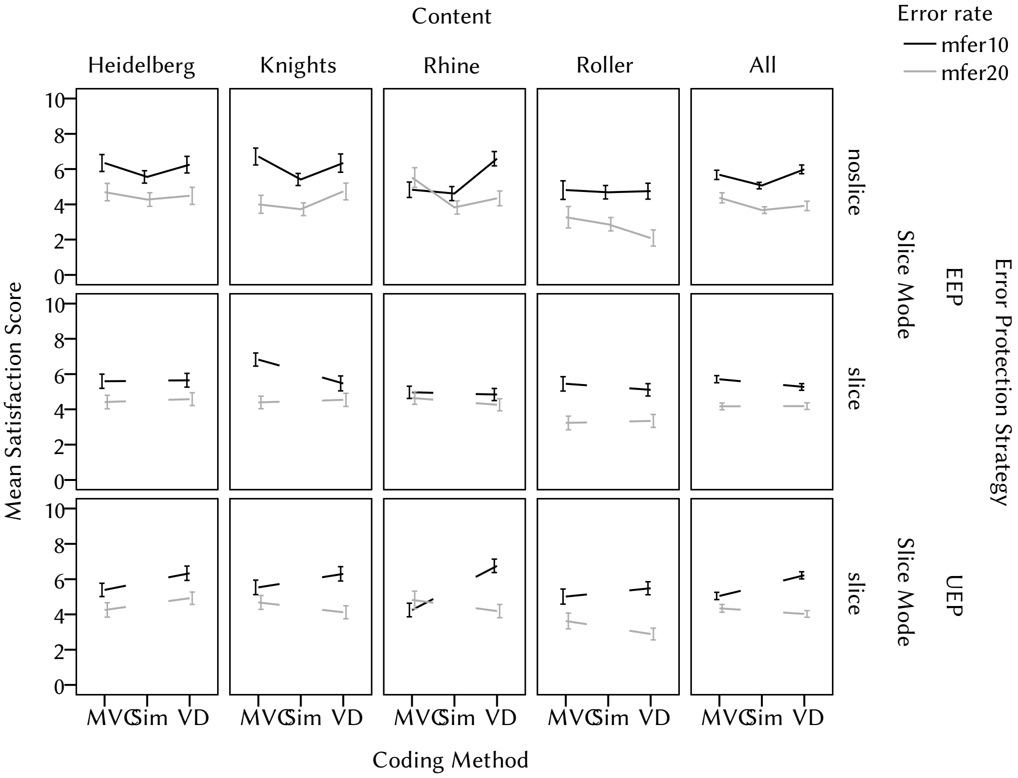


Figure 4.13. – Overall quality for all variables in total and content by content.

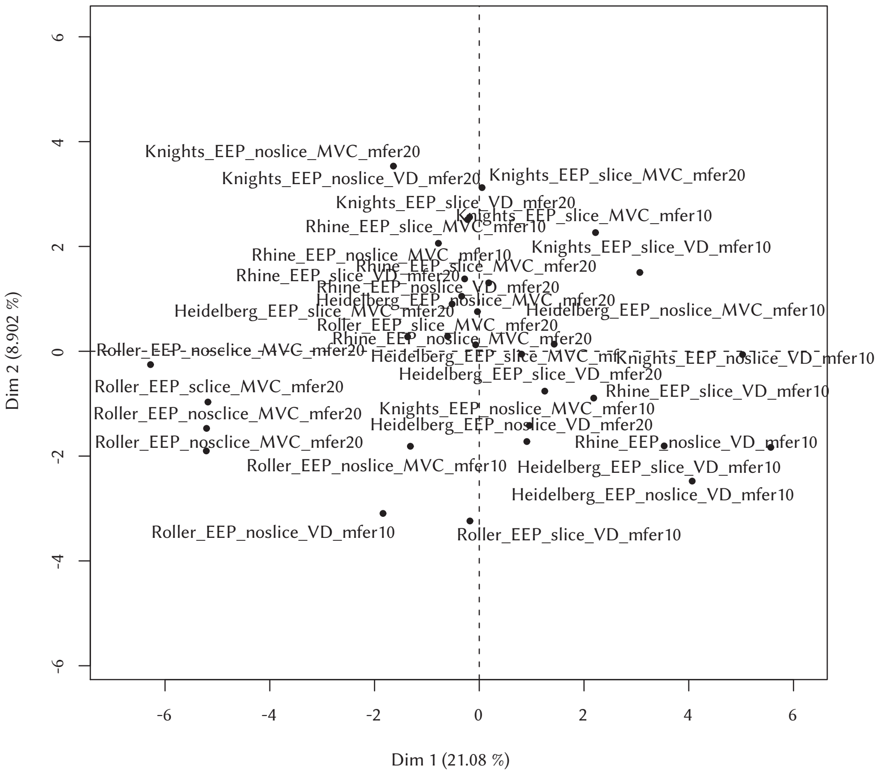


Figure 4.14. – Item plot of the Multiple Factor Analysis

can be found describing the judder effects of lost video frames during transmission. In contrast, the positive polarity of dimension 1 is described with the terms ’fluent’ and ’perceptibility of objects’, indicating an error-free case in the videos. Confirming the findings of our previous studies, this dimension is also described with 3D-related attributes like ’3D ratio’ or ’immersive’. Dimension 2 is described with attributes like ’motivates longer to watch’, ’quality of sound’, and ’creativity’ on the positive polarity. It also shows partial correlation with ’images distorted at edges’ or ’unpleasant spacious sound’ on the negative side. The identified separation of contents Knight and Roller along dimension 2 in the item plot indicates that dimension 2 must be regarded as a very content-specific dimension. It describes very well the specific attributes that people liked or disliked about the contents, especially the negative descriptions of Roller.

This effect is further supported by the individual factor map (Figure 4.16). The MFA routine in FactoMineR allows defining additional illustrative variables. We defined the different test parameters as illustrative variables. The lower the value of an additional variable, the lower is its impact on the MFA model. The results confirm the findings of the quantitative analysis. Contents Knight (c2) and Roller (c4) were identified as the most impacting variables. An impact on the MFA model can also be seen for the different MFER rates (m1, m2) and for the coding methods (cod1, cod2). The two slices modes (on, off) show only low values, confirming their low impact on perceived quality.

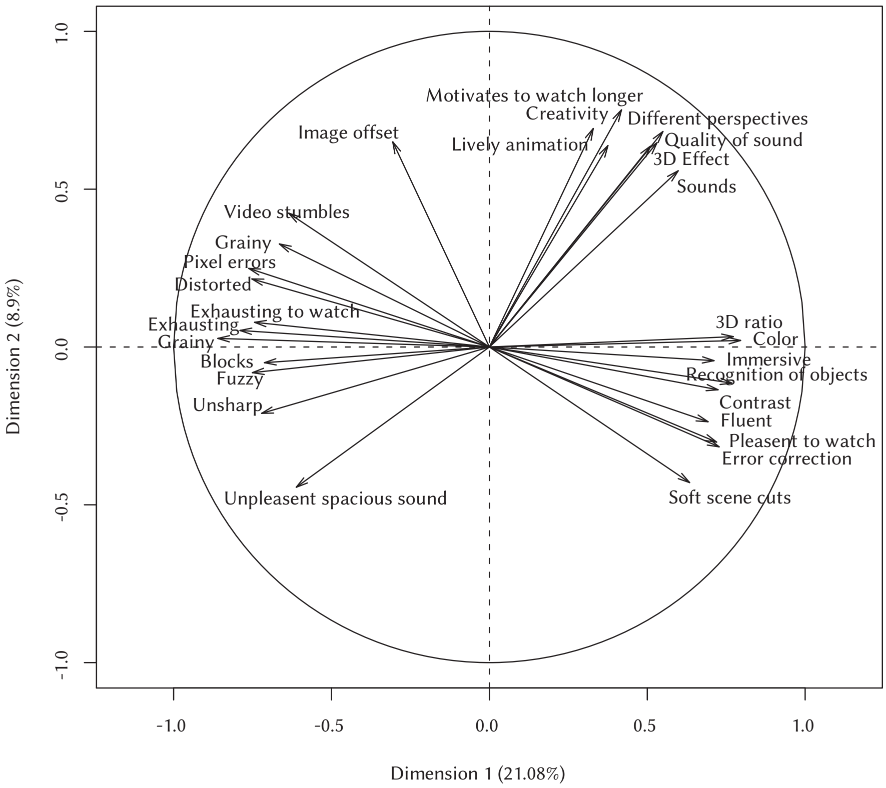


Figure 4.15. – Correlation plot of the Multiple Factor Analysis. For the sake of clarity, only attributes having more than 50% of explained variance are shown.

As an extension of MFA, the Hierarchical Multiple Factor Analysis can be used to further study the significant impact of the content on the perceived quality. For the HMFA, we assumed that each test item was a combination of a set of parameters applied to a specific content. The results are superimposed on the different contents (Figure 4.17).

Each parameter combination is shown at the center of gravity of the partial points of the contents. Figure 4.17 confirms that the test participants were able to distinguish between the different parameters. The parameter combinations are separated in accordance to the MFER rate and the coding method. Only slice mode shows little impact. However, it is noticeable that the different contents affect the evaluation of the test parameters. The lines around the center of gravity show the impact of the contents. While, for the high error rate, the impact of the contents is rather low, as shown by the location of partial points close to the center of gravity, the low error rate does show some impact.

4.4.3.3. External Preference Mapping

The next step of the OPQ approach is to connect users' quality preferences and the sensory data. In the current Extended OPQ approach, a Partial Least Square Regression was applied. To show the differences of the PLS regression and the commonly applied PREFMAP approach, a comparison of

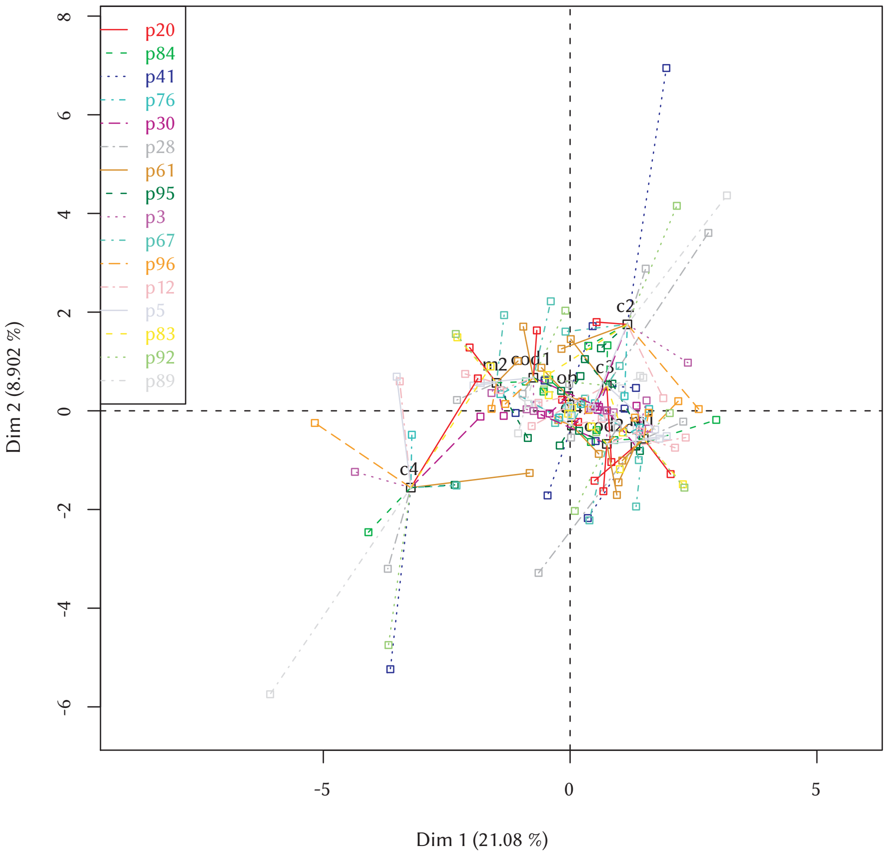


Figure 4.16. – Individual factor map of the MFA. The test parameters were used as supplementary variables in the MFA, and their impact on the MFA results is illustrated by the points of content (c1-c4), coding method (cod1, cod2), error protection (m1, m2), and slice mode (on, off).

both results is presented. For both cases, a clear preference structure can be found in the data set (see Figure 4.18 and Figure 4.19).

The result of PREFMAP is shown in Figure 4.18, which shows that, for all test participants, a clear quality preference exists (red dots in Figure 4.18). Taking into account the MFA correlation plot (Figure 4.15), it can be seen that the preferences are described with terms like ’immersive’ (P12.5), ’contrast’ (P5.10) or ’soft scene cuts’ (P83.4). However, Figure 4.18 also shows that the underlying model of PREFMAP is similar to the MFA because only preferences are regressed onto it. Dependencies from preferences cannot be taken into account.

The PLS result is given as a correlation plot in Figure 4.19. It also shows a clear preference of all test participants. When interpreting the main components of the PLS, two different groups of attributes can be found. The first group relates to artifact-free and 3D perception for good quality (e.g., P5.6 ’perceptibility of objects’, P12.5 ’immersive’). The other group is described with attributes relating to visible blocks and blurriness (P96.7 ’unsharp’, P28.4 ’pixel errors’). Hence, the first component of the PLS model is related to the video quality descriptions with respect to spatial quality. Although this finding supports the findings of the MFA, a second group of attributes influencing

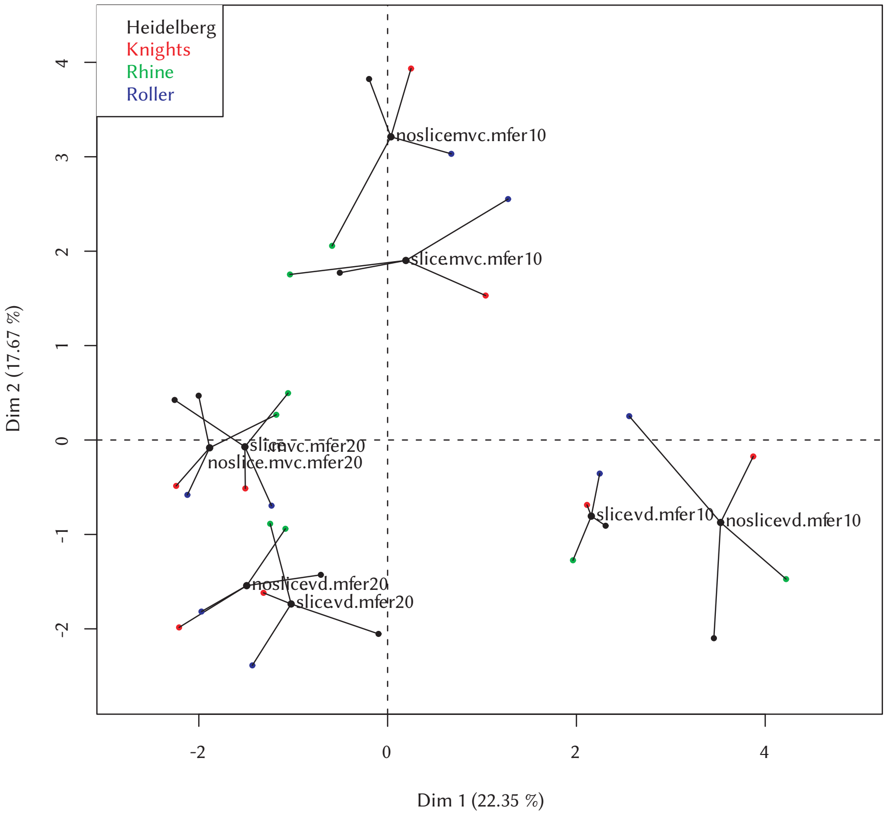


Figure 4.17. – Superimposed representation of the test parameter combinations and the partial clouds of contents.

the PLS model can be found. These attributes describe the video quality related to temporal quality. For good quality, attributes like P30.4 'fluent movement' can be found; for bad quality, attributes like P20.3 'time jumps' or P84.5 'stumble', which correlates only with dimension 2 of the PLS model, can be found. Interestingly, the EPM results are not fully comparable to each other in terms of preferences. This second component cannot be identified in the MFA results. An explanation for the differences between the two approaches can be found in the way the respective latent structures (or models) are developed. A possible interpretation of the result is that, in the quantitative evaluation, test participants evaluated the overall quality more globally. Thus, fluency of the content is the strongest global quality factor. When performing a sensory evaluation, test participants seemed to concentrate on a more detailed evaluation of the content, and spatial errors had greater impact.

4.4.4. Discussion and Conclusion

The aim of this study was to investigate the quality factors in transmission scenarios for mobile 3D television and video. The study highlights the importance of the Open Profiling approach because it allows for studying and understanding quality from different points of view. The results complement each other, and every part of the Extended OPQ approach supports the findings of the previous

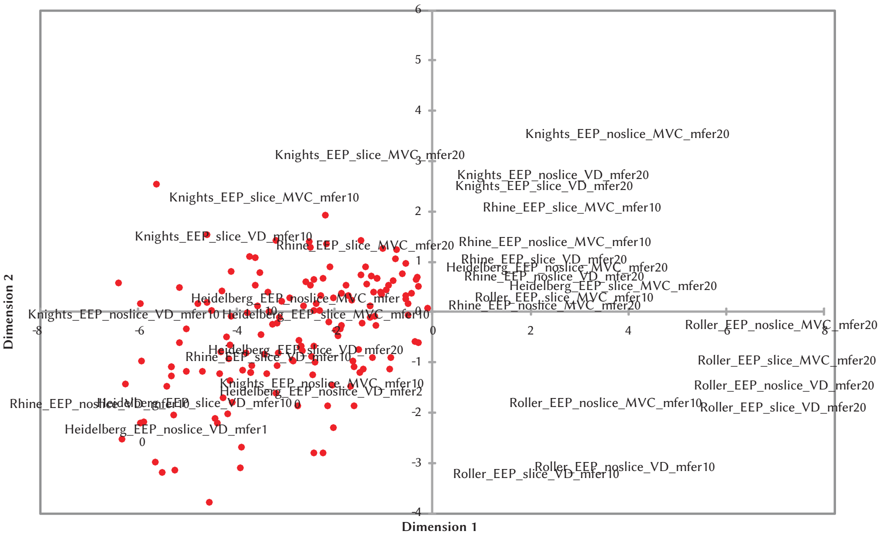


Figure 4.18. – PREFMAP result of the transmission study. The red dots show the quantitative preferences of each assessor regressed onto the MFA result

steps and deepens the understanding of Quality of Experience in mobile 3D video transmission. The author investigated the impact of different transmission settings on the perceived quality for mobile devices. Two different error protection strategies (equal and unequal error protection), two slices modes (off and on), three different coding methods (MVC, Simulcast and Video + Depth), and two different error rates (MFER10 and MFER20) were used as independent variables.

The results of the psychoperceptual evaluation in accordance with ITU recommendations show that the provided quality level of MFER10 videos was good, being at least clearly above 62% of acceptance threshold for all contents while MFER20 videos were not acceptable at all, with only the content Heidelberg having acceptance slightly above 50%. This result indicates that an error rate of 20% is insufficient for consumer products, whereas an error rate of 10% would still be sufficient for prospective systems. The analysis of variance of the satisfaction scores revealed that all independent variables had a significant effect on test participants' perceived quality. The most significant impact was found for the coding methods. MVC and Video + Depth outperform Simulcast as coding methods, a result is in line with previous studies of the production chain of mobile 3D television and video [223]. Interestingly, the quantitative results also show that MVC is rated better than Video + Depth in terms of overall acceptance and satisfaction at high error rates. The findings of the psychoperceptual evaluation were confirmed and extended in the sensory evaluation. The Multiple Factor Analysis of the sensory data with the independent variables as supplementary data showed that, in the sensory data, all test variables had an impact. This result confirms that the test participants were able to distinguish between the different variables during the evaluation.

In addition, the idiosyncratic attributes describe the underlying quality rationale. Good quality is described in terms of sharpness and fluent playback of the videos. Furthermore, 3D-related attributes were correlated with good quality, confirming findings of other related studies [214, 223,

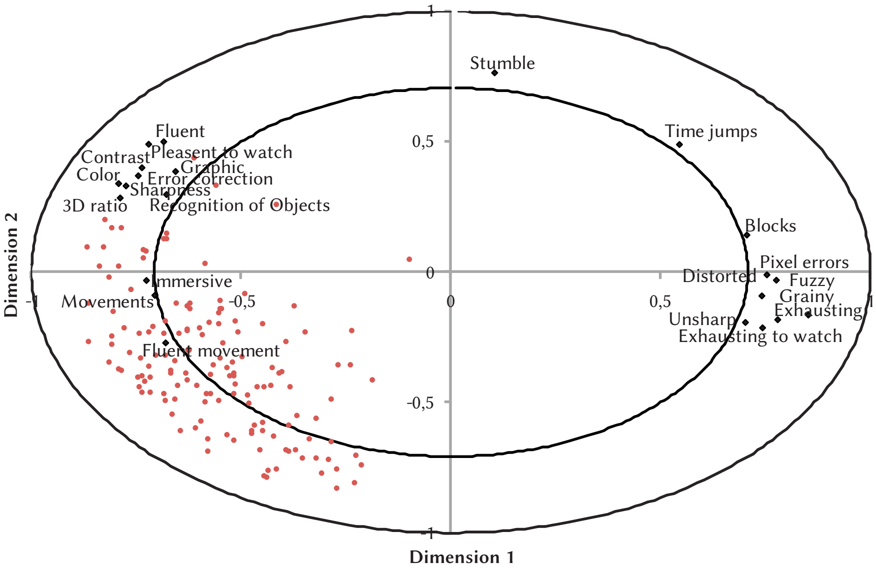


Figure 4.19. – The results of the External Preference Mapping as correlation plot conducted with PLS regression.

224]. Interestingly, bad quality was correlated with attributes that describe blocking errors in the content. These errors can be a result of both the coding method and the applied error protection strategies. The expected descriptions of judder as contrast to fluency of the test items appear rarely. In addition, MFA indicates a strong dependency of quality satisfaction on the contents used in the stimuli.

This finding was confirmed by the applied Hierarchical Multiple Factor Analysis in which a dependency of the transmission parameters from the contents was studied. These results confirm the psychoperceptual evaluation and sensory findings that content plays a crucial role in determining experienced quality of mobile 3D video. The HMFA results deepen the findings to suggest that content seems to become more important when the perceivable errors become are reduced. This finding is supported by the Partial Least Square Regression that linked sensory data and the preference ratings. Preferences all correlate with attributes that represent good quality in the MFA. Interestingly, the importance of judder-free stimuli is increasing in the PLS model. Because PLS takes into account both sensory and preference data to derive the latent structures, the results suggest that fluency was more important in the psychoperceptual evaluation than in the sensory evaluation. We see this result as an indicator that the quality evaluation of test participants differed slightly in the psychoperceptual and sensory analyses. While in the retrospective psychoperceptual evaluation a global attribute like fluency of the videos seemed to be crucial, test participants made a more detailed evaluation of quality in the sensory test and found more quality factors related to spatial details.

4.5. Study 4: External Validation of OPQ by probing the method in the context of use

Methodological contribution: Comparable to other descriptive approaches, great effort was expended in developing Open Profiling of Quality as an evaluation method under controlled laboratory conditions; however, its applicability and validity outside the laboratory are unknown. This study strengthens the external validity of the method by probing OPQ in the context of use and comparing these results to an evaluation in the laboratory. In addition, the study fuses psychoperceptual evaluation and sensory profiling in a single session in which the two tasks are conducted subsequently. Doing so significantly decreases the duration of an OPQ study and minimizes the risk of participants' dropping out in a multisession design.

Research problem for MOBILE3DTV: Artifact-free videos and depth perception were identified as the most important positive features contributing to test participants' Quality of Perception in OPQ studies conducted in laboratory environments. However, MOBILE3DTV has been identified for potential use in quite heterogeneous contexts. Therefore, a goal of this study was to determine whether the identified components of QoE are still valid for use in the system outside the controlled evaluation environment.

4.5.1. Research method

4.5.1.1. Test participants

Thirty-six untrained participants (age: 19-52 years) took part in the study, 16 female and 20 male. All test participants were tested for visual acuity (myopia and hyperopia: Snellen index: 20/40), color vision (the Ishihara test), and stereo vision (the Randot Stereotest 0.6). Five of the participants had been working in the field of video editing or video application. One of the participants had prior experience in subjective quality evaluation, but none of them with 3D video. All other test participants could be classified as naïve participants. The assessors were divided into two groups: a control group with 16 participants, who were tested under laboratory conditions, and a group of 21 participants, who were tested in a user context situation. Fifteen participants of each group were selected randomly for sensory evaluation.

4.5.1.2. Test stimuli

Six different audiovisual clips with a length of 20 seconds were selected for the test according to their audiovisual characteristics and the user requirements for mobile 3D television and video [216] (Table 4.5). They represented different genres such as documentary, sports, music video, and animation. Stimulus material was encoded using three different video qualities. The clips were edited using Premiere Pro CS4 and exported with a resolution of 640px x 480px for each channel. Audio was sampled using a 44.1 kHz rate at 16 bits. Quantization parameters (QP) instead of different bit rates were applied to generate different quality levels [223]: high at QP 30, medium at QP 40, and low at QP 45. The clips were encoded using open source encoders x.264 for video and Nero AAC for audio. Sample rates and resolution remained as in the editing part. Finally, Stereo Movie Maker was used to encode the prepared clips in the 3D-Avi format necessary for the presentation of the stimuli material.

Screenshot	Genre and their audiovisual characteristics
	Animation – Dracula V_{SD}: med, V_{TD}: high, V_{D}: high, V_{DD}: high, V_{SC}: high, A: music, effects
	Documentary – Macroshow V_{SD}: high, V_{TD}: med, V_{D}: high, V_{DD}: low, V_{SC}: med, A: orchestral music, ambience
	Sports – Skydiving V_{SD}: low, V_{TD}: med, V_{D}: med, V_{DD}: low, V_{SC}: low, A: music
	User-created Content – Street Dance V_{SD}: med, V_{TD}: high, V_{D}: high, V_{DD}: med, V_{SC}: low, A: music, ambience
	Documentary – The Eye V_{SD}: med, V_{TD}: med, V_{D}: med, V_{DD}: med V_{SC}: med, A: music
	Sports – 24h V_{SD}: med, V_{TD}: high, V_{D}: med, V_{DD}: high, V_{SC}: high, A: ambient music

Table 4.5. – Snapshots of the six contents under assessment (V_{SD}=visual spatial details, V_{TD}=temporal motion, V_{D}=amount of depth, V_{DD}=depth dynamism, V_{SC}=amount of scene cuts, A= audio characteristics)

4.5.1.3. Stimulus presentation

The tests were conducted in two different contexts (Table 4.6). The first context was a controlled laboratory environment in the listening laboratory at Ilmenau University of Technology, a setting in accordance with the ITU recommendations [5, 39]. For the second context, we chose a coffee bar as the most mentioned usage situation for mobile 3DTV [216]. In the café, the same time slot during the day and same place for each participant were used to obtain similar conditions for the study as defined for quasi-experimental settings [61].

Components/properties	Lab	Café
Physical context		
Functional place	Laboratory conditions	student café at TU Ilmenau
Sensed attributes (Audio, Visual)	A: calm, V: calm, indoors	A: noisy, V: noisy, indoors
Movements (Movement, Position)	M: none, P: straight	M: none, P: lean
Artifacts (other than answer sheet)	none	tea cup
Temporal context		
Duration	1.5 - 2 hours	1.5 - 2 hours
Time of day	Variant	Between 11.45 am and 3 pm
Actions-time	Extra time	Extra time
Task context		
Multitask 1	Quality evaluation	Quality evaluation
Multitask 2	none	Relax, drink tea/coffee
Interruptions	none	possible
Task type	Entertain	Entertain
Social context		
Persons present	Moderator	Moderator, other guests
Interpersonal actions	none	possible
Technical and informational context		
Other systems	none	none
Properties		
Level of dynamism	Static	dynamic
Other related factors		
Motivations	*	Entertain, pass time, relax
Viewing distance	Freedom to adjust	Freedom to adjust
Device volume	Freedom to adjust	Freedom to adjust

Table 4.6. – Characteristics of the contexts, described based on the Model of Context of Use for Mobile-Human-Computer-Interaction [61], operationalized in [9, 58]

The audiovisual clips were presented on an 8” FinePix Real 3D V1 Display based on parallax barrier technology. The FinePix Real 3D V1 Display provides a maximum resolution of 400 x 600 in 3D mode for each channel. Test participants were allowed to adjust their viewing distance so that they could perceive the video as three dimensional. The two stereo speakers integrated into the display were used for audio playback due to the possible inability to connect headphones. Audio playback occurred at a sampling rate of 11 kHz because it is the maximum audio sampling rate available for the device. The order of the clips was randomized to avoid bias effects.

4.5.1.4. Test procedure

Overall, the test procedure of the study followed the Open Profiling of Quality approach [214, 217]. The test procedure was divided into two parts conducted in one single session. The first part started with the visual screening and the explanation of the test procedure. In the following training and anchoring, we presented a subset of test items that covered the full range of quality. Test participants were asked to find their best viewing position and to practice the evaluation task. Then, an Absolute Category Rating (ACR) according to ITU-R P.910 [5] was conducted to evaluate the overall quality quantitatively. The stimuli were presented one by one, and the participants rated the acceptance of the quality on a binary scale and the overall satisfaction on an unlabelled 11-point scale, both retrospectively [57]. Each stimulus was assessed twice.

After a short break of about 10 minutes, the participants filled out a demographic data questionnaire. In the second part, participants were introduced to the sensory evaluation task. Then, during attribute elicitation, the participants watched a second subset of test items to develop their individual quality attributes according to their individual quality perceptions. Attribute refinement took place while the participants evaluated three clips according to their attributes. They were also asked to define their quality attributes and, if necessary, to reconsider whether they perceived some of the attributes as not being unique or whether they could define them precisely. At the end of the refinement, each of the final attributes was attached to a 10 cm long line with the labels 'min' and 'max' at its ends. In the final sensory evaluation, the stimuli were again presented one after the other, and the participants rated overall quality on all of their attributes for each test item. The participants were instructed to mark the sensation quality of an attribute on the line, using 'min' for no sensation of an attribute at all and 'max' for the maximum sensation of an attribute.

4.5.1.5. Methods of Analysis

The quantitative data were analyzed using non-parametric statistical analysis because no normal distribution was given for the test items (Kolmogorov-Smirnov: $p < .05$). The Friedman test was applied to determine whether the independent variables had an impact on the dependent one. Significant differences between two related items could then be measured using Wilcoxon's test. To compare the binary, non-related acceptance data between the contexts, Pearson's Chi-Square test was applied. For the pairwise comparison of satisfaction data between contexts, we applied the Mann-Whitney-U test. All quantitative data analysis was performed using PASW Statistics 18. The sensory data were analyzed by applying Multiple Factor Analysis (MFA). R and its FactoMineR package were used for sensory analysis. For each data set, a PREFMAP was conducted to obtain External Preference Mapping. Finally, Hierarchical Multiple Factor Analysis (HMFA) was applied to compare the two sensory data sets from the laboratory and external context, looking for similar information as a final step of checking validity.

4.5.2. Results

4.5.2.1. Psychoperceptual Evaluation

Acceptance of Overall Quality On average, all presented stimuli at qp30 provided a highly acceptable quality level of 93%, qp40 stimuli reached an average level just above 50%, and qp45 received an acceptance rate of 12% (Figure 4.20). Comparison of the results from the two contexts did not reveal a significant difference, with the exception of the contents of Streetdance and The Eye at medium quality level (Pearson's χ^2: $p < .05$).

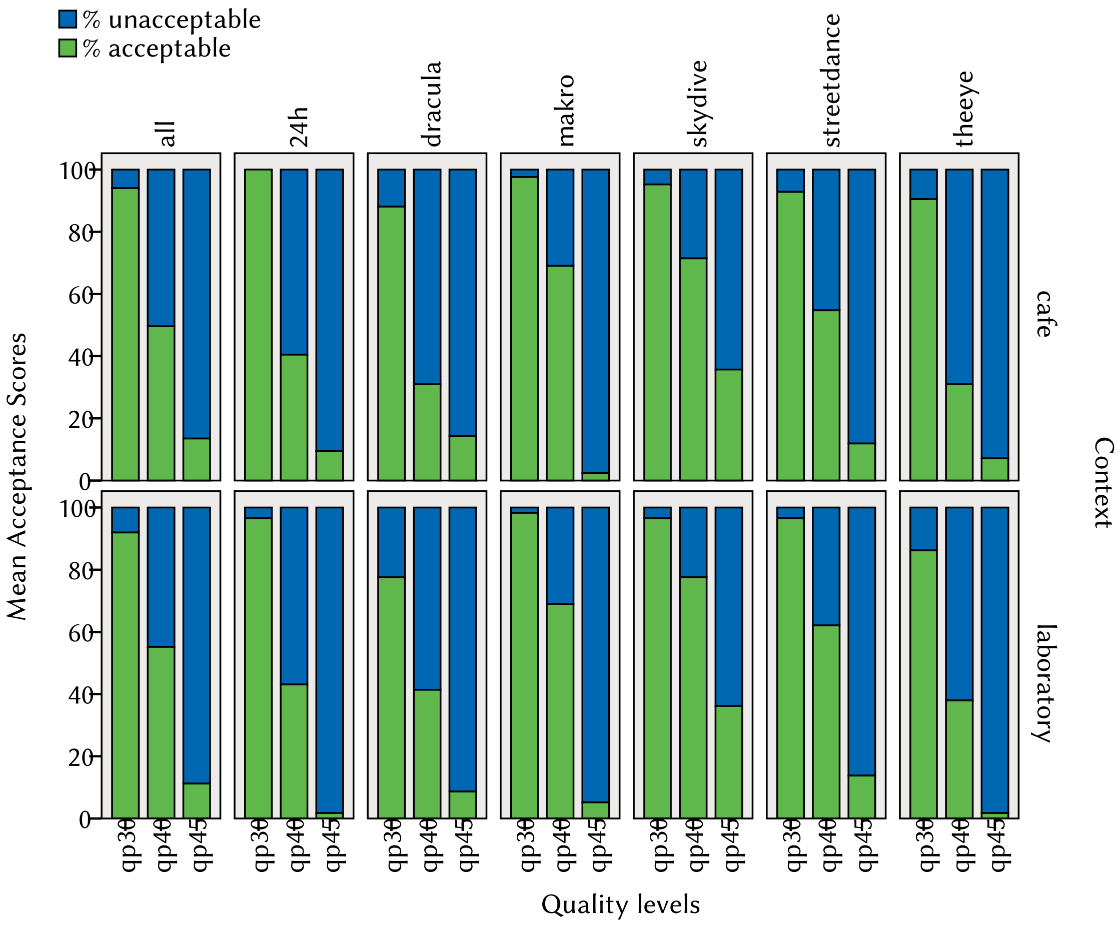


Figure 4.20. – Overall acceptance scores for the items tested

Satisfaction with overall quality Parameter combinations influenced overall quality satisfaction when averaged over the content for each of the contexts (laboratory: F_r = 627.705, df = 2, p < .001; café: F_r = 419.846, df = 2, p < .001). In a comparison of satisfaction scores between contexts, no significant difference was found except for content The Eye. For The Eye, the quality at QP45 was slightly better rated in the café than in the laboratory (Mann-Whitney-U: U = -2.305, p < .05). Figure 4.21 shows the overall quality scores averaged over contents and content by content for the different QPs and contexts. QP30 provided significantly higher quality satisfaction, with the rating for QP45 being worst (all comparisons: p < .001). The results of content-by-content analysis follow the overall tendency for different QPs (all comparisons: p < .001).

The analysis of differences among contents per QP revealed interesting findings in differences of satisfaction scores (Friedman: all comparisons: p < .001). Although in both contexts 24h was among the highest rated contents at QP30, it was among the worst rated contents for QP45. Comparable results can be found for Makro. In contrast, Streetdance was the significantly best rated content at QP45 (Skydive vs. Streetdance: laboratory: Z = -5.103, p < .001; café: Z = -4.118, p < .001), although it was rated average for QP30.

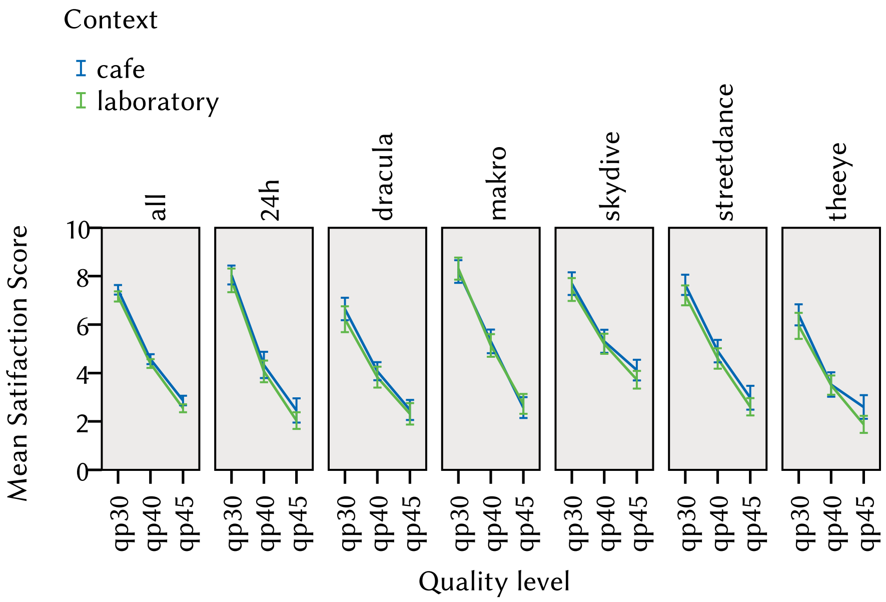


Figure 4.21. – Mean satisfaction scores for the items tested in Study 4. Error bars show 95% CI of mean.

4.5.2.2. Sensory Evaluation and External Preference Mapping

Fifteen participants per context developed 91 individual quality attributes for the laboratory and 78 attributes in the café. The average number of attributes per participant was 6 (min: 4, max: 7) and 5 (min: 4, max: 8) for laboratory and café, respectively.

Laboratory The results for the sensory data from the laboratory are shown by item and correlation plot in Figure 4.22 and Figure 4.23. The item plot (Figure 4.22) shows the loadings of the test items on the first and second components of the MFA. The first two components of the MFA explain 56.42% of the variance in the individual data (also called explained variance) with 44.25% and 12.17%, respectively. For the first component, the items separate along the different QPs. Items close to the origin have less impact on the component than those with high (positive or negative) loadings. For the second component, a clear separation of all items of content Dracula can be found. These items show high impact for the second component of the model.

Further insight can be obtained from the correlation plot (Figure 4.23), which shows the correlation of each individual attribute with the first and second component of the MFA model. The first component is mainly described with attributes like 'blocky' or 'artifacts' on its negative polarity while the positive polarity correlates strongly with attributes like 'clear, 'sharpness of edges', or '3D effect'. These items describe the differences in the perception of video quality and are in accordance with the separation of QPs along the first dimension. The second dimension is correlated with such attributes as 'double images' on the one polarity, with few attributes like 'color-fast' and 'perceivable as one image' on the other. Considering that, along this component, content Dracula separated from the other content, a problem with obtaining a proper 3D perception is evident. The double

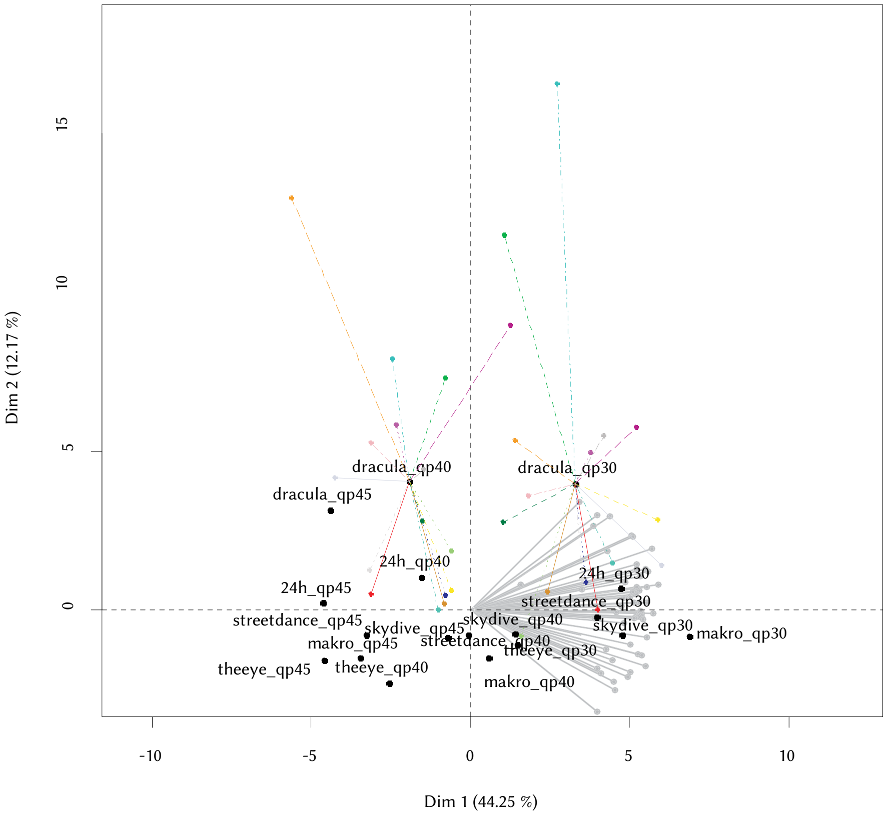


Figure 4.22. – Item plot and partial loadings for the laboratory. The partial loadings show individual participant's impact. The gray arrows mark the quantitative preferences of test participants from PREFMAP.

images may have been caused by high disparity in this content. A few attributes correlate with both dimensions, such as '3D effect', 'depth', and 'amount of 3D'. However, the partial plot (Figure 4.22) shows that this problem occurred for only some participants. While some participants indicated very high loadings for their individual configurations on dimension 2, others had none.

As the final step in the full Open Profiling of Quality, an External Preference Mapping was conducted. The results of the PREFMAP are included as gray errors in Figure 4.22. The PREFMAP reveals a clear preference structure towards the high quality levels, a result in line with the psychoperceptual findings. The variation of preferences along dimension 2 can be explained by the different perceptions of the Dracula content. Either lower preferences were indicated due to perception of crosstalk, or high preferences were indicated due to a good 3D perception of this content.

Café The sensory results for the café are comparable to those obtained in the laboratory (Figure 4.24 and Figure 4.25). The first two components of the MFA model account for 47.76% of explained variance (component 1: 33.48%, component 2: 14.28%). As for the laboratory results, the items separate along the first dimension according to their QPs (Figure 4.24). The separation of content Dracula along the second component can be seen as well.

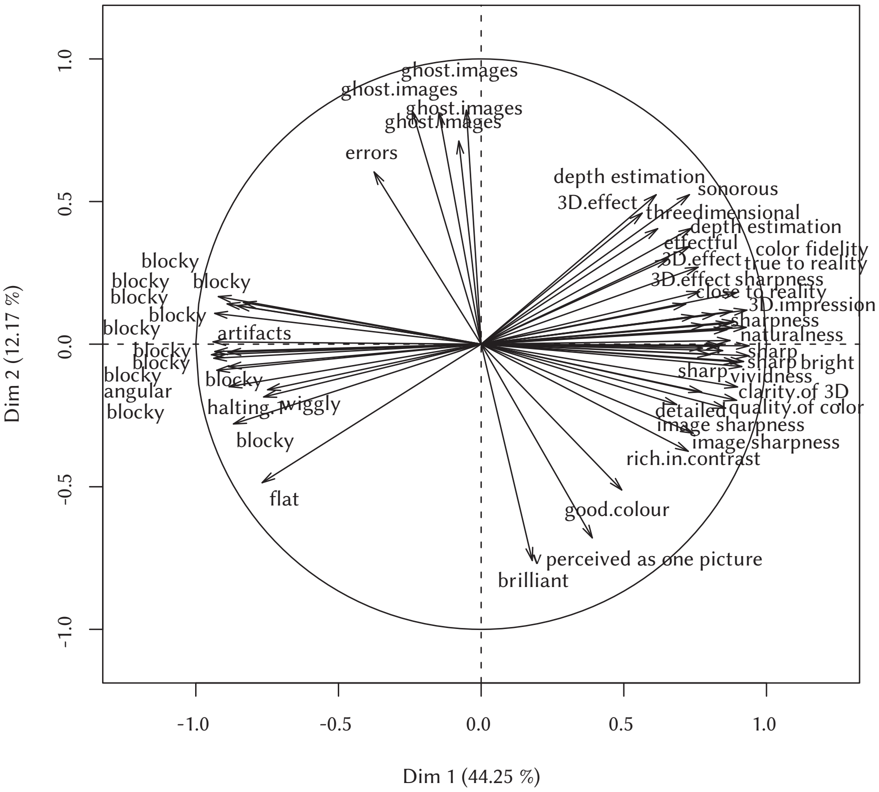


Figure 4.23. – Correlation plot of the laboratory evaluation.

The correlation plot (Figure 4.25) shows high correlation of attributes like 'blocky' or 'blurry' and, in contrast, of attributes like 'spacious', 'rich in details', and 'clear' with the first components. The second component is correlated with attributes like 'double effect', 'dark', and 'annoying' on its one polarity. The other polarity correlates with 'bright' and 'realistic'. Again, the partial plots show the differences of individual contributions to the second dimension. The External Preference Mapping for the contextual data is included in Figure 4.24.

Comparison of results The final step of the analysis is the comparison of the two MFA models obtained from laboratory and café. A simple solution for comparison is the description of differences and similarities between the results of the individual models. Overall, the separate analysis has shown that the first two dimensions of each model describe similar things. While the first component relates to video quality, the second component refers to quality factors in relation to display and disparity problems. In addition, the attributes, also developed by different participants, are very similar in describing the two components. However, a difference can be found in the number of attributes. In general, attributes that have a correlation higher than 0.5, that is, contributing to at least 50% of the explained variance, are regarded as more important than the rest. For the laboratory, 61.5% of all attributes contribute to this criterion while, for the café, only 44.6% of the attributes contribute to 50% of the explained variance.

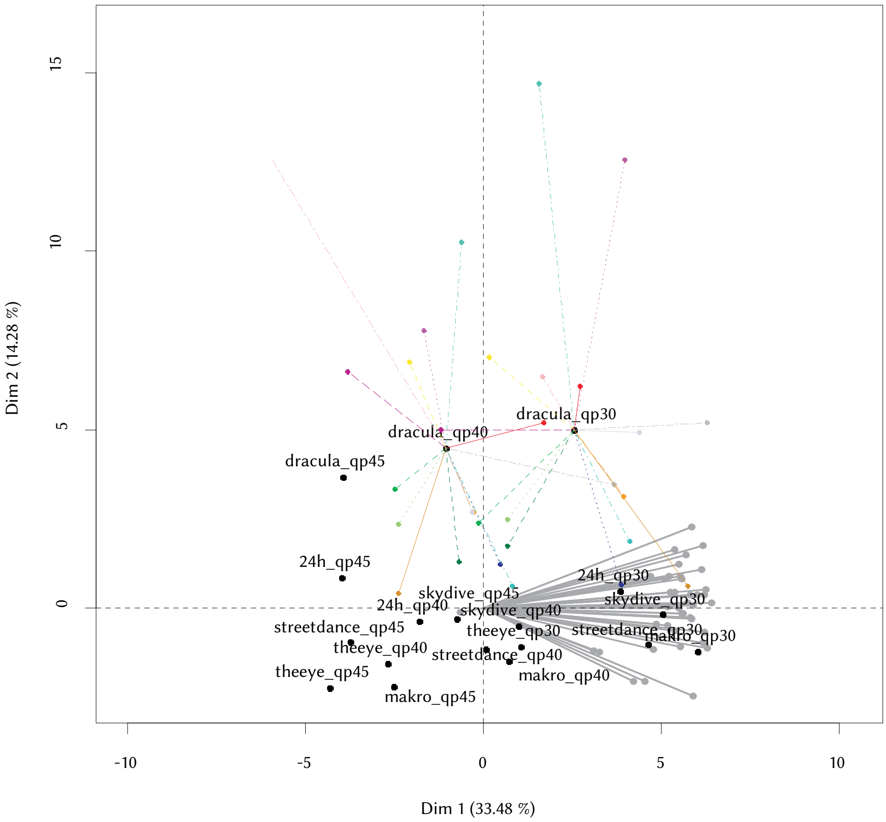


Figure 4.24. – Item plot and partial loadings for the café. The partial loadings show individual participant's impact. The gray arrows mark the quantitative preferences of test participants from PREFMAP.

A majority of the attributes for the laboratory MFA model show high correlation with the first dimension while only few correlate with the second. For the café MFA results, it is noticeable that the number of attributes along the first component is lower. In addition, there are more inter-dimensional attributes; that is, the dimensions are not as well separated as in the laboratory model.

Although the differences have already been noted, we want to determine whether the models are different. The HMFA results confirm the previous findings and allow modeling the comparison in a joint analysis of both data sets. In the HMFA results, each test item is plotted at the center of gravity between both data sets. In addition, the partial clouds for each data set are plotted to see the separate impact of laboratory and café data. The HMFA model levels the separate models in terms of explained variance (51.12%; 37.93% for component 1 and 13.19% for component 2) and loadings of the items (Figure 4.26). In this joint model, again the different QPs separate along the first component. The analysis shows that the deviation between the data sets along the first component of the HMFA model is low. Along the second dimension, we can identify differences in the impact of the partial clouds. The deviation between the data sets along this component is much higher, especially for content Dracula. Along the second dimension, the café data show higher loadings than does the laboratory data set.

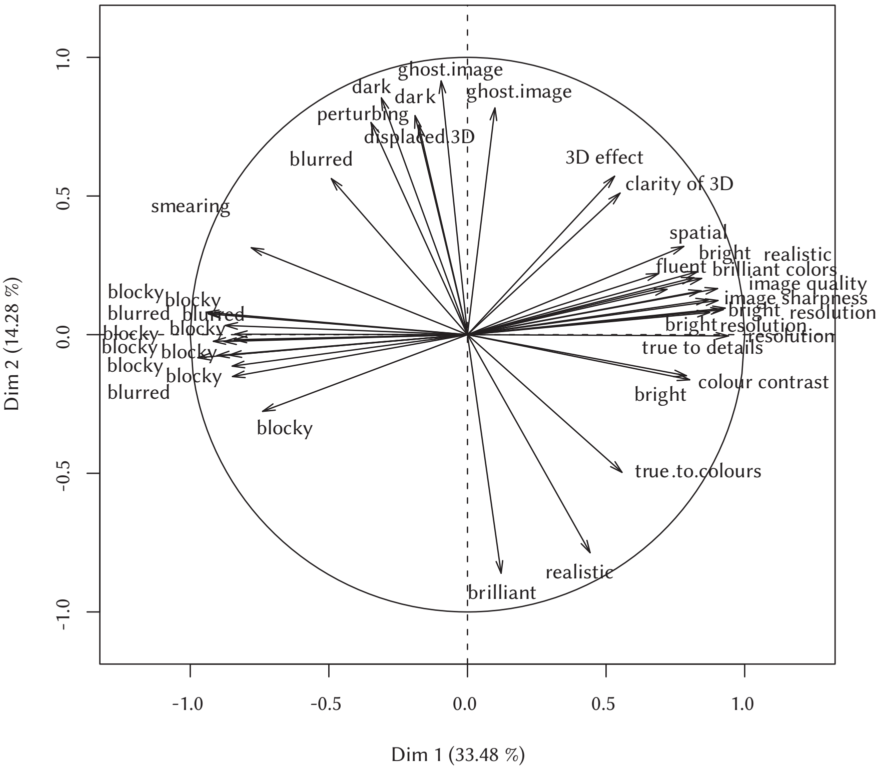


Figure 4.25. – Correlation plot of the evaluation in the context of use.

4.5.3. Discussion and Conclusion

The goal of this study was to validate the quality models that can be obtained using the Open Profiling of Quality approach in a comparison between data from a laboratory environment and a context-of-use environment. Within the User-Centered Quality of Experience approach, the descriptive evaluation of quality and its evaluation in the context of use were the two key approaches combined in this study.

The OPQ approach allows for a combined evaluation of quality using quantitative evaluation and qualitative descriptive sensorial profiling methods. The quantitative results show high impact of the different quality levels on users' quality perception in both contexts. In addition, a content dependency of the results was identified. However, a difference between the two contexts could not be identified. The two descriptive models obtained from the evaluations in the laboratory and in a café confirm both previous findings about experienced quality of mobile 3D television [58][214, 217]. They give deeper knowledge to explain the quantitative results. From the models, the author was able to identify two main components that describe users' perceived quality. The most important component is video quality. As in previous studies, good video quality includes descriptions of 3D perception, explaining the content dependency among different QPs. Contents Makro and 24h are rich in detail and offer better 3D perception than other contents. When video quality decreases, the

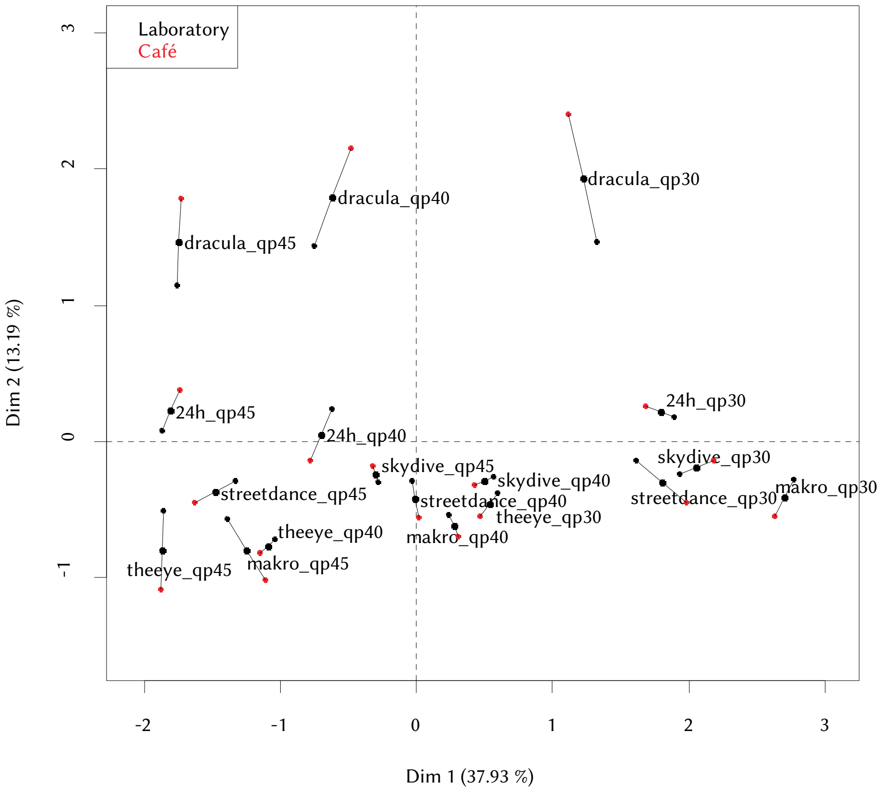


Figure 4.26. – Result of the Hierarchical Multiple Factor Analysis. The partial clouds show the impact of the laboratory data set (black) and the data set from the café (red) on the joint model

added value is no longer evaluated, and the satisfaction of quality with these two contents decreases with the loss of details.

Beside the video quality, the impact of perceivable double images was found. These descriptions largely correlated with content Dracula and arose from a high amount of disparity and resulting crosstalk between the left and right channels. For this dimension, two important findings were made. First, the analysis shows that crosstalk was perceived only by several test participants although participants were screened for the same visual abilities at the beginning of the test. This finding deepens the need for tools for better screening and description of the test sample [220]. In addition, test participants showed higher sensitivity to this component in the context of use. This finding is confirmed by previous studies in which the ease of use and the viewing comfort for mobile 3D television were identified as important components of quality in the context of use [9, 58].

Limitations to this study exist in the missing detailed recording of the characteristics of the context of use during the study. Although we tracked special events during the study by writing them down, a detailed recording over time, for example, a video recording, is missing. This gap does not allow for deeper analysis of shared attention and makes it impossible to report this concept more accurately as reported in other studies on the context of use [9]. However, the results of the study give valuable knowledge for understanding more deeply the interaction of quality perception with the interaction of perception and shared attention in the context of use.

In summary, the results show that descriptive quality models obtained by applying the OPQ method are, overall, very similar between the two contextual settings. From both models, two main components, video quality and crosstalk, were identified, and the loadings of attributes and the correlation of individual attributes were equal to each other. This finding contributes to the overall validity of OPQ results as attested in several other studies [214, 217]. However, differences were still identified between the evaluation in the laboratory and in the context-of-use environment, underscoring the importance of contextual evaluations in User-Centered Quality of Experience evaluations.

4.6. Discussion

4.6.1. Convergence and complementation

The four studies highlight the complementation and convergences of the results acquired with different methods and underline the positive features of the Open Profiling of Quality approach. The results, summarized in Table 4.7, complemented each other in all studies. In addition, each study allowed for the explanation of the quantitative quality preferences according to sensory descriptions. For example, when quantitative excellence between stimuli was not identified, the qualitative results showed the detectable differences between the used variables, inferiority nullified the positive influence of quality (audiovisual depth), and the participants' sensorial preferences contributed to final multimodal quality evaluations.

Furthermore, the results suggest an explanation for the excellence of mobile 3D video parameters through understanding the relationship between quality and depth perception. The descriptions of depth and being error free were attached to good quality when visual presentation mode and coding factors were varied. Without qualitative data, the reasons behind the quantitative data had been based on assumptions [152], even though sensory evaluation as a single method cannot be used to explain users' preferences. The results concerning the dependency of video quality and added value from perception of depth confirm previous findings and extend the existing 3D Quality of Experience models with the hierarchical dependency structure. Added value is only experienced when the number of perceivable visual artifacts is low [152]. In addition, the application of OPQ also allowed for better understanding of the impact of user characteristics on 3D quality perception. The studies showed that consumers have different preferences of modalities for deriving audiovisual quality factors. In addition, the impact of quality factors, like large disparity, is different. While for some test participants large disparity results in the perception of crosstalk, for others it increases the 3D experience and leads to high acceptance and quality satisfaction ratings.

4.6.2. Aspects of reliability and validity

The application of Open Profiling of Quality in four studies on MOBILE3DTV underscores different aspects of the reliability and validity of the method. Although the results of the studies are a valuable contribution to quality research on mobile 3D video, the main purpose in applying OPQ was validation of the research method.

Validity The internal validity of OPQ was shown in all studies. The results (or cause-effect relationships) of all studies can be related to the independent variables that were varied in each study. Naïve test participants were able to identify the independent variables, indicating that OPQ is applicable for use with naïve test participants. Threats to internal validity were thoroughly addressed. To prevent sampling bias, test participants in all studies were recruited in accordance with the user requirements for mobile 3D television and video [216]. All test participants were screened for visual acuity, color vision, and ability to see 3D. Equal distribution of test participants by age and gender were observed. Furthermore, instrumentation effects were handled through standardized test conditions [5, 39], thorough reporting of test stimulus production, and extensive training of research personnel to assure similar calibration of instrumentation over time. During analysis, the requirements for statistical methods, for example, normal distribution, were checked to overcome the threat of incorrect applied statistics. In addition, different methods of analysis for the results of sensory evaluations were used, leading to comparable results and conclusions of causal impact of the independent variables on users' ratings [6].

External validity assures that Open Profiling of Quality results can be generalized beyond the experimental context across samples, settings, and time [4, 6]. Within the studies, external validity is highlighted by converging results among complementary research questions, evaluations in different contexts of use, different test devices, and a selection of test participants, variables, and test contents according to previously established user requirements [216]. Although differences in the results occur among the variations, the main OPQ models show good agreement between the conditions and lead to the same cause-effect relationships of independent variables and assessors' quality perceptions. Eventually, construct validity of OPQ can be assessed in accordance with the three-step approach of Carmines and Zeller [183]. The measurements of naturalness and the resulting models of 3D Quality of Experience of Seuntïens [152] are well-established measures for evaluating 3D quality. If construct validity of OPQ exists, then good correlation or, at least, agreement should be found between the models of Seuntïens and the results within this thesis. The studies indicate construct validity of OPQ such that OPQ can identify video quality and depth perception as the strongest impacting factors for 3D Quality of Experience, similar to the findings of Seuntïens [152]. Furthermore, it can be used to extend the knowledge about the rationale with better modeling of dependencies and affective dimensions.

Reliability The results of the different studies with varying research questions confirm the internal reliability of consistency of OPQ results. Although no explicit test-retest study was conducted, the evaluation of the same research question in two different contexts in Study 4 (section 4.5) as a quasi-test-retest application shows good agreement in the joint Hierarchical Multiple Factor Analysis in terms of the identified quality factors and their impact on perceived quality. Further, the results obtained from quantitative and sensory evaluations show good agreement on quality models and preferences towards artifact-free videos in all studies. To achieve internal consistency of Open Profiling of Quality, the thesis introduces a standardized test description and methods for common training of test participants within OPQ studies. The experience of the author and other researchers has shown that such standard training is an important step in consistent application of the method and reliable results. While internal reliability describes the reproduction of empirical findings within OPQ studies, external reliability refers to the equivalence of results in comparison to other evaluation approaches. The quality rationales that OPQ identified to explain quantitative quality preferences are confirmed by other studies on (mobile) 3D video and television [60, 152]. Especially, the dominance of visual artifacts and 3D perception and the bipolar description of quality (blocky - sharp; blurry - visible details) are also described in Jumisko-Pyykkö and Utriainen's interview-based quality component models [60]. Finally, the last point of discussion is the role of interrater reliability in the interpretation of OPQ models. The quality models for Study 4 were therefore analyzed independently by two researchers, the author of this thesis and a master student for her thesis [184]. A comparison of the descriptions of the dimensions of the OPQ models shows good agreement between the raters, although Kappa values [4] are hard to calculate for these interpretations. For example, the following is a juxtaposition of the interpretations of the MFA models for the context data set:

Eulenberg [184]:

- "The first dimension can be connected to attributes of 'good visual quality' on the one end of the axis and with 'bad visual quality' on the other end of the axis." [1]
- "The second dimension can be associated to bad 3D visualization. Thereby, the second dimension of mainly affected by content 'Dracula', and to a minor degree by contents 'The Eye' and 'Makro'." [2]

The author (section 4.5):

- "The first component is mainly described with attributes like 'blocky' or 'artifacts' on its negative polarity, while the positive one correlates strongly with attributes like 'clear', 'sharpness of edges', or '3D effect'. These items describe the differences in the perception of video quality and are in accordance to the separation of QPs along the first dimension."

[1] "Die 1. Dimension kann daher mit den Begriffen 'gute visuelle Qualität' an einem Ende der Achse und 'schlechte visuelle Qualität' am anderen Achsenende verknüpft werden."

[2] "Die 2. Dimension lässt sich mit schlechter 3D-Darstellung assoziieren. Dabei ist die 2. Dimension hauptsächlich durch den Content 'Dracula' geprägt, in geringem Maße aber auch durch die Contents 'The Eye' und 'Makro'."

- "The second dimension is correlating with attributes such as 'double images' on the one, with few attributes like 'color-fast' and 'perceivable as one image' on the other polarity. Having in mind that along this component content Dracula separated from the other content, one can see that there was a problem with getting a proper 3D perception."

4.6.3. Conclusion

This section concludes the first part of the thesis in which the author explored the development of a validated mixed-methods research approach for audiovisual quality evaluations. Finalizing the development process of Open Profiling of Quality, the method was submitted as a proposal for standardization to ITU-T SG12 [185]. "Proposal on open profiling of quality as a mixed method evaluation approach for audiovisual quality assessment" [225] was accepted as a submission for Question 13/12 "QoE, QoS and performance requirements and assessment methods for multimedia including IPTV" [186] and was presented successfully in the general ITU-T SG12 meeting in January 2011.

However, during the work towards standardization of Open Profiling of Quality, some shortcomings and limitations were identified. First, during a series of OPQ studies on a set of related research questions, a great deal of sensory data are collected. These data have individual characteristics in terms of meanings. Interpretation of these idiosyncratic attributes is very difficult for the researcher and depends on definitions given by each assessor [117]. However, for further development, a common vocabulary or quality terminology is desirable because it may significantly shorten the length of studies by eliminating the need for vocabulary development. However, Open Profiling of Quality does not offer tools for creating a common terminology from the individual quality factors. Second, systematic comparisons between OPQ and existing methods are needed to provide guidelines for the effective use of these methods by practitioners. To probe aspects in the comparisons, OPQ can provide a relatively easy data-collection and analysis procedure that does, however, require multiple evaluation sessions. In contrast, interview-based methods require good interviewing skills in personnel and a relatively time-consuming data analysis to be completed in a one-session design. Systematic comparisons are needed to verify performance-related aspects (e.g., accuracy in different quality ranges, validity, reliability, and costs), complexity (e.g., ease of planning, conducting and analyzing, and interpreting results), and evaluation factors (e.g., number of stimuli, knowledge of research personnel) [187–189]. The long-term goal is to support the idea of safe development of these instruments by understanding their benefits and limitations when capturing deeper understanding of experienced multimedia quality. These two aspects will be covered in the following sections within the work on the Extended-OPQ approach (chapter 5) and the development of a holistic comparison model for descriptive research methods (chapter 6).

Summary of the results	Summary of the methodological contribution
Study 1: How does perception of audiovisual content change when it is presented in 2D and 3D?	
Although the results of the psychoperceptual evaluations did not reveal any significant differences between the 2D and 3D conditions, OPQ results show that the independent variables in the test were identified and evaluated. In addition, the results revealed different preferences of participants for one modality from which they derived their quality attributes.	This study was the first study using OPQ. It shows the applicability of OPQ to evaluations by naïve participants. It shows that perceived quality goes beyond technical descriptions and that OPQ can be used to study individual differences in sensory evaluations.
Study 2: What is the optimum coding method for mobile 3D video?	
The results of psychoperceptual evaluation showed that Multiview Coding and Video + Depth provide the highest experienced quality among the tested coding methods. The results of sensory profiling showed that artifacts are still the determining quality factor for 3D. The expected added value through depth perception was rarely mentioned by the test participants. When mentioned, it was connected to the artifact-free video, indicating a hierarchical dependency between depth perception and artifacts. When the visibility of artifacts is low, depth perception seems to contribute to the added value of 3D.	The study shows convergent results of quantitative and sensory results and confirms the methods of External Preference Mapping. Further, the study shows that naïve test participants can evaluate a large test set and a wide quality range with OPQ.
Study 3: What are the optimum transmission conditions for transmitting mobile 3D videos over a DVB-H channel?	
The results show that the provided quality level of videos with a low error rate is clearly above 50%. Still, the different coding methods had the highest impact on the experienced quality of test participants. In the sensory evaluation, the results show that the expected descriptions of judder in contrast to fluency of the test items are found rarely. In contrast, descriptions are dominated by artifacts relating to blockiness or blur. Again, impact of 3D perception was identified only for artifact-free videos. For the system development process of MOBILE3DTV, the study identified MVC as the preferred encoding method for mobile 3D video due to its error robustness.	The study extends OPQ with new methods of analysis to overcome threats of internal and external validity. The results show a clear benefit from these changes because MFA offers more flexibility in analyzing and interpreting the data. The results of the study fix HMFA and PREFMAP as the methods of choice for analyzing sensory data within OPQ.
Study 4: Do sensory profiles change when evaluations are conducted in the actual context of use?	
The sensory profile obtained from the evaluation in the café is comparable to the laboratory evaluation result. Additional impact of crosstalk was identified. However, the results also show clear differences among participants in perceiving this crosstalk.	The study increases the external validity of OPQ by compared evaluations in different contexts of use. Further, the methodological approach was merged into a single-session design to shorten the duration of an OPQ study minimizing the risk of participants' dropping out.

Table 4.7. – Summary of the results of the four presented OPQ studies with respect to answering the research question and making progress in the methodological development of Open Profiling of Quality

5. Extended-OPQ: From individual vocabulary to general components of QoE

In this section, the component model as an extension to Open Profiling of Quality is introduced. The extension allows creating common terminology from a set of individual quality attributes from different OPQ studies. The obtained components of QoE for mobile 3D television are presented, and the use of fixed terminology in comparison to OPQ's original individual vocabulary profiling approach is presented. Parts of this chapter have been published in Strohmeier et al. "The Extended-OPQ method for User-centered Quality of Experience evaluation: A study for mobile 3D video broadcasting over DVB-H," EURASIP Journal on Image and Video Processing, special issue on Quality of Multimedia Experience, vol. 2011, Article ID 538294, 24 pages, 2011, doi:10.1155/2011/538294 [217], Jumisko-Pyykkö et al. "Descriptive Quality of Experience for Mobile 3D Video", in Proc. of the 6th Nordic Conference on Human-Computer Interaction (nordiCHI), Reykjavik, Iceland, 2010, and Kunze et al. "Towards a comparison model for audiovisual quality evaluation methods", submitted to Third International Workshop on Quality of Multimedia Experience (QoMEX), Mechelen, Belgium, 2011.

5.1. Fixed vocabulary and terminologies in descriptive analysis

In contrast to individual descriptive methods, fixed vocabulary approaches evaluate perceived quality based on a predefined set of quality factors. The descriptive evaluation with fixed vocabularies has had a long tradition, and several methods have been introduced and applied successfully on different research questions [40, 63]. In general, this fixed vocabulary (also objective language [190], lexicon [191], terminology [192], or consensus vocabulary [46]) is regarded as a more effective way of communicating research results between the quality evaluators and other parties (e.g., development, marketing) involved in the development process of a product [190] compared to individual quality factors. Such approaches also allow for direct comparison of different studies or easier correlation of results with other data sets like instrumental measures [71].

In general, vocabularies include a list of quality attributes to describe the specific characteristics of the product to which they refer. These quality attributes are usually structured hierarchically into categories or broader classes of descriptors. In addition, vocabularies provide definitions or references for each of the quality attributes [190, 191]. Some terminologies in the field of sensory evaluation have become very popular because it allows for defining a common understanding about underlying quality structures. Popular examples are the wine aroma wheel by Noble et al. [192] and Meilgaard et al.'s beer flavor wheel [193], both of which use the common wheel structure to organize the different quality terms.

A fixed vocabulary in sensory evaluation needs to satisfy different quality aspects that were introduced by Civille and Lawless [71]. Especially, the criteria of discrimination and non-redundancy need to be met so that each quality descriptor has no overlap with another term. In descriptive evaluation methods that apply these vocabularies, a consensus about the meaning of each of the attributes is needed among assessors [40]. While sensory evaluation methods like Texture Profile [194] or Flavour Profile (see [195]) apply vocabularies that have been defined by underlying physical or chemical properties of the product, Quantitative Descriptive Analysis (QDA) (see [63]) makes use of extensive group discussions and training of assessors to develop and sharpen the meaning and consensus of the set of quality factors.

In research related to audiovisual quality evaluations, Bech and Zacharov [46] provide an overview of existing quality attributes obtained in several descriptive analysis studies. Although these attributes show common structures, Bech and Zacharov indicate that they must be regarded as highly application specific, so they cannot be regarded as terminology for audio quality in general. A consensus vocabulary for video quality evaluation was developed in Bech et al.'s RaPID approach [69].

5.2. The component model as extension of the OPQ method

5.2.1. Open definition task and qualitative descriptions

Within a set of OPQ studies in a specific research area, test participants develop a large number of attributes that all relate to their individual descriptions of perceived quality in the specific domain. As descriptive analysis targets a broad evaluation of a specific research area with respect to different research problems [63], these descriptors cover a multifaceted view on experienced quality in this domain. During the development of Open Profiling of Quality, the question arose concerning whether it was possible to develop a common vocabulary from these individual attributes to describe and evaluate the experienced quality of audiovisual 3D media. In fact, OPQ is a suitable approach to investigate and model individually experienced quality factors, but higher level descriptions of these quality factors for communicating the main impacting factors to engineers or designers have been missing.

In a related approach, Samoylenko et al. [196] introduced the Verbal Protocol Analysis method. The goal of this approach was to analyze descriptions of the timbres of musical sounds and place them in a common structure. The approach has three levels of classification. In the first level, Samoylenko et al. classify each of the verbal descriptors according to its 'logical sense', that is, whether it describes similarities or differences between two stimuli. The second phase clusters descriptors according to their 'stimulus relatedness,' which refers to either global or specific descriptions. The third hierarchical level groups descriptors according to 'semantic aspect'. This level differentiates each descriptor either by single features or by a holistic, conceptual description. Overall, ten classifications are made for each descriptor in the three levels. The final result is a classification of each descriptor according to these ten classes. Samoylenko et al. use this classification to switch from a descriptor-related analysis to a more general analysis of results within different groups in the classification. Although this approach is promising for a generalized step of analysis of data obtained from free verbalization tasks, it does not allow for development of general vocabulary that can be used in prospective evaluation studies.

The component model is a qualitative data extension that allows for identifying the main components of Quality of Experience in the OPQ study and organizing them into a logical structure of categories and subcategories. The component model is included in the Extended-OPQ approach [217], which extends OPQ with a fourth step of data analysis using data collected during the OPQ test (Figure 3.1). Within the attribute-refinement task of the sensory evaluation, a free definition task is conducted. The task completes the attribute refinement, and test participants are asked to define each of their idiosyncratic attributes. As in the attribute elicitation, participants are free to use their own words, but definitions must clarify what an attribute means for them or to which aspect of experienced quality it relates. In addition, participants are asked to define minimum and maximum values of sensation for each attribute if possible. My experience has shown that this task is rather simple for the test participants compared to attribute elicitation. After the attribute-refinement task, participants were all able to define their attributes very precisely (Table 5.1).

Attribute	Participant's Definition	Minimum	Maximum
fluent movement	movements and action get blurry and get stuck in the background	movements get very blurry	n/a
image blurred	frames are not layered correctly	image not displaced	image seems to be highly displaced
constant background	background does not change when there is a non-moving image	n/a	colors and outlines do not change at all

Table 5.1. – Examples of attributes and their definitions obtained in the transmission study of Mobile3DTV (Section 4.4)

Collecting definitions of the individual attributes is not new within the existing Free-Choice profiling approaches, and definitions are collected in related methods [117]. However, those definitions have only served to interpret the attributes in the sensory data analysis [214]. In the Extended-OPQ approach, I see these definitions as a second level of descriptions of the experienced quality factors with the help of the free definition task. These descriptions are short (one sentence), well defined, and precise. While the individual attributes are used for sensory analysis, the component model extension finally applies these qualitative descriptors to form a framework of components of Quality of Experience. By applying the principles of Grounded Theory framework [66] through systematic steps of open coding, concept development and categorizing, researchers develop a descriptive Quality of Experience framework that shows the underlying main components of QoE in relation to the developed individual quality factors. Comparable approaches have been used in the interview-based mixed-methods approaches also included in the UC-QoE evaluation framework [9, 58]. This similarity makes it possible to directly compare (and combine) the outcomes of the different methods into a joint model.

5.2.2. Components of Quality of Experience for mobile 3D video

From the data sets obtained in the evaluations of mobile 3D television, the chosen three studies that represented a large variety of research problems [226]. The characteristics of these studies are summarized in Table 5.2.

Experiment and Research problem	Experiment Variables	Stimuli Characteristics
Experiment 1: [214] 2D-3D COMPARISON Sample size: 15	Video: presentation mode (2D/3D) Audio: presentation mode (mono/stereo) Content: 6 contents	Length: $\sim$ 18 s Videos: Synthetic and Natural Presentation mode: 2D and 3D Quality level: Highly acceptable Video: mp4v, 10-22Mbit/s, 25fps Audio: WMA 9, 48kHz 16bit
Experiment 2: [214] 3D CODING METHODS Sample size: 15	Video: 4 coding schemes 2 quality levels(low: 74-160kbps bit rate, high: 160-452kbps); Content: 6 contents	Length: $\sim$ 10 s Videos: Synthetic and natural Presentation mode: 3D Quality level: Highly acceptable Video: H.264/AVC (JMVC 5.0.5) Audio: none

Experiment 3: [217]
3D DVB-H TRANSMISSION
Sample size: 17
	Video: 3 coding schemes @ slice and nonslice mode
	2 MFER rates (10%, 20%);
	Audio: clean audio;
	Content: 4 contents
		Length: $\sim$ 60 s
		Videos: Synthetic and natural
		Presentation mode: 3D
		Quality level: Highly acceptable
		Video: H.264/AVC (JM 14.2), MVC (JMVC 5.0.5)
		Audio: WMA 9 11/44.1 kHz

Table 5.2. – Characteristics of the experiments chosen for development of the QoE component model

For each of these studies, test participants developed a set of individual definitions in the free-definition task at the end of OPQ’s attribute-refinement task. These definitions were taken as independent descriptive data sets for experienced quality and analyzed following the concept of data-driven framework in accordance to the principles of Grounded Theory [66] and the instructions given by Jumisko-Pyykkö [9]:

1. **Open coding towards concepts**: Usually, this steps starts by extracting meaningful pieces of data from the transcribed data sets. In the analysis of the free definition data, each definition can be treated directly as codes in the analysis as the definitions are short, well-defined, and precise in comparison to, for example, interview data. From these codes, concepts and their properties are identified.
2. **Categorization**: All concepts developed are further categorized into major categories and probably subcategories

3. **Frequencies of mention**: The frequency in each category is determined by counting the number of participants who mentioned it. Several mentions of the same concept by the same participant are counted just once.
4. **Interrater reliability**: A second researcher performs coding and categorization for a randomly selected 20% of each data set, and interrater reliability is calculated using Cohen's Kappa [4].

The results of the data-driven analysis of the free-definition task data shows that, in general, experienced quality for mobile 3DTV transmission is constructed from components of visual quality (depth, spatial, temporal), viewing experience, content, audio, and audiovisual quality (Table 5.3).

The component model provides converging results to the results obtained in the sensory evaluations in terms of components and their importance. The most important category of the component model obtained is the visual quality, confirming the findings of the sensory analysis. Although the weighting of its subcomponents spatial, temporal, and depth differs among the different studies, the overall findings show that especially the artifact-free perception of the video (clarity, fluency, excellence of 3D) determines participants' components of quality. In addition, the model shows that test participants often use complementary descriptions of quality that lead to contrary subcategories comparable to descriptions along dimensions in the sensory results. Thus, either visual spatial quality is described positively in terms of detection of objects and their details, or the same effect is described negatively due to different structural imperfections such as blocking impairments and visible pixels. This juxtaposition can also be identified for other components, for example, fluent motion as opposed to influent motion or eye strain as opposed to ease of viewing. Finally, it is remarkable that the results of the component framework analysis confirm the findings of the sensory evaluations for audio and audiovisual components. While in the sensory analysis one could still argue that the audio-related components are simply overwhelmed by the high impact of visual components, the model shows that few attributes are developed in relation to audio and audiovisual quality components. This finding confirms the sensory results. However, the inclusion of audio and audiovisual as separate components is important for the holistic view of the developed component model.

Within the work in the UC-QoE framework development, the component model was used in a joint analysis with qualitative data obtained through interviews in contextual studies [58]. The comparable characteristics of the data and the resulting separate component models allowed combining the different data sets into one descriptive Quality of Experience model for mobile 3D video. The results presented by Jumisko-Pyykkö et al. [226] confirm the results of the OPQ component model and generalize the model. Table 5.4 lists the final components of QoE for mobile 3D video [226]. Especially by the contextual data, new emphasis is placed on context-dependent components within the category of Viewing Experience. Overall, the developed joint results present a general descriptive model of QoE for mobile 3D media. Jumisko-Pyykkö et al. [226] conclude that important steps for further work on the descriptive model are related to validation and operationalization. Following, I present the results of a comparison study between Open Profiling of Quality and a newly introduced method called Conventional Profiling [226], which the Free-Choice Profiling task is substituted for by using the QoE component model as fixed vocabulary for sensory evaluations.

5.3. Study 5: Comparison of the perceptual model of OPQ and Conventional Profiling

Operationalization of the developed QoE components can be accomplished through several methods. In this study, the components were used as fixed attributes in a conventional profiling approach. For validation of results, the laboratory part of the OPQ study presented in section 4.5 was repeated in the use of conventional profiling.

5.3.1. Research method

5.3.1.1. Test participants

Sixty-three test participants took part in the study. All test participants were screened for normal or corrected-to-normal vision, color vision, and 3D vision. All test participants can be classified as naïve assessors because they had experience neither in the domain of research nor in subjective quality evaluation studies. In the study, each test participant passed the psychoperceptual evaluation. For the qualitative part of the study, 15 randomly selected participants were assigned to Conventional Profiling (CP) and 16 participants to OPQ.

5.3.1.2. Test stimuli

Six different contents and three different video qualities (QPs) were used. The stimuli of the study were the same as those produced and applied in the contextual validation study of OPQ. Description of contents and production of the variables is reported in section 4.5.

5.3.1.3. Stimulus presentation

The tests were conducted in a laboratory at Ilmenau University of Technology, and test conditions were arranged according to the specifications in ITU-T P.910 [5]. As a playback device, a digital Viewer FinePix REAL 3D V1 from FUJIFILM was used with a resolution of 640x480 pixels for the 3D videos. The viewing distance was set to 50 cm initially, but test participants were allowed to adjust their viewing distance for the best stereoscopic experience. The integrated loudspeakers of the FinePix V1 were used for audio playback due to a missing headphone connection. According to the speakers' maximum sampling rate, audio was represented with a sampling rate of 11 kHz. Different playlists in pseudo-randomized orders were used for video presentation. During the psychoperceptual evaluation, each test item was presented twice. In the OPQ and CP approaches, each video was presented once and again at the request of the test participant.

5.3.1.4. Test procedure

Psychoperceptual evaluation and Open Profiling of Quality This study is based on the data set obtained in Study 4. Psychoperceptual evaluation and OPQ are reported in section 4.5.

Conventional Profiling Conventional Profiling (CP) [226] is an adaptation of OPQ with an evaluation of quality on fixed vocabularies comparable to QDA (see section 2.4.1). In contrast to classic consensus vocabulary methods, we did not conduct any group discussions for attribute elicitation but used the developed component model of QoE for 3D mobile media. At the beginning of the conventional profiling, test participants received a list with the quality components and their descriptions (see table 5.4). Test participants first read this list and were asked to become familiar with the attributes. They were allowed to ask questions for clarification, but no references were provided by the researcher. In the second step, comparable to OPQ's attribute refinement, a score card on which each attribute was displayed with a 10 cm line labeled with 'min' and 'max' was provided (see section 3.3.1.3). In a training task, assessors watched and rated a subset of 6 videos. Finally, a sensory evaluation of the whole data set was conducted, and all items were rated on the fixed attributes. The conventional profiling was conducted after a 5 min break following the psychoperceptual evaluation.

5.3.1.5. Methods of Analysis

For the psychoperceptual evaluation nonparametric methods (Kolmogorov-Smirnov: $p < .05$) and for OPQ, a Multiple Factor Analysis was used (see section 4.5). The data of the conventional profiling were first transformed into quantitative measures by measuring the distance from the 'min' to the participant's rating for each attribute and test item. To maintain comparability with the results of OPQ and to take into account individual differences of the test participants, a Multiple Factor Analysis was applied to the CP data set. Although the common method for analyzing consensus data is to perform a PCA on the mean over participants, common individual profiling methods like GPA or MFA have been applied because of their ability to display individual differences in the use of the consensus vocabulary [132, 133, 135]. Finally, a Hierarchical Multiple Factor Analysis was conducted for joint analytical comparison of the OPQ and CP data sets.

5.3.2. Results

5.3.2.1. Psychoperceptual evaluation

Acceptance of overall quality The presented stimuli reached an acceptance level of 52.8% in total. Items with qp30 were accepted with a minimum of 82.5% and with 94.2% over all contents. Items with qp40 reached an acceptance level of 52.4% over all contents, whereas items with qp45 were not acceptable at all (11.9%)(Figure 5.1).

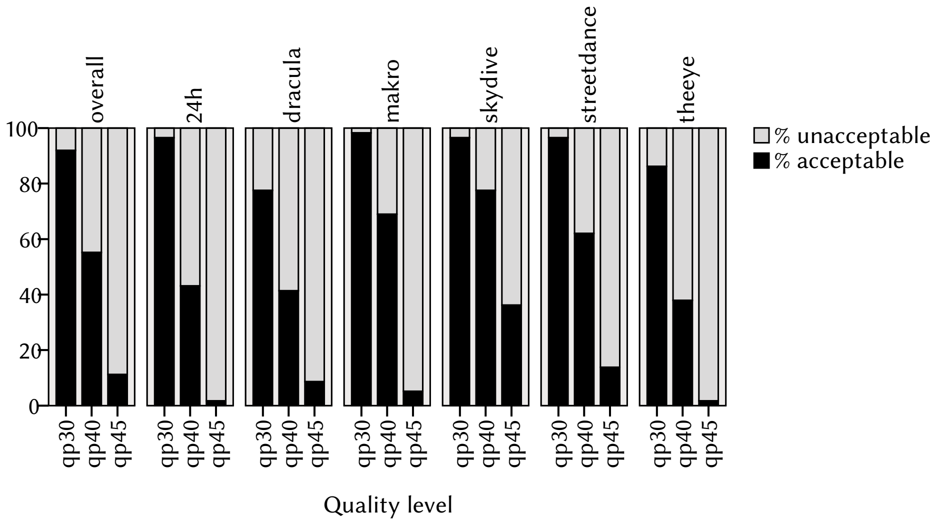


Figure 5.1. – Overall acceptance ratings averaged over contents as well as by individual content.

Satisfaction with overall quality Analysis of the distributions of OPQ and CP data set did not reveal significant differences, so the data were analyzed jointly (Mann-Whitney U: all comparisons: p >.05). The coding quality parameters influenced the overall quality perception when averaged over all contents (F_r = 743.433, df= 2, p < .001) as well as by individual content (all comparisons: p < .001). Figure 5.2 shows the mean satisfaction scores averaged over all participants for the different contents and quality parameters and for the separate contents tested. Videos with qp30 provided the most satisfying quality (qp30 vs. qp40: Z = -17.021, p < .001). The qp45 quality level received the worst ratings (qp40 vs. qp45: Z = -15.832, p < .001). This finding is also valid for a content-wise analysis (all comparisons: p < .001). The results also show differences in the ratings of the different contents at the same quality levels (all comparisons: p < .001). Contents 24h and Makro received the highest ratings for qp30. The eye and Dracula were worst rated for the highest quality level (Dracula vs. Theeye: Z = -1.692, p = .091, ns). At the low quality level, skydive was rated best (all comparisons: p < .001). Dracula, 24h, and the eye were worst rated for the low quality level (Dracula vs. 24h vs. the eye: all comparisons: p > .05, ns; vs. other contents: all comparisons: p < .05).

5.3.2.2. Sensory evaluation

The results of the OPQ are reported in detail in section 4.5. For comparison, the loadings of the test items on the MFA components of the OPQ data set are shown again in figure 5.3 (black labels). Summarizing the results, the MFA of the OPQ data revealed a perceptual model of two dimensions that account for 44.25% explained variance. The first dimension was identified as descriptions of video quality while the second component correlated with attributes that referred to a proper 3D perception.

For the CP data set, we also calculated the MFA over all participants. The CP model resulting from the MFA accounts for 53.46% explained variance (dimension 1: 44.04% and dimension 2: 9.42%). The

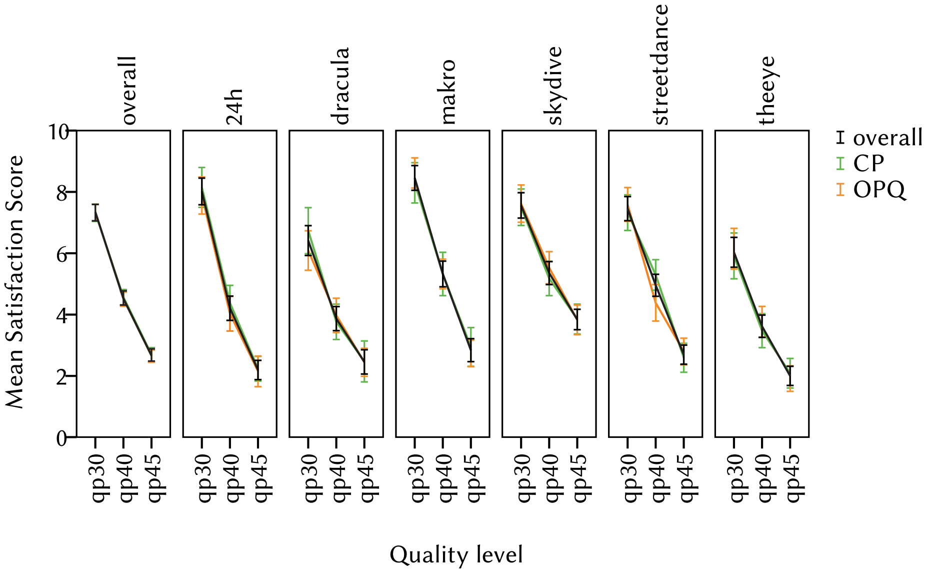


Figure 5.2. – Mean satisfaction scores of the psychoperceptual rating given for the joint data set as well as for the individual results of psychoperceptual evaluation within OPQ and CP.

item plot (Figure 5.3; red labels) shows that the test items separate along the first component according to the different QPs. Along the second component, especially the items of content Dracula separate from the other test items. For the sake of clarity, the author averaged the resulting correlations for all 19 attributes over the participants and presents the averaged attribute correlations in Figure 5.4. While all these attributes show high correlation with dimension 1 and low correlation with dimension 2, the non-averaged results (Figure 5.4; gray arrows) also reveal high correlation of some attributes with dimension 2. For dimension 1, the highest correlation is given for attributes like 'clarity of motion'; 'objects and edges'; 'color, brightness, and contrast'; and 'clarity of image'. For dimension 2, the correlation plot shows that the attributes with high correlation differ and fall into classes like 'ease of viewing', 'fluency of motion', or 'perceivable depth'. Further, MFA calculated the correlations of each individual PCA result with the overall MFA model (Figure 5.5). For all participants, the dimension 1 ('F1') correlates with MFA's Dim1 and many of the individual 'F2's with the MFA's Dim2. Thus, the structure of the individual data sets are quite comparable. However, no correlation of the overall attributes with Dim2 can be found within the averaged MFA results. The patterning of the individual attributes (gray arrows in Figure 5.4) suggests that participants may have understood and used the attributes in different ways and that the used attributes may not have been adequate for describing all perceived quality characteristics.

5.3.2.3. Comparison of OPQ and CP data sets

Both data sets show comparable impact on the HMFA model (Figure 5.6). The model shows the loadings of each test item in the HMFA model as well as the partial plots of each hierarchy level, that is, OPQ and CP. The separation of test items along the first dimension is in accordance with

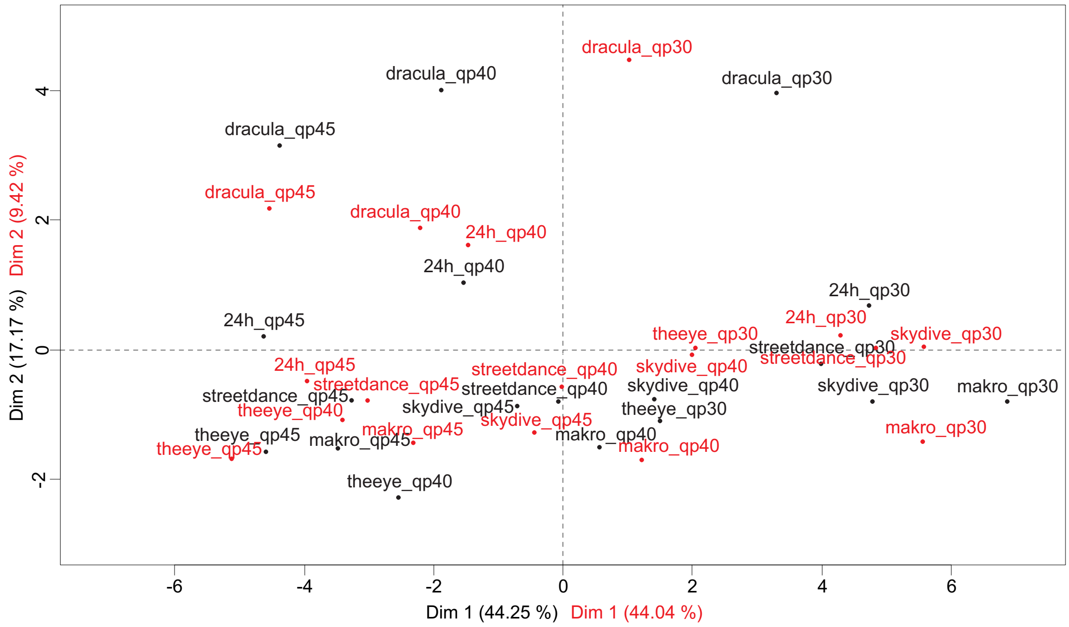


Figure 5.3. – Item plots of the MFAs on OPQ (black) and CP (red) data sets.

the QPs applied. Along the second component, the Dracula items separate from the rest of the test items. These findings confirm the individual results of OPQ and CP, which can also be seen in a comparable explained variance of 54.19% of the HMFA model. Further analysis of the partial plots shows that the deviation along the first dimension between the two data sets is very low. Along the second component, higher deviation can be found. OPQ results tend more towards the extremes for most items, with loadings on dimension two (e.g. dracula_qp40, theeye_qp45, makro_qp45), than the CP results.

5.3.3. Discussion

The purpose of this study was twofold. First, I wanted to show an initial approach to operationalizing the vocabulary obtained from the application of the Extended-OPQ methods, especially the component model, on the data sets from quality evaluations on mobile 3D video. The presented Conventional Profiling approach adapts the method of OPQ and replaces the individual vocabulary evaluation with an evaluation with fixed vocabulary. Second, the study targeted the validation of the descriptive QoE component model for mobile 3D video. This study presents the comparison of the Conventional Profiling approach with an Open Profiling of Quality.

The results of the study confirm the intended goals and show that both data sets describe the test items in a comparable way. In general, the results of OPQ and CP are very similar in terms of dimensions as well as the strength of the two models in terms of explained variances. CP and OPQ both indicate the importance of the video quality as the most important quality factor. A description of the individual results as well as the joint analysis of the data sets in the HMFA shows

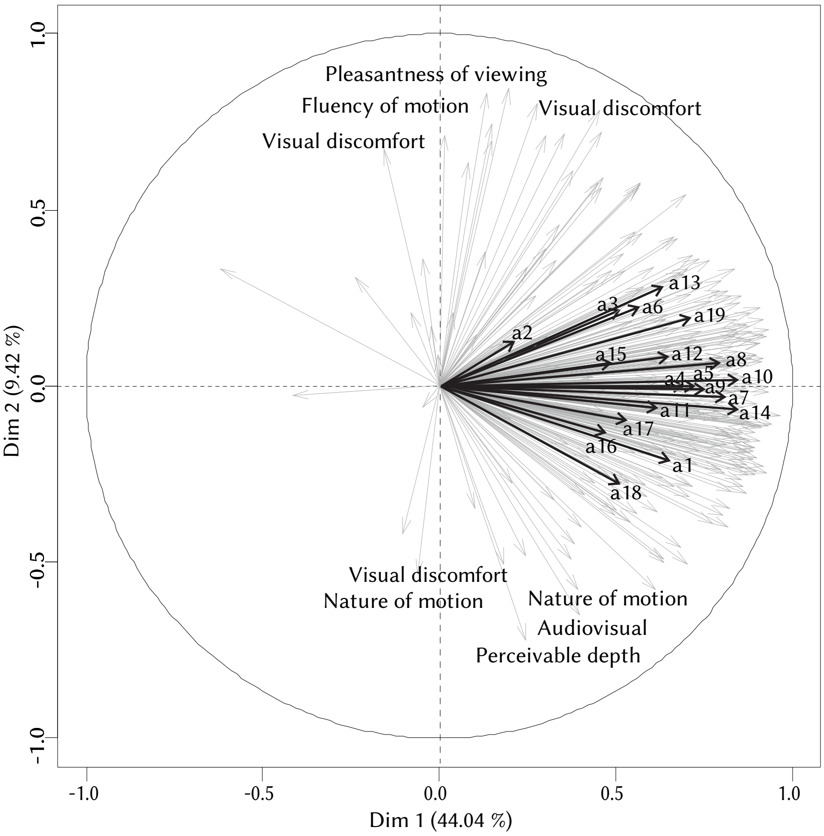


Figure 5.4. – Correlation plot of averaged MFA correlations over all participants (black) as well as individual correlations (gray) of exemplary attributes with Dimension 2 (a1: perceivable depth; a2: impression of depth; a3: fore/background layers; a4: balance of fore/background quality; a5: clarity of image; a6: block-free image; a7: color, brightness, and contrast; a8: objects and edges; a9: fluency of motion; a10: clarity of motion; a11: nature of motion; a12: ease of viewing; a13: pleasantness of viewing; a14: enhanced immersion; a15: visual discomfort; a16: comparison to existing technologies; a17: overall quality; a18: audio; a19: audiovisual)

only low deviation along this dimension between the data obtained from OPQ and CP. Although the separation of test items between the two methods is also comparable along dimension 2, there exist larger deviations between the data sets. The results of the individual MFA analyses show a large difference between the explained variance of dimension 2 for OPQ (17.17%) and CP (9.42%). The analysis of the correlation plot of the CP reveals that there is still an inconsistency in descriptors used to describe the second dimension, explaining the low percentage of explained variance for this dimension. This inconsistency also explains why there is no clear correlation of attributes in the averaged data with dimension 2 (Figure 5.4). Although test participants were able to describe the second component as an important quality factor, they applied different attributes to rate the quality of crosstalk and double images, which are covered by dimension 2. In this case, the agreement in the individual descriptions in the OPQ approach is higher, and the importance of the impact of double images is better emphasized.

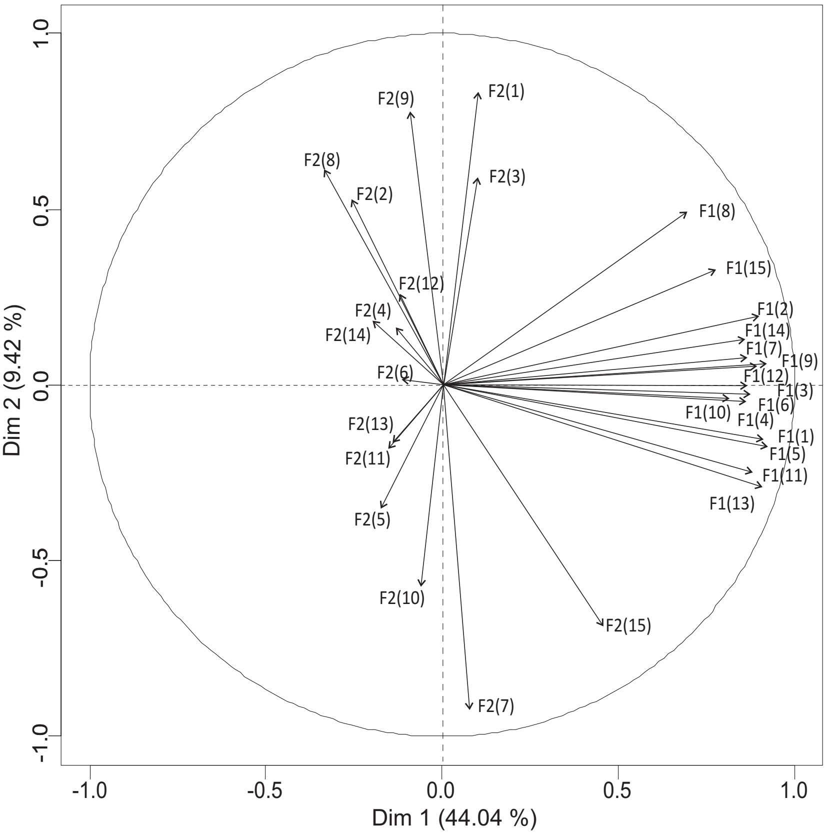


Figure 5.5. – Partial correlations of the individuals' PCAs and the result of the MFA. 'F1' and 'F2' mark the PCA components of the individual PCAs. The numbers in brackets mark individuals' IDs during the test.

The results of the Conventional Profiling suggest that further adaptations and refinement are needed for the attributes. Originally, sensory analysis based on fixed vocabulary spends a lot of time on creating a consensus among test participants on the different attributes. These missing discussions can be identified as a weakness of the Ext-OPQ approach and further operationalization of its created terminologies. However, the results show very good agreement along the first dimension of the CP model and good separation of the test items along the second component. Further work needs to address the refinement of attributes that can be achieved by decreasing the number of attributes in the vocabulary and by redefining and better specifying the definitions per attribute.

5.4. Conclusions

The component model of the Extended-OPQ method describes an approach to develop a fixed vocabulary of descriptive components for Quality of Experience from a series of OPQ tests. As outlined above, sensory analysis targets the global evaluation of a field of research that includes studies on variant research questions within the domain (see section 4). From these studies, many attributes and their definitions are obtained and can be integrated into a terminology of components of QoE within the Ext-OPQ approach. Because the data analysis of the free definitions in the Ext-OPQ ap-

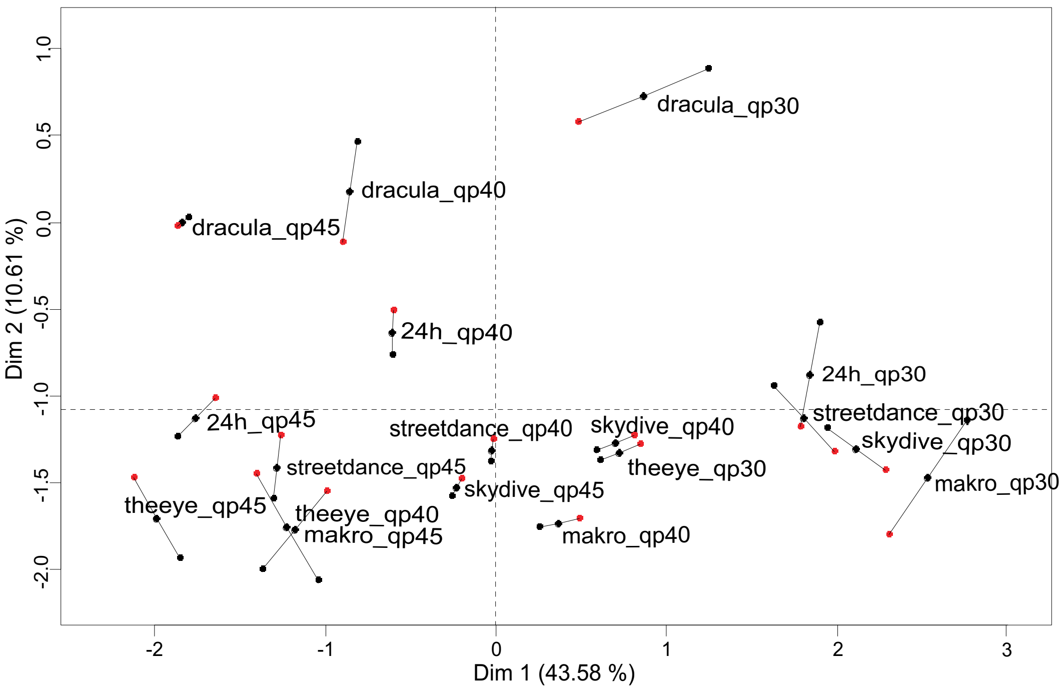


Figure 5.6. – Result of the Hierarchical Multiple Factor Analysis on the joint data set of OPQ and CP. The partial plots show the impact of the individual data sets.

proach is compliant with the interview-based descriptive methods within the User-centered Quality of Experience evaluation framework, the models can be validated in a first step parallel analysis of different data sets.

However, terms developed in the Ext-OPQ need operationalization to make them and the Ext-OPQ approach a valuable tool within descriptive evaluations. By applying the terminology in an adapted approach of the conventional profiling method [226], the author presented a first approach for creating new tools for quality evaluations based on the vocabulary. Further operationalization will increase the validation of the terminology and can create a basis for communication about QoE within system development. Thus, the Extended-OPQ approach with its component model offers researchers a tool to reuse the individual attributes and descriptions from OPQ studies.

Components (major and sub)	Definition (examples)	Exp1 in %	Exp2 in %	Exp3 in %
	Total number of attributes developed	130	129	128
	Interrater reliability; Cohen's Kappa	0.8	0.8	0.8
VISUAL SPATIAL	**Descriptions of spatial video quality factors**			
Clarity	Good spatial quality (clarity, sharpness, accuracy, visibility, error-free)	73.3	80.0	76.5
Color	Colors in general, their intensity, hue, and contrast	66.7	100.0	52.9
Brightness	Brightness and contrast	26.7	80.0	17.6
Blurry	Blurry, inaccurate, not sharp	46.7	40.0	47.1
Visible pixels	Impairments with visible structure (e.g. blockiness, graininess, pixels)	33.3	73.3	70.6
Detection of objects	Ability to detect details, their edges, outlines	73.3	80.0	47.1
VISUAL TEMPORAL	**Descriptions of temporal video quality factors**			
Motion in general	General descriptions of motion in the content or camera movement	26.7	53.3	29.4
Fluent motion	Good temporal quality (fluency, dynamic, natural movements)		60.0	52.9
Influent motion	Impairments in temporal quality (cut-offs, stops, jerky motion, judder)	6.7	40.0	88.2
Blurry motion	Experience of blurred motion under the fast motion	20.0	6.7	17.6
VISUAL DEPTH	**Descriptions of depth in video**			
3D effect in general	General descriptions of a perceived 3D effect and its detectability	86.7	80.0	58.8
Excellence of 3D effect	Artificial, strange, erroneous 3D descriptions (too much depth, flat planes)	66.7	6.7	
Layered 3D	Depth is described having multiple layers or structure	26.7	33.3	23.5
Foreground	Foreground related descriptions	46.7	26.7	17.6
Background	Background related descriptions	33.3	66.7	35.3
VIEWING EXPERIENCE	**User's high level constructs of experienced quality**			
Eye strain	Feeling of discomfort in the eyes	20.0	20.0	35.5
Ease of viewing	Ease of concentration, focusing on viewing, free from interruptions	40.0	6.7	52.9
Interest in content	Interest in watching the content	40.0	13.3	11.8
3D Added value	Added value of the 3D effect (advantage over current system, fun, worth of seeing, touchable, involving)	53.3	33.3	17.6
Overall quality	Experience of quality as a whole without emphasizing one certain factor	20.0	40.0	11.8
CONTENT	**Content and content dependent descriptions**	13.3	6.7	17.6
AUDIO	**Mentions of audio and its excellence**	13.3		11.8
AUDIOVISUAL	**Audiovisual quality (synchronism and fitness between media)**			29.4

Table 5.3. – Components of Quality of Experience, their definitions and percentage of participants' attributes in this category per study

Components (Major and Subcategories) Bipolar impressions	Definition (examples)
VISUAL QUALITY: Descriptions of quality of visual modality, divided into depth, spatial and motion quality	
Depth	*Descriptions of depth quality in video, characterized by perceivable depth, its natural impression, composition of foreground and background layers, and balance of their quality*
Perceivable depth	Ability to detect depth or variable amount of depth as a part of presentation
Impression of depth	3D effect creates a natural, realistic and error-free impression instead of an artificial and erroneous impression (e.g. too much depth, double objects, shadows, seeing through objects)
Foreground-background layers	Depth is composed of foreground and background layers and the impression of the transitions between these layers can vary from smooth to distinguishable separate layers
Balance of foreground-background quality	Balance between the excellence of foreground and background of image quality (e.g. sharp foreground, blurry background or vice versa, or they are otherwise not in balance)
Spatial	*Descriptions of spatial image quality of video, characterized by clarity, block-freeness, colors, brightness, contrast and ability to detect objects and edges*
Clarity of image	Clarity of image overall – Clear (synonyms: sharpness, accuracy, visibility) vs. unclear (synonyms: blur, inaccurate, not sharp)
Block-free image	Existence of impairments with visible structure in image (e.g. blockiness, graininess, pixels)
Color, brightness and contrast	Excellence of colors, brightness and contrast
Objects and edges	Ability to detect necessary objects and details, their edges and outlines
Motion	*Descriptions of motion of video, characterized by fluency, clarity and nature of motion*
Fluency of motion	Excellence of natural fluency of motion – Fluent (dynamic, natural) vs. influent (cut-offs, stops, jerky)
Clarity of motion	Excellence of clarity of motion (e.g. accuracy under fast movement or movement out of screen) – Clear, sharp vs. blurred, pixilated
Nature of motion	Nature of motion in the content or camera movements - Static (synonym: slow) vs. dynamic (synonym: fast)
VIEWING EXPERIENCE: Descriptions of viewing experience, characterized by ease and pleasantness of viewing, enhanced immersion in it, visual discomfort and impression of improved technology and overall quality	
Ease of viewing	Easy to concentrate on viewing (e.g. free from extra effort and learning, viewing angle does not interrupt viewing)
Pleasantness of viewing	Pleasurable viewing experience, also for a longer period of time (e.g. 15min)
Enhanced immersion	Feeling of enhanced immersion into the viewing experience (impression of becoming a part of the events in the content, involvement, fun and improved impression of naturality, like-likeness, tangibility and realism)
Visual discomfort	Feeling of visual discomfort (eye-strain) and descriptions of related discomfort symptoms (headache, general discomfort)
Comparison to existing technology	Impression that provided quality of new technology (3D) is higher than quality of comparable existing technology (e.g. 2D video on a mobile device)
Overall quality	Impression of excellence of quality as a whole without emphasizing a certain factor (e.g. excellence over the time, relation between erroneous/error-free)
CONTENT: Descriptions of content, their content dependency, and interests in viewing content	
OTHER MODALITIES INTERACTIONS: Descriptions of quality of audio modality and interaction between quality of audio and visual modalities	
Audio	Audio and its excellence
Audiovisual	Bimodal audiovisual quality (synchronism and fitness between media) and its excellence

Table 5.4. – Components of Quality of Experience for 3D video on mobile devices and their definitions

6. Performance of OPQ in comparison to related research methods

In this final research chapter, I compare Open Profiling of Quality to other descriptive research approaches in the domain of video quality evaluations. The comparison of the method is a very important final step towards a well-validated research method. For a systematic comparison of descriptive research methods, a comparison model for holistic understanding of differences and similarities of approaches is developed. I apply this framework in two studies in which I compare OPQ to interview-based approaches and free-sorting methods. Parts of this chapter have been published in Kunze et al. "Comparison of Two Mixed Methods Approaches for Multimodal Quality Evaluations: Open Profiling of Quality and Conventional Profiling", Proc. of the Third International Workshop on Quality of Multimedia Experience (QoMEX 2011), Mechelen, Belgium, 2011 [227].

6.1. Towards a model for systematic comparison of research methods

6.1.1. Comparison criteria for experimental research methods

The systematic comparison of different research approaches is important to selecting a proper research method for a specific research problem. In addition, it is a key aspect in the methodological work on new research approaches. The principle of parsimony as one principle of good experimental research only allows creating new methods if the need for new approaches is clearly identified. Thus, Open Profiling of Quality needs to be compared to related research methods so as to describe the purpose, appropriateness, and abilities of each of the approaches.

Comparisons have been made for different psychoperceptual quality evaluation methods in the ITU recommendations [5, 39]. In those like ITU Recommendation ITU-T P.910 [5], different research methods are described, and short guidelines are offered for purpose-directed selection of the appropriate method. Within these guidelines, mostly stimulus-related factors, for example, perceivable quality range or discrimination power, are taken into account to direct the selection process. Similar approaches can be found in the juxtaposition of different sensory evaluation methods in food sciences [40, 195]. Here, the offered guidelines are oriented along three main criteria in a research-problem-related approach and differentiated in accordance with the "three primary questions about products": 1) acceptability, 2) sensory analysis, and 3) the nature of differences [40]. Although these two approaches provide guidelines for key aspects in the comparison of research methods, there is the need for further comparison attributes as questions about issues like detail of results, costs, or needed research personnel are not yet addressed. However, further comparison criteria can be identified from other fields of research beyond the domain of quality evaluations.

The most general comparison criteria are described in the social sciences. In that field, performance indices are well-established tools to measure differences between methods in terms of their degree of scientific nature. These criteria are primarily validity and reliability [2, 3] (see section 2.1), but generalizability, replication, and objectivity are also found to be important criteria in related research approaches. Especially for qualitative research, Lincoln and Guba [197] summarize these criteria in their concept of trustworthiness. Trustworthiness, in general, concerns the real value of a method or study for the audience for which it is intended. It assesses this value in terms of such criteria as underlying research questions, applicability of the method in other domains, consistency of results among different studies, and minimization of any bias effects [197].

Studies on usability extend the criteria of validity and reliability with other performance-related criteria like effectiveness, efficiency, and robustness related to economical aspects [198]. In addition, Markopoulos and Bekker [198] list criteria for describing different usability methods: purpose of the test, the artifact tested, the interaction tasks, participants, facilitator, environment/context, procedure, capture of data, and the characteristics of the test participants. Other studies on comparison of usability tests extend the definitions of effectiveness and describe it in terms of cost effectiveness and effectiveness of results (number of usability problems identified) [188, 199].

In the food sciences also some effort has been put forth to compare different sensory evaluation methods. While many of the comparisons focus on the pure juxtaposition of the results [82, 84, 141], McTigue et al. [187] describe a set of requirements for holistic comparison of descriptive methods. In a comparison of four descriptive analysis approaches, they applied the following criteria: subject selection, number of subjects, training, samples evaluated, replications, method of measurement, analysis of data, outcome, and professional personnel [187]. In addition to results, their comparison criteria describe similarities and differences in terms of requirements for time, test items, personnel, and need for technical equipment. The importance of including test personnel and technical equipment in a systematic comparison model was also found by Yokum and Armstrong [189]. In a comparison of several forecasting methods, implementation-related criteria like ease of use, ease of interpretation, and cost/time were rated as very important to the overall comparison of validity, reliability, and objectivity. Stecher et al. [200] found similar criteria when comparing assessments in vocational education.

6.1.2. A holistic comparison model for systematic comparison

The comparison model organizes the identified comparison attributes from the literature review into main classes and subclasses. The model as a schematic description of a theory accounts for its known properties and targets further study of its characteristics in different applications [201]. So the second goal towards a holistic comparison model is its operationalization based on selected comparison criteria and appropriate measures per criterion.

6.1.2.1. Modeling

Targeting a systematic comparison of research methods, the comparison model classifies each of the identified comparison classes. The structure of the model goes from a particular criterion to more general categories identified during the development process (Figure 6.1). This structure is beneficial in the comparison of methods because comparison and description of research methods can be accomplished based on either the general categories or particularly selected criteria of each group for more detailed comparisons. The model consists of classification of the comparison criteria according to four key aspects of method descriptions identified during the literature review: excellence, economy, implementation, and assessment.

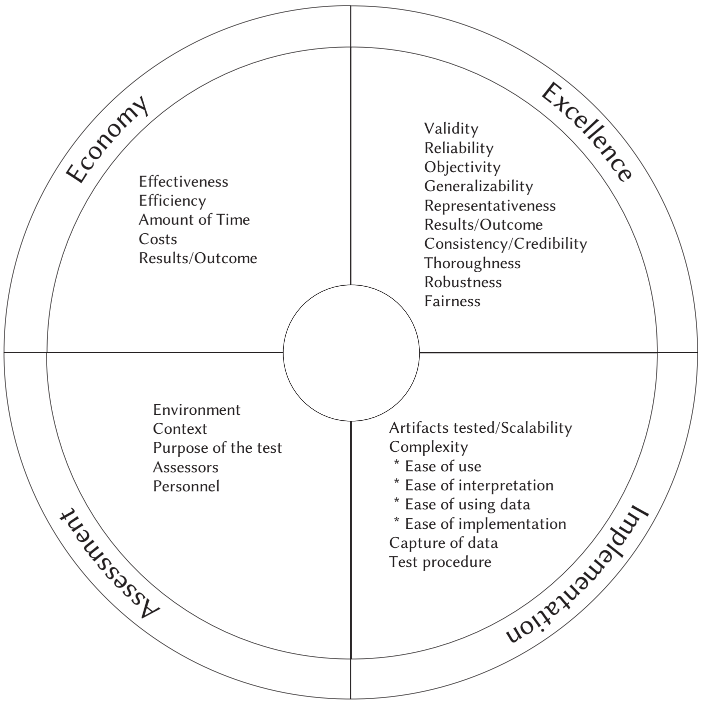


Figure 6.1. – Holistic comparison model containing four main categories of Economy, Excellence, Assessment, and Implementation as well as the criteria per class.

Excellence Excellence-related comparison criteria measure the quality of a test. This class refers to the principles of good experimental research. Beyond comparison, each test method should fulfill these criteria to be a useable research method. The excellence criteria themselves are hard to measure for an individual method and have to be discussed with respect to the whole related field of research. The criteria validity and reliability are known as quality criteria in the social sciences [3]. Validity is related "to our confidence that a given finding shows what it purports to show" [2]. Furthermore, a valid finding has been logically and correctly interpreted. In the literature, validity is discussed in terms different aspects, of which internal, external, construct, and content validity are among the most prevalent (see section 2.1). The general practice for measuring validity is a careful and thoughtful test design. During the test design process, it is important to consider possible threats to validity such as history effects, sampling bias, or confounding variables, which require thorough consideration in the test development [4]. Reliability as another very important test quality criterion can also be differentiated into different types: internal, interrater, and external reliability (or stability) [3, 4]. Coolican [4] describes how reliability is measured by correlation coefficients. A correlation coefficient above 0.75/0.8 is considered to represent good reliability. In addition to validity and reliability, further excellence-related criteria are included in the comparison model (Figure 6.1). One best practice for a good test design is to describe such criteria carefully in the test-development process, discussing and interpreting them.

Implementation Implementation-related criteria concern the implementation of a test and offer a very practical basis for comparison of research methods. A detailed description of the test procedure is crucial to replicating a test, so the implementation-related criteria are connected to the class of excellence-related criteria. Implementation-related criteria address characteristics of items being tested, as well as general descriptions about the complexity of conducting the test. Test stimuli can be compared in terms of number of parameters tested, assessed quality range, or very generally, the number of items being tested. The complexity of a test then extends from stimuli to general test complexity. Important criteria are ease of implementation, ease of application of the method, ease of using the data, and ease of interpreting the results. There is a strong link between the aspects of complexity and the class of economy-related criteria.

Assessment Assessment-related criteria concern the global characteristics of the test. They extend the aspects of implementation-related criteria with criteria beyond the test procedure. Thus, a key question is the purpose of the test applied within a research method to answer a specific research question. Criteria concerning the context of the test and appropriate evaluation environment, test participants, and the characteristics of the chosen sample, as well as the personnel, are relevant for this category.

Economy This category comprises criteria that measure the economic potential of a method and is based on the comparison aspects introduced in the field of usability [198]. While assessment-related criteria compare methods based on their scientific values in terms of results and obtained knowledge with respect to the research question, economy-related criteria evaluate methods with respect to working efforts and costs. Thus, the amount of time and the costs related to the results and efficiency of a method can be estimated from that juxtaposition. Furthermore, effectiveness assesses the performance of a method, its completeness and accuracy, and the achievement of its desired goals. Time and costs of a method have to be measured with regard to the test results to compare the efficiency of methods [188, 198, 199].

6.1.2.2. Operationalization

Operationalization of the comparison model allows for creating a unified tool for holistic research methods comparison among studies and applications. To make the model applicable for researchers and to assure reliability and validity of the comparison results of different studies, the tool provides defined criteria and appropriate measures per criterion. In the current approach, the author restricts the operationalization of comparison attributes to a selected set of attributes. In this approach, comparison attributes are classified as primary, secondary, and general criteria based on the possibility of measuring them during a test (Table 6.1). Primary criteria are directly measurable during the application of a method. Examples of primary criteria are the outcome of a method, the conduction time per method, and the assigned research personnel needed to run the study. Secondary criteria depend on the primary criteria and can be derived directly from them. Thus, the costs of a study depend on a combination of the conduction time and the research personnel assigned for each subtask within a method. In addition, the final results of a method can be classified as secondary criteria because they are dependent on the chosen methodology as well as the underlying research question, for example, with respect to the degree of detail which they provide to answer it. General criteria can be discussed only in terms of the specific results of one method. Examples are measurements of reliability and validity. These criteria need comparison beyond one specific method or study and need to be discussed with regards to results of previous work. Thus, the operationalized model in Table 6.1 will be used as a first conceptual tool for holistic comparison of descriptive evaluation methods in audiovisual quality research.

Comparison of test outcomes and test results The test results are key criteria for describing the applicability and power of a certain research method to answer specific research questions. The results of such comparison do have an impact on other general comparison classes, such as economy and excellence. While the outcome of a method can be directly compared as a primary criterion, the final comparison of test results must include the underlying research purpose or research question. Comparisons can be conducted based either on a discussion of outcomes or, when the data allow, on statistical analyses. Possible methods are either multivariate statistics or measures of correlations [4, 109]. Results as secondary criteria are generally compared in a discussion relating outcomes and research questions.

Comparison of durations for a method The conduction time of a method is a second important indicator for the economy aspects of a research method. The time that it takes to get results on a research question is a significant factor to calculate the costs for these results. While the time to prepare and to analyze a study is hard to quantify as it depends on the number and experience of the responsible researchers, the time for conducting a study is a relatively constant value. This first operationalization, defines milestones for conducting the study (Pre-Post-test, ACR, Descriptive evaluation, total) and for data analysis (Data preparations, analysis, total). During the each phase of

Comparison criterion	Definition	Selected measure
Primary criteria		
Research personnel	Classification of personnel into helpers, assistants, and experts based on their knowledge and skills	Identify the assigned research personnel for different tasks within the method taking into account the needed skills and knowledge to fulfill the task successfully
Duration	Duration for planning, conducting and analyzing the study	Definition of milestones that mark consecutive tasks within each method and measurement of time in minutes to fulfill each milestone. Comparisons can be done based on duration per participant for each milestone as well as for the overall duration
Outcome	Comparison of the outcome of each method after all analysis is finished	Comparisons can either be done on interpretational basis, but statistical comparisons are preferred if available (e.g. correlation, multivariate analysis)
Secondary criteria		
Costs	Comparison of results after all analysis is finished	Multiplication of durations for each task and the related (relative) costs per assigned research personnel per task
Test results	Results of the test with respect to the research question, e.g. in terms of amount or degree of gained information from the methods' outcome	Comparison of test results on interpretational basis. Within interpretation, the gained information related to the research question shall be discussed.
General criteria		
Validity	Comparison of results, similarities and differences, which allows a statement about whether the methods were able to measure the intended construct	Discuss the results of the method, the test development and how threats to validity were eliminated
Reliability	Comparison of test results of different methods and discussion of their consistency	Comparison of test results among the methods and comparison of results with related studies

Table 6.1. – Operationalized comparison criteria, their definitions, and measures towards a holistic comparison model for audiovisual quality evaluation methods.

the study, the time to achieve the following milestone is measured in minutes and later the measures taken can be analyzed and compared statistically.

Comparison of needed research personnel Beside the conduction time, the needed personnel that must be assigned to conduct a study is an important criterion for the selection of a specific research method. The number of personnel and the needed experiences determine both working effort and the total costs of a test. Although the topic of classification of research personnel is complex in terms of experience, training, and knowledge, the author suggests a classification according to three basic classes: helpers, assistants, and experts (Table 6.2).

Comparison of costs for a study The costs of a study mainly consist of the effort of research personnel per task of the study and the time spent for the phase. As costs for research personnel may differ between countries or academia and industry, the current approach uses relative costs.

Class	Definition	Tasks
Expert	An expert is familiar with the related research methods from planning a study until reporting of results. He has knowledge and experience in conducting and analyzing different research methods of psychoperceptual evaluation and descriptive techniques. In addition, he can give instructions about all stages of a study to other personnel.	Planning of an study Data analysis Interpretation of sensory results Reporting of results Responsibility for test design
Assistant	An assistant is able to conduct tasks during a study on his own after getting detailed instructions and training. Assistants guarantee a certain quality level and their tasks only need to be refined by experts from time to time.	Conduction of tests Data processing and basic analysis Contact to test participants
Helper	Helpers support the study by doing simple tasks that can be done after short instruction by an expert. They can fulfill very specific tasks on their own without continuous supervision of an expert. Helpers may not have any knowledge about the study or work beyond their tasks.	Recruitment of test participants Publicity Laboratory preparations Data preparations Transcriptions

Table 6.2. – Classification of research personnel for audiovisual quality evaluations.

In Germany, the relation of costs for helper:assistant:researcher is approximately 1:2:4 [1]. Based on this, relative costs can be calculated per method by *task cost = costs for assigned personnel* $*$ *duration of task*. The total costs of a study are then the cumulated task costs.

6.2. Study 6: Comparison of OPQ with related mixed-methods research approaches

During the recent work towards mixed-methods research approaches, some other methods in addition to Open Profiling of Quality have been presented [37, 38]. Although all methods use the standardized methods of ITU recommendations for the quantitative evaluation, the applied descriptive methods differ and make use of either interviews techniques or sensory evaluations. Until now, no systematic comparison of the methods has been conducted to identify strengths and weaknesses of the different approaches. The goal of this study is an initial systematic and holistic comparison of different mixed-methods approaches by applying the developed comparison model. The study will demonstrate the operationalization of the model as a research tool in a comparison of four mixed-methods approaches for audiovisual quality assessment.

[1] This ratio is based on the current regulations for payment of researchers and (non-)scientific assistants in accordance to the labor agreement 2011 for the German public service.

6.2.1. Research method

6.2.1.1. Test participants

Fifty test participants took part in the study (see Table 6.3). All participants were recruited according to the user requirements for mobile 3D television. They were screened for normal or corrected-to-normal visual acuity (myopia and hyperopia: Snellen index 20/40), color vision using the Ishihara test, and stereo vision using the Randot Stereo Test (≤60?arcsec). All test participants could be classified as naïve assessors because they had experience neither in the domain of research nor in subjective quality evaluation studies. All test participants passed the psychoperceptual evaluation. Subgroups of the test sample were assigned randomly to one of the descriptive methods (Table 6.3).

	Psychoperceptual evaluation	Open Profiling of Quality	Post-task interview	Descriptive Sorted Napping
Participants	N = 50	N = 16	N = 17	N = 17
Gender and age [years]	f / m: 26 / 24 Mean age: 23 (2.5)	f / m: 8 / 8 Mean age: 24 (1.8)	f / m: 9 / 8 Mean age: 23 (2.9)	f / m: 9 / 8 Mean age: 23 (2.5)

Table 6.3. – Characteristics of the test samples per method under comparison.

6.2.1.2. Test stimuli and stimuli presentation

The test stimuli and the stimuli presentation were similar to those in Study 4 (section 4.5). Eventually, there were six different contents presented at two different quality levels. Audio was not impaired. The tests were conducted in a laboratory at Ilmenau University of Technology, and test conditions were arranged according to the specifications in ITU-T P.910 [5]. Different playlists in pseudo-randomized orders were used during video presentation. During the psychoperceptual evaluation, each test item was presented twice. In OPQ and the Descriptive Sorted Napping, each item was presented once and a second time upon request of the test participant. In addition, test participants were allowed to ask for later replay of specific test items during the Napping procedure.

6.2.1.3. Test procedure

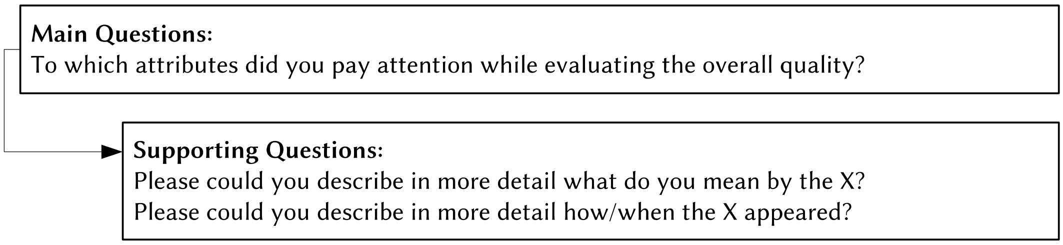
For the different mixed-methods approaches being tested, we combined a psychoperceptual evaluation with one of the three approaches. The three mixed-methods approaches of Open Profiling of Quality, post-task interview, and Descriptive Sorted Napping were selected in accordance with previously reported descriptive research methods in the field of audiovisual quality evaluations.

Psychoperceptual evaluation Psychoperceptual evaluation started with the test participants' screening and the explanation of the test procedure. In the following training and anchoring, we presented a subset of test items that covered the full range of constructed quality. Test participants were asked to find their best viewing position and to practice the evaluation task. We applied

Absolute Category Rating (ACR) according to ITU-T P.910 [5] for the psychoperceptual evaluation of the overall quality, rated with an unlabelled 11-point scale. The stimuli were presented one by one, and the participants rated retrospectively their overall satisfaction. In addition, the acceptability of overall quality was rated on a binary (yes/no) scale [57]. Each stimulus was assessed twice. After a short break of about 10 minutes, during which the participants filled out a demographic data questionnaire, the sensory evaluation was conducted.

Open Profiling of Quality Open Profiling of Quality was implemented in accordance with its originally proposed four-step structure of 1) introduction, 2) attribute elicitation, 3) attribute refinement, and 4) sensory evaluation. During the introduction, test participants were familiarized with describing sensations using their own words using the 'apple task'. During the attribute elicitation, a subset of 10 test items was presented one by one, each item twice. The participants were asked to write down their individual attributes on a blank sheet of paper. They were not limited in the number of attributes, nor were they given any limitations on describing sensations. During the attribute refinement step, participants were asked to rethink (add, remove, change) their attributes before defining their final list of words. In addition, we collected definitions and labels for the minimum and maximum sensation for each attribute. The final attributes were transferred to score cards. The final evaluation was then conducted on each test item based on the individual attributes. The test participants rated the sensation of each attribute. The whole OPQ study was conducted in one session in contrast to its original description.

Post-task Interview The post-task interview was conducted in accordance with the Experienced Quality Factors approach of Jumisko-Pyykkö et al. [37]. A semi-structured interview was conducted after the psychoperceptual evaluation based on a free-description task. During the interview, the test participants were encouraged to describe their impressions of the overall quality as broadly as possible. The semi-structured interview included main and supporting questions (Figure 6.2). While main questions were asked several times during the interview with slight variations, the supporting questions helped to clarify the answers given in response to the main questions. Only terms introduced by the participant were used during each interview. No additional stimulus material was used.



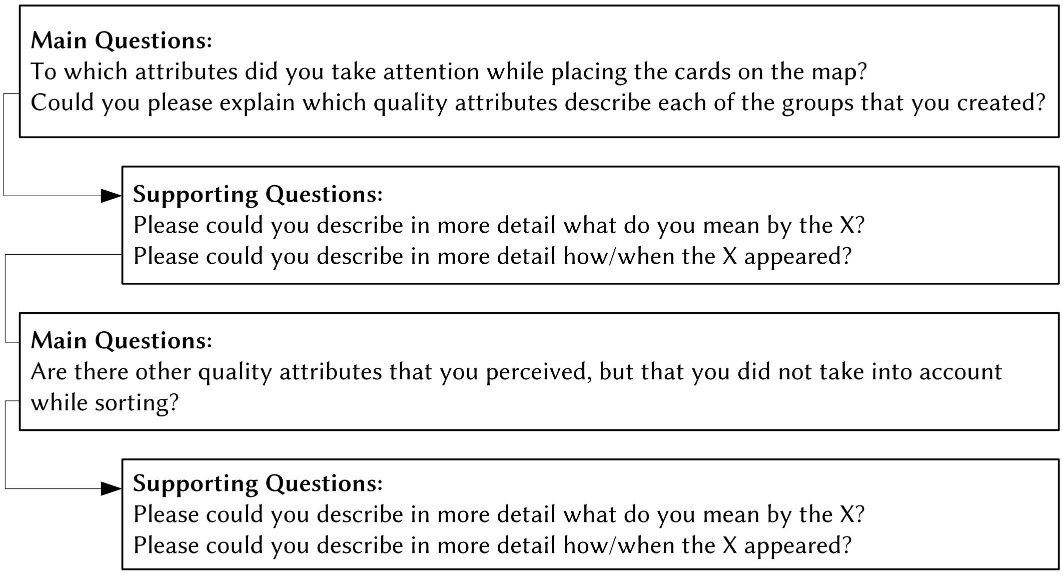
Main Questions:
To which attributes did you pay attention while evaluating the overall quality?

Supporting Questions:
Please could you describe in more detail what do you mean by the X?
Please could you describe in more detail how/when the X appeared?

Figure 6.2. – Structure of the semi-structured interview in the post-task interview method, including main and supporting questions.

Descriptive Sorted Napping The descriptive sorted napping procedure is based on the ideas of free perceptive mapping tasks and sorted napping [91, 96, 97, 100], but extends these methods with a short post-task interview. In the first step, test participants received a sheet of blank paper (nappe: French for tablecloth) 65x47.5 cm in size (landscape orientation). In addition, they received small white cards (6x5cm) numbered from 1 to 18. The test participants were told that the cards represented the test items that would be presented one after another on the device. After each video, they were place the card representing that video on the nappe in such a way that similarly perceived videos were placed close together, while differently perceived videos were farther away from each other. No limitations were placed on sorting the items on the nappe. After this instruction, we played one video after the other. When a new video started, the test participant took a new card, watched the video once, and placed the card on the nappe. The video was played again upon request. Test participants were also allowed to request a replay of previous test items if necessary.

After the presentation of all test items, participants were asked to describe their nappe with their individual attributes. This use of extended sorting tasks is known as Ultra-flash Profiling [141] or Sorted Napping [97]. In Sorted Napping, test participants draw a line around a group of cards and describe the group with a set of individual attributes. Intersections between groups are allowed. Because previous applications of perceptive mapping approaches had shown that the results were limited to the two dimensions given by the nappe, we conducted a short post-task interview after the sorted napping task. This semi-structured interview targeted 1) an additional descriptive/explanatory level of the developed sorting and 2) an evaluation of additional attributes that were not taken into account when creating the sorting (Figure 6.3).



Main Questions:
To which attributes did you take attention while placing the cards on the map?
Could you please explain which quality attributes describe each of the groups that you created?

Supporting Questions:
Please could you describe in more detail what do you mean by the X?
Please could you describe in more detail how/when the X appeared?

Main Questions:
Are there other quality attributes that you perceived, but that you did not take into account while sorting?

Supporting Questions:
Please could you describe in more detail what do you mean by the X?
Please could you describe in more detail how/when the X appeared?

Figure 6.3. – Structure of the semi-structured interview in the Sorted Napping method, including main and supporting questions and a two-stage structure asking for attributes included or not included in the sorting of the maps.

6.2.1.4. Methods of Analysis

Psychoperceptual evaluation For the psychoperceptual evaluation, nonparametric methods (Kolmogorov-Smirnov: $p < .05$) were used for the analysis. Within methods, the results were analyzed using a combination of the Friedman and Wilcoxon tests. The unrelated data among methods were analyzed applying Kruskal-Wallis, and pairwise comparisons were conducted with the Mann-Whitney-U test [4]. Frequencies were counted for the acceptance ratings. PASW Statistics 18 was used for quantitative data analysis. Analysis was conducted on the joint quantitative data set as well as separately for the samples of each descriptive method.

Open Profiling of Quality The OPQ data were analyzed using multiple factor analysis comparable to that used in Study 3 and Study 4 (sections 4.4 and 4.5).

Post-task Interview The analysis procedure of the interview data followed the ideas of data-driven frameworks. Our analysis approach was based on the Grounded Theory framework by Strauss and Corbin [66]. First, the interviews were transcribed, and meaningful pieces of data were extracted. Second, we applied open coding to identify concepts and categories among the extracted pieces of data. In a final step, the codes were categorized into major categories and subcategories [9]. The developed framework was reviewed by a second researcher and categorization was then repeated by the researcher to calculate interrater reliability (Cohen's Kappa [4]).

Descriptive Sorted Napping The maps of the descriptive sorted napping were transformed into two different data sets. The first data set contains the position of each test item on the nappe in terms of X and Y (Figure 6.4). The second data set describes whether a test item was described with an individual attribute or not. By doing so, one gets a two-dimensional (X/Y) quantitative description of the test items and a binary classification based on the individual attributes. From these data sets, a MFA based on the X-Y data with the attributes data set as supplementary variables was conducted [202]. The interview data were analyzed in a manner similar to the post-task interview method. With respect to the two main questions in the semi-structural approach of the interviews, separate classifications were created for each question.

6.2.2. Results

6.2.2.1. Psychoperceptual Evaluation

Acceptance of Overall Quality Overall, the quality levels differed in their acceptance scores. The highest quality level qp30 reached an acceptance level of over 90%. While qp40 was close to the threshold of an acceptance level of 50%, qp45 did not reach the 20% acceptance level. Figure 6.5 presents the acceptance levels averaged over methods as well as method-by-method. For all methods, comparable acceptance levels were found.

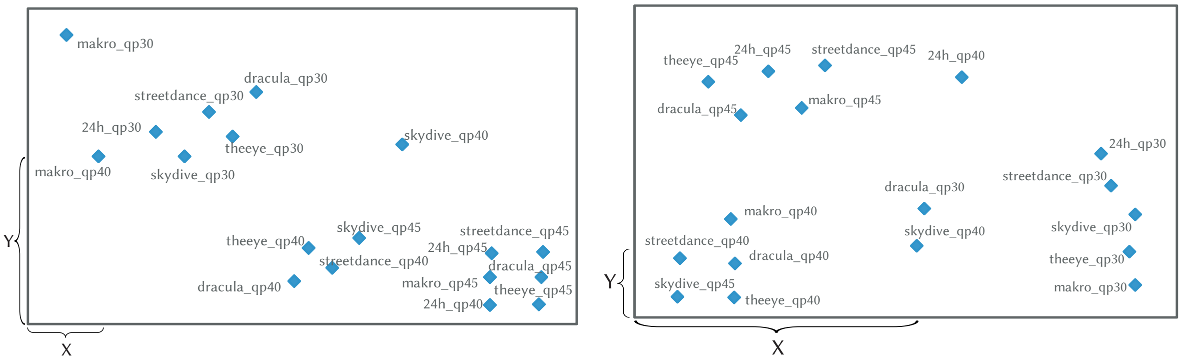


Figure 6.4. – Examples of two sortings obtained during the Sorted Napping task.

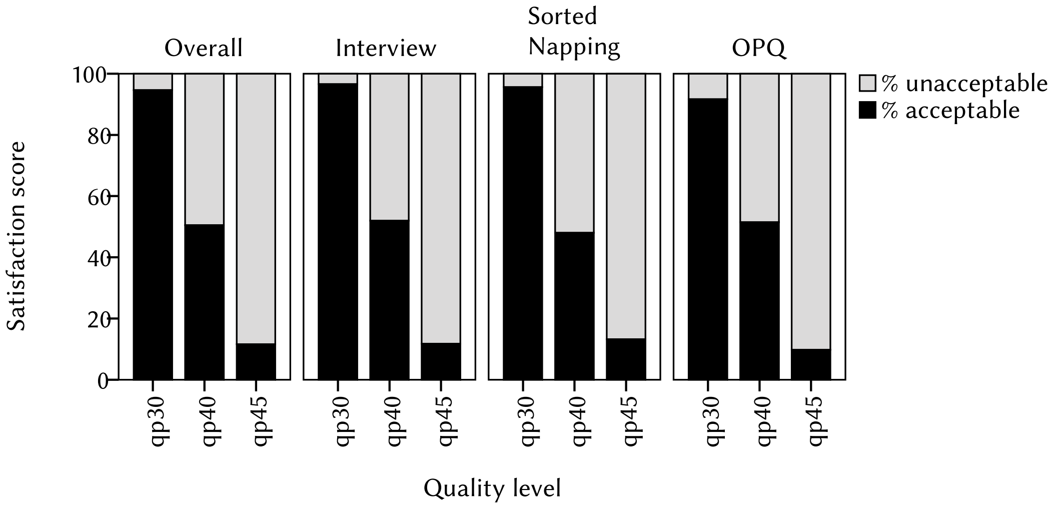


Figure 6.5. – Results for acceptance scores of overall quality.

Satisfaction with Overall Quality Overall, the quality level had a significant impact on the perceived overall quality when averaged over the contents (all comparisons: <.001). Figure 6.6 shows that qp30 received the highest ratings while qp45 was significantly rated worse on average as well as in content-by-content comparison. To compare the descriptive results of the different methods, a common quality satisfaction among methods is necessary so that the same quality being described can be assumed. The methods-wise analysis shows that only in five cases can slight differences in the overall quality ratings be identified (dracula_qp30, theeye_qp30: P<.05; makro_qp30, makro_qp40, skydive_qp45: P<.01). Nevertheless, the overall differences among the quality levels are comparable across methods.

6.2.2.2. Open Profiling of Quality

The results of the Open Profiling of Quality evaluation are discussed in detail in Study 4 (see section 4.5). Briefly summarizing for this comparison, the findings of OPQ show that the first dimension

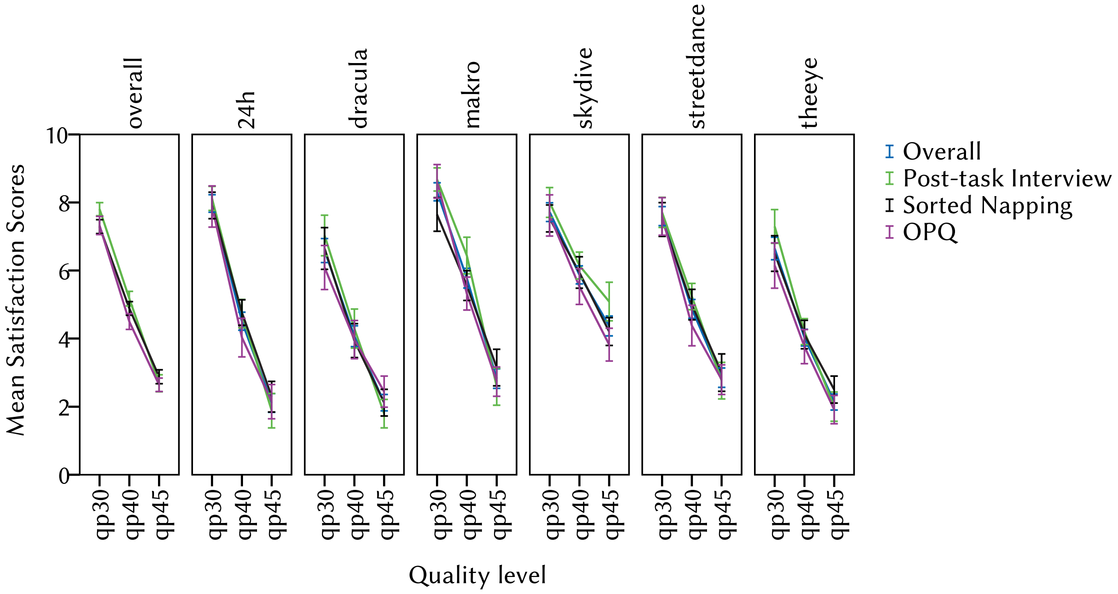


Figure 6.6. – Results for satisfaction with overall quality. The bars show 95% CI of mean

of the MFA model separates the test items along the video quality (Figure 6.7). This dimension is described with attributes like 'blocky' or 'artifacts' on its negative polarity and 'clear' or '3D effect' on its positive polarity (Figure 6.8). Along the second dimension, Dracula items were separated from the other contents. This dimension is described with attributes like 'double images' or, conversely, 'perceivable as one picture' and correlates mainly to problems caused by perception of crosstalk due to high disparity. The External Preference Mapping shows clear preference for items of high quality level (Figure 6.7).

6.2.2.3. Post-task Interview

Experienced quality is constructed from components of visual quality (depth, spatial, temporal), viewing experience, and interactions with other modalities (Table 6.4). Visual quality consists of the three subcategories of spatial, temporal, and depth. Visual depth quality is characterized by the test participants' ability to detect depth and the impression that the depth perception creates. In addition, depth is also described as different layers of foreground and background and the ability to detect the composition of the content on these layers. Furthermore, the erroneous nature of depth perception in terms of double pictures was mentioned. Visual spatial quality is comprised of the subcategories of sharpness, brightness, color, and resolution of the video, which all contribute to the ability to detect objects, outlines, and details in the content. In addition, negative categories like blur and visible artifacts were identified. The annoying impact of visible pixels and blocking artifacts was mentioned by every test participant. Many of them also said that the blocking artifacts hindered them from detecting details or correcting outlines of objects: "And because of the pixels it became very hard to detect certain details" (male, 21). Finally, visual temporal quality summa-

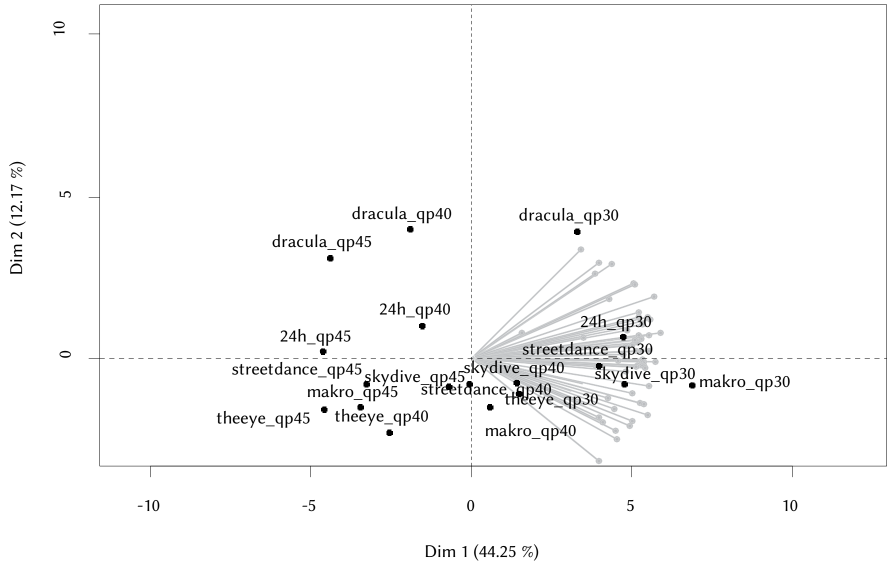


Figure 6.7. – Item plot for the Open Profiling of Quality. The gray arrows mark the quantitative preferences of test participants from PREFMAP.

rizes the characteristics of motion from general mentions of motion to its fluency and unimpaired characteristics. This category was the least mentioned of the three main categories of visual quality.

Viewing experience described the users' high-level interpretations of the system (media, content) or the system's influence on users' perception. It extends the descriptions of quality beyond direct characteristics of the stimuli (e.g., details, colors, artifacts). However, they emphasize the interpretation of stimuli, including users' knowledge, emotions, or attitudes, as a part of quality experience. The most mentioned subcategory was Pleasantness and Interest which relates to the motivation of a user to watch the content: "I don't know if it was content-dependent or if it is just my personal impression that I prefer this kind of videos over others" (female, 24). The comfort of viewing is highlighted by the ease of viewing as well as the comparison of the system being tested to other available audiovisual systems. However, 3D still caused negative effects for some of the test participants, summarized in the category of visual discomfort. Finally, the overall quality illustrates the total impression of quality. Few attributes related to descriptions of audio and audiovisual attributes were found.6

6.2.2.4. Descriptive Sorted Napping

The MFA result of the Sorted Napping analysis can be interpreted as a common perceptual sorting map among test participants. The first two dimensions account for 52.72% explained variance (Dim1: 39.75%, Dim2: 12.97%). The item map (Figure 6.9) shows a separation of items into three groups according to the different QPs. Along dimension 1, QP30 and QP45 are the determining parameters; QP40 separates along dimension 2 from the other two groups. The correlation plot (Figure 6.10) underscores this trisection of items with three visible clusters of attributes. In general,

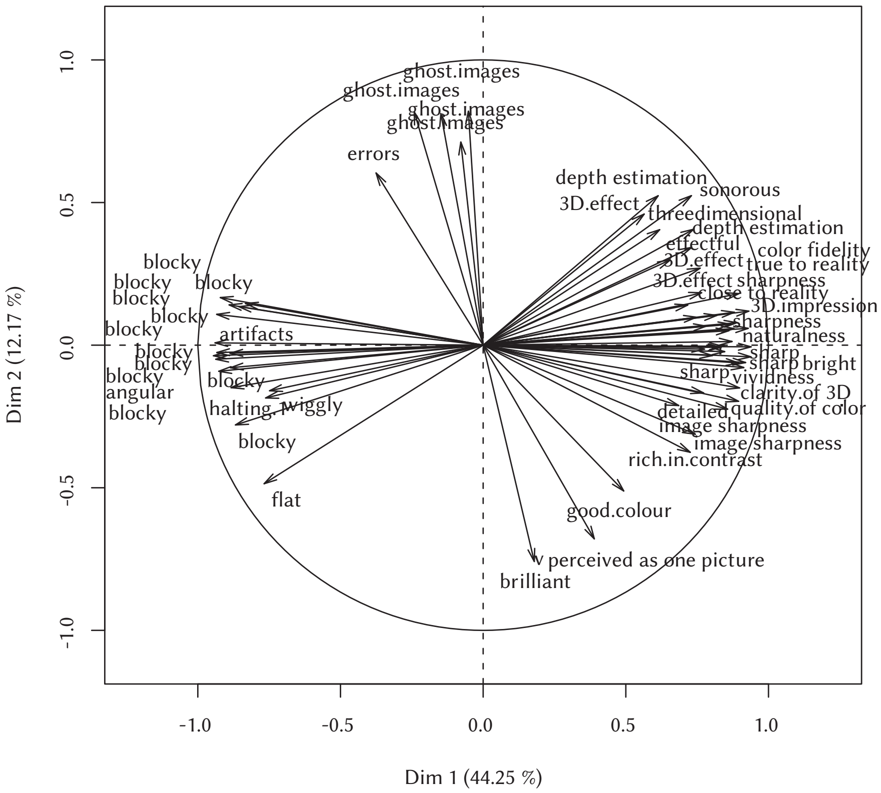


Figure 6.8. – Correlation plot for the Open Profiling of Quality.

the three groups are described with attributes like ’best quality’ (QP30), ’medium quality’ (QP40), and ’bad/worst quality’ (QP45). The group of QP30 items is correlated with attributes like ’sharp images’, ’3D perceivable well’, and ’would watch it again’. QP40 items are described with attributes like ’acceptable quality’, ’objects still detectable’, or ’makes no fun to watch’. The final group of QP45 is described with attributes like ’very blocky’, ’details not detectable’, and ’very exhausting to watch’.

The classification of the items and the underlying quality attributes are confirmed in the results of the post-task interview analysis (Table 6.4). We developed two classifications, one for explaining the sorted maps and one for extending knowledge about additional quality factors not taken into account while creating the sorting. The classes for descriptions of the map construct experienced quality from components of visual quality (depth, spatial, temporal), viewing experience, and content. The component of visual depth relates to the perceivable 3D sensation and related artifacts caused by perception of double pictures. For visual spatial quality, the dominating categories are sharpness, resolution, and detection of objects and details for a positive description, as well as blurry and visible pixels to describe degraded excellence of the stimuli. Visual temporal classes comprise descriptions of motion in general as well as fluent and clear (artifact-free) perception of motion. In the main category of viewing experience, the description of the overall quality in general as well as test participants’ interest in the content and related pleasantness are mentioned. The effect of visual

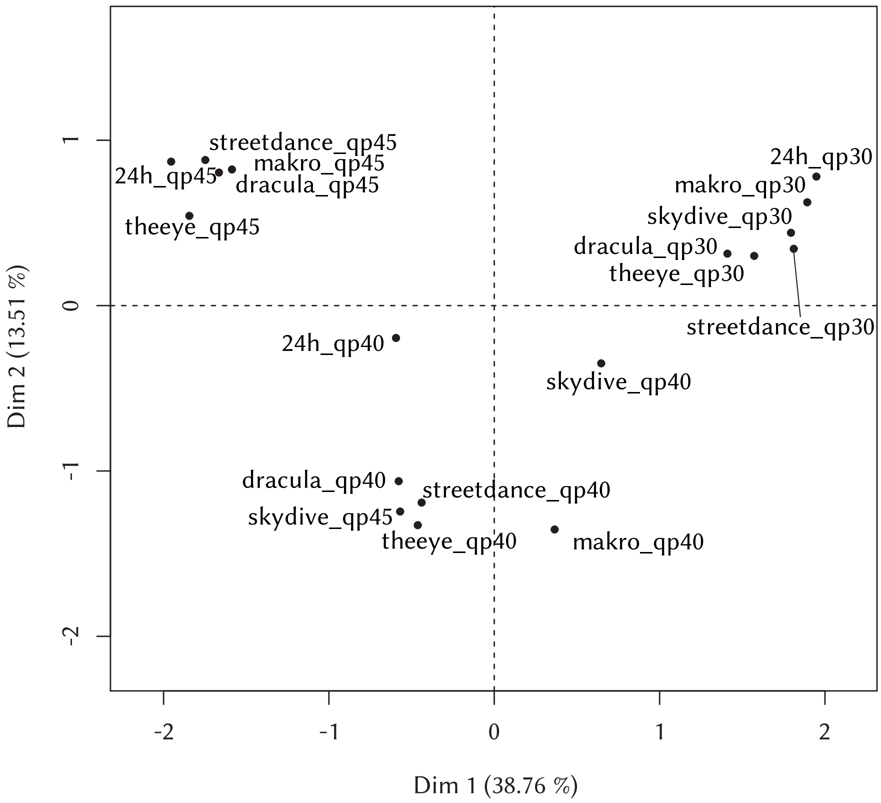


Figure 6.9. – Item plot of the Sorted Napping based on the MFA analysis.

discomfort resulting in a feeling of discomfort in the eyes was mentioned by most test participants. The content class finalizes the descriptive classes in relation to the sorted maps.

The attributes not represented in the creation of the perceptive maps include only a few categories. The largest class is visual depth with the subcomponents of perceivable depth and related perception of double pictures. Other visual categories are color and brightness and descriptions of motion in general. For the category of viewing experience, some test participants described their interest in the content as well as ergonomic aspects of the ease of viewing. In addition, visual discomfort and 3D added value were mentioned. The content category concludes the classification of the interview data.

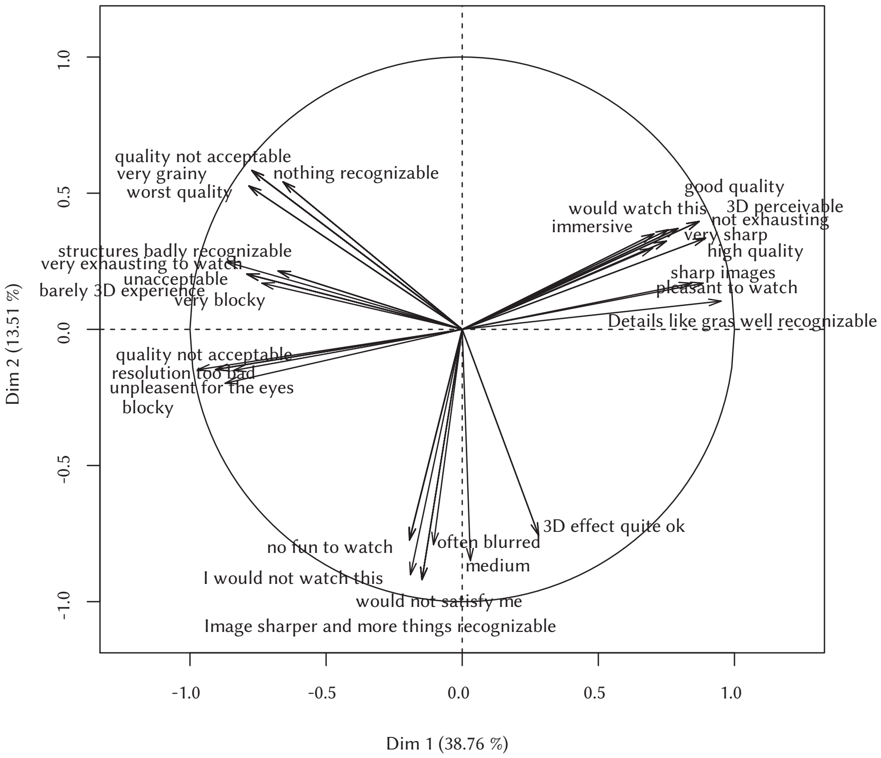


Figure 6.10. – Correlation plot of the Sorted Napping based on the MFA analysis.

COMPONENTS (major and sub)	DEFINITION	Interviews Kappa = 1.0	Napping Included Kappa = 0.8	Napping Not included Kappa = 1.0	Rank in Post-task Interview	Rank in Napping Incl./not incl.
Visual Depth	Descriptions of depth in video					
Perceivable depth	Ability to detect depth or variable amount of depth as a part of presentation	70.6	56.3	37.5	2	3/1
Impression of depth	3D effect creates a natural and realistic impression	70.6	*	*	2	*
Double pictures	Seeing separate left and right channels, crosstalk of the two channels	41.2	25.0	18.8	7	8/2
Separation Foreground/Background	Depth is composed of foreground and background layers	29.4	*	*	9	*
Visual Spatial	Descriptions of spatial video quality factors					
Sharpness	Descriptions of sharp and clear images; artifact-free	41.2	43.8	*	7	5/*
Color	Descriptions of the color and its excellence	52.9	*	12.5	5	*/3
Brightness and contrast	Excellence of brightness and contrast	41.2	*	*	7	*
Resolution of the video	Descriptions of the resolution of the video	35.3	37.5	*	8	6/*
Blurry	Blurry, inaccurate, not sharp	17.6	31.3	*	11	7/*
Visible pixels	Impairments with visible structure (e.g. blockiness, graininess, pixels).	100.0	81.3	*	1	1/*
Detection of objects	Ability to detect and identify objects	35.3	56.3	*	8	3/*
Detection of edges and outlines	Ability to detect edges between objects and excellence of the outlines	35.3	*	*	8	*
Detection of details	Ability to detect details and fine structures in the video	58.8	25.0	*	4	8/*
Other visual artifacts	Other visual artifacts are mentioned, but not explained	17.6	*	*	11	*
Visual Temporal	Descriptions of temporal video quality factors					
Motion in general	General descriptions of motion in the content (slow, fast)	47.1	25.0	18.8	6	8/2
Fluent motion	Good temporal quality (fluency, stops judder)	23.5	12.5	*	10	6/*
Clarity of motion	Artifact-free videos in temporal quality (motion blur, natural movements)	11.8	37.5	*	12	6/*
Viewing Experience	User's high level constructs of experienced quality					
Visual discomfort	Feeling of discomfort in the eyes	47.1	62.5	6.3	6	2/4
Comparison to other technology	Quality of the current system is compared to other existing video technology (HD, 2D, MobileTV, Internet)	35.3	37.5	*	8	6/*
Pleasantness and Interest	Pleasurable viewing experience and interest in the content, also for a longer period of time (e.g. 15min)	64.7	56.3	12.5	3	3/3
3D Added value	Added value of the 3D effect (advantage over current system, fun, worth of seeing, touchable, involving)	*	12.5	6.3	*	9/4
Ease of viewing	Easy to concentrate on viewing (e.g. free from extra effort and learning, viewing angle does not interrupt viewing)	23.5	*	12.5	10	*/3
Overall quality	Experience of quality as a whole without emphasizing one certain factor	35.5	50.0	*	8	4/*
Other Modalities Interactions	Descriptions of quality of audio modality and interaction between quality of audio and visual modalities	5.9	18.8	27	13	13
	Total number of attribute descriptions	**225**	**113**			

Table 6.4. – Results of the qualitative analysis for post-task interview and Descriptive Sorted Napping (* means that attribute was not mentioned).

6.2.3. Systematic comparison of the methods being tested

For a systematic and holistic comparison of the five methods under assessment, the developed comparison tool with its selected attributes is applied to the results obtained from each method.

6.2.3.1. Comparison of test results

The test results are key criteria in describing the applicability and power of a certain research method to answer specific research questions. The results of this comparison do have an impact on other general comparison classes such as economy and excellence. The outcome of each method and its impact on answering the research question are summarized in Table 6.5.

In general, all three descriptive methods extend the understanding of the preference order from the psychoperceptual evaluation (Table 6.5). The descriptive results highlight the complementation and convergence of results from different research methods and underscore the positive features of mixed-methods research. All results can basically explain the excellence of test items and the impact of the underlying parameters. All results agree in terms of which components have the greatest impact on the perceived quality and emphasize the importance of visual quality in the perception of 3D Quality of Experience. The results show descriptions of components by pairs of antonyms and describe video quality based either on artifacts (blur, pixels) or on positive characteristics (detection of details, sharpness). Other general components like depth perception and fluency of content were identified by all methods as well. However, there are also differences between the results of methods, which will be discussed further.

Interview-based approaches The results obtained from the interviews (post-task interview and Descriptive Sorted Napping) show the most detailed description of underlying components of quality (Table 6.4). The results underscore the manifold dimensions that impact the perceived quality of mobile 3D video. Dimensions are weighted according to how frequently they were mentioned. The analysis of open coding allows for taking into account attributes that were mentioned only a few times. The findings underscore the interaction between the perception and description of excellence of the video according to Visual Depth, Visual Spatial, and Visual Temporal and the users' high-level interpretations of the system (Viewing Experience). The semi-structured approach of the post-task interview resulted in very detailed descriptions of main categories and subcategories. The categories complement each other very well and form the most holistic quality descriptions. They provide detailed information on different aspects like 3D perception and its connection to foreground-background separation as well as to the difference between perception and 3D impression or more detailed information on object detection. In contrast, the napping interview allowed for the creation of a pseudo-hierarchical structure based on the descriptions of sorted attributes and non-respected attributes of the sorted map. This hierarchy underscores the dominance of visual quality and its superiority over depth perception. However, the interview results do not provide a direct link to the psychoperceptual results beyond interpretation.

Method	Outcome	Results
Psychoperceptual evaluation.	Acceptance Scores for each item; Mean Satisfaction Scores for each item	QP30 received the significantly highest ratings in overall acceptance and satisfaction scores; QP40 was worst rated in both scores
OPQ	Two-dimensional MFA model; Correlation plot of individual attributes and MFA dimensions; External Preference Mapping	Video quality and perception of crosstalk identified as the crucial quality factors; descriptions of low artifact level and perception of depth combined in good video quality; different impact of large disparity in Dracula content - either perception of crosstalk or perception of enhanced 3D impression; clear preference of users towards artifact-free stimuli
Post-task interview	Framework of components of 3D Quality of Experience for mobile 3D video	Identification of general underlying quality parameters; Perceived quality consists of components of Visual Depth, Visual Spatial, Visual Temporal, Viewing Experience and Other Modalities; highest frequencies for visible pixels, perception and impression of depth, and pleasantness of viewing
Descriptive Sorted Napping	Two-dimensional MFA model representing mean sorting of users; independent components of 3D Quality of Experience with respect to whether the component is taken into account for sorting	Mean sorting shows discrimination of test items according to the different QPs; correlation of individual attributes reveals discrimination in terms of perceived video quality and differs mainly between artifact-free perception and perception of blockiness; Interviews confirm the general classes of post-task interview; most important categories included in sorting are visible pixels and visual discomfort; most important categories not included in maps are perceivable depth and double pictures

Table 6.5. – Summary of the different outcomes and results of the methods under comparison. The table shows how each method extended the knowledge about the underlying quality rationale and deepened the results of the psychoperceptual evaluation.

Profiling-based approaches Both profiling-based approaches can separate and discriminate the test items in comparable ways. However, the characteristics of the developed maps are different between Sorted Napping and OPQ. The maps of the Sorted Napping can be interpreted as a mean sorting of all test participants. Although they all separate the items only along the quality level, they are distributed over two dimensions representing the two degrees of freedom on the nappe. Each of the three obtained groups of items is described with a set of correlating attributes. In contrast, the obtained item maps from OPQ a common low-dimensional model of individual higher dimensional configurations. These results represent the most salient factors common to the individual configurations and model these together with descriptive attributes. Lower impacting test items and attributes (and other possible components of the final model) are not very well represented in the results and are very often neglected for the sake of clarity of results. Although the PCA helps to identify the most salient and therefore the most critical items and components, it omits information from the results that can be found in the interview data of the Descriptive Sorted Napping. In contrast, regression methods allow for simple linking of sensory profiles from OPQ with the psychoperceptual results.

Overall, the results show that both interview-based and profiling-based approaches can identify the critical components of 3D Quality of Experience. The strengths of the interview-based approach lies in the creation of a very general component model of Quality of Experience that takes into account the manifold dimensions and weighting of dimensions in terms of frequencies. Its weakness is the possibility of modeling relationships between attributes and components. This shortcoming is overcome by sensory profiling methods that allow for detailed modeling of critical components and description of the components through correlation of attributes. However, these models are derived only from the dominating attributes. Individual differences are often underestimated and not represented in the final models.

6.2.3.2. Comparison of durations

The time from the point when the test participant entered the laboratory until leaving it after the test was measured to determine the conduction time (Table 6.6). The time for analysis was measured from the beginning of the analysis until the final results were available (Table 6.7).

	Total	Pre-Post test	ACR	Descriptive task
OPQ	103.4 (12.6)	24.1 (5.9)	33.3 (5.1)	40.5 (7.8)
Post-task interview	67.6 (10.1)	21.4 (6.4)	30.6 (6.0)	9.8 (2.0)
Sorted Napping	96.3 (11.8)	21.9 (5.6)	31.9 (5.5)	37.1 (8.2)
Napping / Interview				*28.9 (8.2) / 8.2 (4.2)*

Table 6.6. – Durations [in minutes per participant] for data collection for the different methods. Values in brackets show standard deviations.

Statistical analysis of the conduction time reveals significant differences between methods for the overall conduction time (Kruskal-Wallis: $\chi^2 = 37.12$, $p < .001$) and for the conduction time of the descriptive tasks (Kruskal-Wallis: $\chi^2 = 48.67$, $p < .001$). No differences were found for the pre+post test time and for the ACR (all comparisons: $p > .05$). Because no differences were identified for the pre-post section as well as for the ACR, the differences in the overall time arise from different durations in the descriptive evaluations. A detailed analysis of methods shows that the post-task interview was significantly the shortest method (Interview vs. Napping: Z = -4.99, $p < .001$). No differences were identified between OPQ and Napping (Interview vs. Napping: Z = -0.885, $p > .05$, ns). However, a comparison of the Sorted Napping (without the interview) and the OPQ shows significantly shorter conduction time for the Sorted Napping (all comparisons: P< .001).

While the conduction time is rather easy to measure, measurements for analysis are hard to determine. They depend on the skills of the researcher responsible for analysis. Another factor is the way data are collected. Statistical analysis of the durations of analysis reveals significant differences between methods for durations (all comparisons: <.001). The duration for analysis of Absolute Category Rating was not taken into account for the comparison because the length can be seen as constant. No changes to ACR in terms of methods and analysis of conduction time revealed any differences.

	Data preparations	Analysis	Total
OPQ	11.2 (1.8)	2.0 (est.)	**13.2**
Post-task interview	50.3 (9.5)	15 (est.)	60.3 (est.)
Sorted Napping	32.7 (10.8)	12 (est.)	44.7 (est.)
Napping / Interview	*6.4 (0.9) / 26.3 (10.8)*	*2.0 (est.) / 10 (est.)*	(8.4 / 36.3)

Table 6.7. – Durations [in minutes per participant] for data preparation and data analysis for the different methods. Values in brackets show standard deviations. Durations for qualitative data analysis and the sensory analysis of FCP data are estimated (est.) based on my experience and knowledge.

6.2.3.3. Comparison of needed research personnel

For the presented comparison of research methods, the author assigned personnel to the different tasks according to experience obtained in previous studies. The comparison of methods in Table 6.8 led to the presented results, and all assigned workers were able to conduct their tasks as expected.

	Planning of Study	Recruitment	Conduction of test	Preparation of data	Quantitative data analysis	Sensory data analysis	Descriptive data analysis	Interpretation of results
Open Profiling of Quality	Expert	Assistant	Assistant	Helper	Expert	Expert	-	Expert
Post-task Interview	Expert	Assistant	Expert	Helper	Expert	-	Expert	Expert
Descriptive Sorted Napping	Expert	Assistant	Expert	Helper	Expert	Expert	Expert	Expert

Table 6.8. – Assigned research personnel for the subtasks within each method being compared.

6.2.3.4. Comparison of costs per study

Finalizing the comparisons of methods, the costs for each method can be calculated. I calculated the relative costs as the product of the duration of each task and the relative costs per task for the assigned research personnel. As relative costs, a ratio of 1:2:4 for helper:assistant:expert is taken as a basis for the calculations. For the comparison, I disregarded costs for planning the study and recruitment because I assumed them to be constant for each method. The costs can be found in Table 6.9.

6.3. Discussions and Conclusion

The goal of this section was a holistic comparison of the Open Profiling of Quality approach to other related mixed-methods approaches. The literature review on criteria for comparisons showed that holistic comparisons need to take into account a set of criteria rather than only one aspect at a time to offer a complete picture of strengths, weaknesses, and limitations of the different methods, from planning to interpretation of results. Based on a first conceptual attempt to operationalize the component model for structured comparison of methods, three different methods were contrasted within this section: Open Profiling of Quality, post-task interview, and Descriptive Sorted Napping.

	Conduction of test	Preparation of data	Sensory data analysis	Interview data analysis	**Total**
Open Profiling of Quality	**206.8** (103.4 * 2)	**11.2** (11.2 * 1)	**8** (2 * 4)	-	**226**
Post-task Interview	**270.4** (67.6 * 4)	**50.3** (50.3 * 1)	-	**60** (15 * 4)	**380.7**
Descriptive Sorted Napping	**385.2** (96.3 * 4)	**32.7** (32.7 * 1)	**8** (2 * 4)	**40** (10 * 4)	**483.5**

Table 6.9. – Relative costs per participant [costs per participant] for the methods being compared. Each cost was calculated as the product of the duration per task and the relative costs per assigned researcher per task (see brackets).

The discussion of the results of the methods comparison is constrained by the following limitations. First, the applied comparison model is a first conceptual model. The operationalization of the studied comparison components is restricted to easily measureable aspects. More complex aspects as well as further refinement of the current concepts need to be considered. Second, latest developments in the compared methods are not yet considered. The post-task interview methods have recently been extended with methods of correspondence analysis that have not been conducted [9]. Third, the measures of duration for the qualitative analysis of interview data are based only on estimations because thorough measures of durations were not possible due to analysis of data being divided among several researchers. However, the results of the conducted study offer valuable information about strengths and weaknesses of the methods being compared and can be used to eventually formulate guidelines for practitioners to select descriptive (mixed) research approaches.

Decisive for the principles of validity and reliability of the methods is the comparison of test results. In the currently operationalized comparison model, test results need to be discussed as the outcome of a method in relation to the underlying research question. In general, the results of the methods are comparable to each other and are also consistent with the findings of other studies [58, 152][226] emphasizing validity and reliability of the methods. All methods were able to identify the most crucial quality factors either by contribution to the MFA model or by frequencies in the qualitative analysis. However, the results also show differences in the details of components. While sensory evaluation-based approaches like OPQ or the Sorted Napping restrict the models to the most salient factors and neglect the impact of lower dimensions, interview-based approaches offer possibilities for eliciting a wide and general set of components of Quality of Experience. However, they are not able to provide models that can be linked to quantitative preferences.

Beyond test results, the methods were compared based on criteria of durations, research personnel, and relative costs. Although these comparisons are only a first attempt to extend the comparisons towards more holistic aspects, the results show differences between the methods but also offer possibilities for improving these towards a final set of evaluation methods. While the time needed to plan the studies was not taken into account and was assumed to be constant for all methods, the comparison of durations for conducting and analyzing the studies shows large differences. While, for sensory evaluation approaches, the time to conduct a study is significantly higher in contrast to interview-based approaches, the durations for analysis of interview data exceeds those of the sensory evaluations. Durations for analyzing sensory data may even be shorter when evaluations can be GUI-based because the measuring task for the ratings can be omitted [2].

Second, the assigned research personnel and the resulting relative costs per method were compared. The assigned research personnel do have impact on the overall costs of the study. The personnel comparison shows possibilities of decreasing costs for each method. The results in relation to the relative costs indicate that training of personnel (assistants) to conduct the study can significantly decrease costs for the methods. In the current comparison, interviews were conducted and analyzed by experts. Training assistants to conduct a study may lead to half the cost for these methods. Such a decrease of costs for conducting the interview-based studies may compensate for the greater expense, allowing the final decision on a research method to depend on the targeted test results. Based on this information, the following recommendations for choosing a method can be made:

Open Profiling of Quality
OPQ is the method to choose if practitioners target a statistical model of quality factors that can be used to explain users' quality preferences. Open Profiling of Quality may be chosen for evaluations of technological milestones within a system development process to identify the impact of related system parameters on the users' perceived quality in models.

Strengths: Development of statistical models; large set of methods of analysis to study different aspects for broad understanding of the test results; focus on common results as well as individual differences; analysis based on statistics and interpretation of practitioner after all analysis is completed.
Weaknesses: Strength and completeness of the models depends on the abilities of test participants to evaluate quality based on their individual attributes; GPA/MFA analysis builds models on high impact attributes and restrains low impact components for the model.

Post-task Interview Interview-based approaches are a light-weight data collection procedure that can be used to elicit a detailed set of quality factors to understand the general characteristics of quality. Post-task interviews may be chosen at the very beginning of a system development process to be able to identify the range of users' quality factors.

Strengths: Provides a wide and general set of quality factors; can be used for evaluations in the context of use to identify quality factors beyond system-related parameters; light modeling of attributes could be achieved by further analysis (e.g., correspondence analysis).

[2] Within these studies, the OPQ evaluation was done using pen and paper to avoid offering an unintentional 2D reference to the 3D screen on which the stimuli were presented.

Weaknesses: Requires good skills of practitioner for conducting semi-structured interviews; large effort of data analysis; interpretation of data starts during the open coding analysis.

Descriptive Sorted Napping The proposed approach of Descriptive Sorted Napping is a good compromise between OPQ and post-task interview that allows identifying the most crucial quality components in the sorted map and enriching the data with a set of descriptive components. Descriptive Sorted Napping may be chosen for quick insights into dominating quality factors and light hierarchical knowledge about the set of general components.

Strengths: Method can be easily modularized (Sorted Napping only) to offer quick 'from-time-to-time' evaluation of dominating quality factors; easy to understand for users and no extensive development of individual vocabulary needed.
Weaknesses: Nappe limits the dimensions for sorting so that more complex relationships (e.g., audiovisual interactions) may not be identified; abstraction through cards necessary because videos cannot be sorted in the mobile scenario.

7. Discussion and Conclusion

7.1. Summary

The overall goal of this thesis was the development of a validated mixed-methods research approach for audiovisual multimedia quality assessment. The Open Profiling of Quality approach was developed through constructive research in a series of studies in the field of mobile 3D television and video. A total of six experiments were thoroughly reported. More than 300 test participants took part in these evaluations. The results of these studies have been published in 8 main and 12 supplementary scientific publications in international peer-reviewed journals and scientific conferences. The eventually developed method of Open Profiling of Quality was accepted as a proposal for standardized activities of ITU-T SG12 on Q13/12 "QoE, QoS and performance requirements and assessment methods for multimedia including IPTV" in January 2011 [225].

Two research questions formed the basis for the presented work of the thesis. The main research problem targeted the development of a mixed-methods research approach for audiovisual quality evaluations. This methodological approach led to the development of Open Profiling of Quality. The second research question targeted the constructive application of OPQ in the evaluation of mobile 3D television and video systems. The results deepened the knowledge about underlying quality factors of Quality of Experience.

> **"How can quantitative and descriptive data collected in audiovisual quality assessment be combined into a mixed-methods research approach that is applicable for quality evaluations with naïve assessors?"**

The developed Open Profiling of Quality combines standardized quantitative, psychoperceptual quality evaluations and descriptive analysis adapted from Free-Choice Profiling. Based on a literature review, different research approaches to study subjective quality in multimedia systems were identified and contrasted. This review underscored the importance of including the users in modern quality research to arrive at a deeper knowledge about their individual quality perceptions using descriptive evaluation approaches.

Further, the literature review revealed the aspects of multidisciplinary approaches. Based on the socio-scientific, theoretical mixed-methods research model, standardized quantitative methods from multimedia engineering have been combined with descriptive methods stemming from the food sciences in the proposed Open Profiling of Quality approach. Open Profiling of Quality is a mixed-methods approach that combines the evaluation of quality preferences and the elicitation of idiosyncratic experienced quality factors using quantitative psychoperceptual evaluation and, subsequently, an adaptation of Free-Choice Profiling. The two data sets are finally linked by using techniques of External Preference Mapping, which was missing in related descriptive methods [69, 118]. OPQ was constructed under the constraints of good experimental research with respect to validity and reliability.

The theory of sensory analysis demands an application of a new method within a series of studies to validate its applicability and to show its potential [63]. OPQ was developed in constructive research in a series of studies on the subjective quality of mobile 3D television and video. Each study focused on a different aspect of research and underlying research questions varied widely. Beyond the development of Open Profiling of Quality, the methodological work within this thesis was complemented with the development of the Extended-OPQ approach and a comparison model using systematic comparison of mixed-methods research approaches. The developed Extended-OPQ method enables researchers to develop a common set of components of Quality of Experience from the individual quality factors collected in a series of OPQ studies. The set of individual definitions per attribute, which test participants develop during their OPQ sessions, is used in an open-coding approach to form main categories and subcategories of QoE components. The application of qualitative methods of analysis enables researchers to combine the Ext-OPQ approach with other descriptive (interview-based) data sets. The validity of the Ext-OPQ method and its outcome was evaluated in a consensus-vocabulary profiling approach. In this method, the ideas of Quantitative Descriptive Analysis were adapted by substituting the time-consuming vocabulary development with an operationalization of the QoE components developed in the Ext-OPQ. The juxtaposition of OPQ and CP results show good agreement between the methods and underscore the validity of the Ext-OPQ approach and its operationalization for consensus-vocabulary-based quality evaluations.

Concluding the methododological work of the thesis, an extensive between-methods comparison was conducted to increase the knowledge of benefits, applicability, and limitations of OPQ and related mixed-methods approaches. The basis of the comparison was a model that extends the focus of one part (mostly results) of the method to a holistic comparison of aspects of excellence, economy, implementation, and assessment to take into account the whole complexity of a method from planning to analysis [187–189]. This thesis operationalized the model in a first conceptual approach to compare Open Profiling of Quality to an interview-based evaluation and a sorting-based Descriptive Napping approach, which were identified in the literature review as the three main approaches within related mixed-methods quality research [37, 38]. Overall, the comparison shows converging results in terms of identified quality factors. It indicates also the good applicability of OPQ to identify and model individual quality factors (c.f., interview-based methods) in which it does not restrict the dimensions of quality as sorting-based approaches do. The results of the comparison offer guidelines for practitioners to select and use a proper mixed-methods approach with respect to research questions, finally leading to safe long-term development of these methods [9].

A summary of the results of all conducted studies shows converging and complementing results that indicate the validity and reliability of the developed method. The first four studies underscore different aspects of internal, external, and construct validity and show that the method can be applied to study subjective audiovisual quality with naïve participants (section 4.5.3). The results of the OPQ method are reliable internally and externally in a comparison among OPQ studies as well as related work on (mobile) 3DTV quality evaluations, respectively. The thesis includes extensive guidelines for planning, conducting, and analyzing an OPQ study. With respect to validity, a detailed guideline is necessary to avoid instrumentation effects and incorrect applied statistics that may otherwise threaten the internal validity of OPQ. In addition, all studies were conducted by a team of trained assistants or experts so that the threat of researcher expectancies was avoided. Further discussion of validity and reliability can be made based on the Extended-OPQ approach (Study 5, section 5.3) and the comparison of related research methods based on the comparison model (Study 6, section 6.2). The results of the External-OPQ approach and the comparison of OPQ and an operationalized CP method based on consensus vocabulary overcome the threat of mono-method bias and confirm the identified OPQ models for other sensory evaluation approaches. Further, the analysis shows that naïve participants are able to consistently apply fixed attributes to evaluate good and bad video quality although problems in describing crosstalk and ghosting effects still exist. Finally, Study 6 compares three related descriptive evaluation approaches and shows that all methods measure the same construct of quality though differences in the detail of the models contribute to validity and reliability of OPQ in comparison to other methods of the state of the art.

> **"Which are the critical quality factors in mobile 3D video and television, and how do they impact the overall Quality of Experience of the system?"**

Within a constructive research approach of developing Open Profiling of Quality, the author applied OPQ in a series of studies to the field of quality evaluations for mobile 3D television and video. The results of these studies deepened the knowledge about quality aspects of these systems and showed that Open Profiling of Quality can explain quality preferences of naïve participants beyond quantitative ratings.

Most important, the results allow explaining the excellence of mobile 3D video parameters by showing a relationship between video quality and depth perception. The importance of these two components was already identified in the 3D Quality of Experience model by Seuntïens [152], but it is now extended by finding a hierarchical dependency structure in the OPQ models. Descriptions of depth perception are included in the attributes for good video quality within all models, so added value from the 3D perception is only experienced by users when the amount of perceivable visual artifacts is low. Without the sensory extension to quantitative methods, this model was only assumed. In contrast, sensory data alone would not allow mapping users' preferences into the model. The findings were approved by application of other research methods within this thesis as well as in related studies on mobile 3D video [58] that emphasizes the reliability of Open Profiling of Quality.

While the GPA/MFA models of the OPQ studies are dominated by the most salient components, the Ext-OPQ approaches broadened the understanding about underlying quality factors of naïve users for mobile 3D video. This extension was able to identify a set of components for Quality of Experience of mobile 3D video and television that consists of the main components of Visual Quality (Depth, Spatial, Motion), Viewing Experience, Content, and Other Modalities Interactions. The Ext-OPQ approach and the possibilities for joint data analysis of other descriptive data made it possible to derive a general set of QoE components instead of study-dependent characteristics. The results of an operationalization of these components within descriptive evaluations show valid results in comparison to common OPQ models.

Beyond the identification of common quality factors, the results of the OPQ approach show that consumers have different preferences of modalities from which they derive audiovisual quality factors. Especially, the dominance of visual and auditory channels for different users was identified. In addition, the impact of quality factors is different for users. The results show largely differing results for some participants in perception of crosstalk while for others crosstalk leads to an increased 3D experience and, consequently, to high acceptance and quality satisfaction ratings.

7.2. Limitations

The main limitation of the work must be seen in the constructive research approach, which was restricted to the evaluations of mobile 3D video and television. Validity and reliability of the OPQ evaluation depend strongly on the field of research in which the method is applied. Although research questions and quality parameters in the studies varied significantly, the developed models of perceived quality were dominated by specific aspects of 3D video. More general aspects were identified in the Ext-OPQ approach, but still work is needed to generalize the findings of this thesis to Quality of Experience of large-screen 3D television or towards non-stereoscopic 2D television. A second limitation within the practical work arose from the strong emphasis on video within the development of MOBILE3DTV, in which my work for the thesis was embedded. This restricted the audiovisual evaluations in terms of a large set of visual quality parameters and limited the audio parameters being assessed. In addition, the range of perceived qualities within and among studies was quite large so that the assumption that individual profiling may fail due to small impairments and narrow quality ranges could not be studied [89].

In relation to the methodological work, aspects of validity with respect to the ability of users to formulate individual quality attributes need to be taken into account for full validation. Within the development of OPQ, the author identified the importance of thorough introduction of test participants to the sensory evaluation and used the apple task to explain what they were supposed to do. Although this task helped many test participants to understand their task, a large variation in the number of attributes can be found in the studies. Additional screening methods like a verbal fluency test may help to create a common basis of ability to generate quality attributes among test participants [203, 204] for internal validity in terms of test sample selection. Another aspect that needs to be considered is the comparison of the results of OPQ to those given by experts in the domain of 3D video quality. The focus of this study was on the evaluation of quality by naïve participants, but comparable results from naïve assessors and experts further validate that OPQ is applicable for evaluation of quality by naïve participants. With respect to reliability, especially the aspects of internal consistency and interrater reliability need deeper evaluation. Due to the duration of the OPQ method and a multi-session design, test-retest or half-split reliability tests were not conducted within this thesis, thereby limiting the internal consistency of the method. Further, the order of quantitative and sensory evaluation was not varied, which may also affect OPQ's internal reliability. With respect to interrater reliability, comparison of researchers' interpretations of MFA models was exemplarily conducted, but no statistical approaches have yet been identified for calculating Kappa values or similar measures available in other descriptive methods.

7.3. Further work

Further work need to focus on methodology-related work on optimization of the Open Profiling approach. Still, the use of the original OPQ method requires participation of assessors in multiple sessions or long one session designs (Study 3 and Study 4 in section 4). Further research needs to investigate order effects of quantitative and sensory evaluations that may allow for a combined psychoperceptual and sensory evaluation within the same task. Optimized OPQ assessment in terms of duration can also decrease the risk of drop-out rates, which can be a problem of construct validity in multi-session studies [6]. Although this problem did not occur in the presented studies, it is good for practitioners to keep this limitation in mind if considering small sample sizes for the sensory profiling task.

Other aspects for further development of the OPQ method largely concern the sample selection and the test procedure. For multimodal quality evaluation studies with naïve participants, it is worth considering a well-validated tool for identifying the groups of different information processing styles [114]. The author's experiences during the development of OPQ have repeatedly highlighted the importance of training and careful attribute development for sensory evaluation. Individual differences in the ability to describe properties accurately are not only a typically reported challenge in the food sciences [77], but they also seem to be present in multimedia quality studies. Problems can occur when inaccuracy adds noise to the sensory data, which can limit the quality of results [82]. Further work needs to take into account possible improvements of individual's vocabulary by adding supporting tasks beyond the suggested 'apple task' to facilitate assessors' attribute elicitation. Possible methods are the Repertory Grid Method [78–81, 117] or Natural Grouping [205][228]. Other researchers have proposed additional screening tools like a verbal fluency test [203] during the sample selection to take into account different abilities in expressing sensations [204]. However, not only selection of samples but also identification of outliers within the sensory evaluations must be considered. Guidelines for detecting outliers in sensory data are needed. While in quantitative results, outliers can be detected and removed [5, 39], sensory evaluation methods do not provide robust methods that can be applied to that end. However, the residuals given for each configuration after GPA or MFA show large differences between the most important (low residual) and the least important configurations (high residual). These residuals may provide the possibility for outlier detection [206].

Beyond the methodological optimization of OPQ, a set of mixed-methods evaluation tools can also help to identify the most salient individual differences among users and lead to the creation of user profiles that take into account different processing patterns in audiovisual quality perception. In further work, the influence of different processing styles on multimodal quality perception under different quality levels and heterogeneous stimulus material needs to be addressed in detail to confirm the phenomenon. Furthermore, for the practitioners of audiovisual quality, a well-validated tool is needed for identifying the groups of different information processing styles and reporting these groups to characterize a test sample.

Finally, this thesis also proposes further adaptations of the Open Profiling of Quality methods with other descriptive evaluation approaches. The adaptation of Conventional Profiling, the introduction of Descriptive Sorted Napping, and the comparison of OPQ to these related research methods have shown that the selection of a proper research method from the available set of tools needs to be conducted carefully. The proposed comparison model and its first conceptual operationalization have further indicated that extensive between-methods comparisons are needed to identify the abilities and limitations of the different methods. Further work needs to be conducted to guide practitioners towards an optimum selection of research methods that will finally lead to a safe long-term development of the different methods.

The application of Open Profiling of Quality in the field of mobile 3D television has shown that OPQ is a valuable research method for research problems in which a complex field of technical quality factors needs to be evaluated. Especially, the challenge to investigate perceptions that only exist subjectively, like the 3D video perception, can be applicable fields for OPQ evaluations. Comparable problems can be found in evaluation of spatial audio and the questions about subjective envelopment or the evaluation of binaural rendering, for which an explicit reference does not exist because it is purely an effect of individual perception [207, 208]. Further applications could be evaluations of text-to-speech algorithms in which also the problem of interaction of artifacts and the intelligibility of text occurs. For the domain of audiovisual quality evaluations, further applications of Open Profiling of Quality may help to deepen understanding of the interactions of auditory and visual perception. Beyond this thesis, audiovisual (3D) stimuli can be evaluated at high-quality levels to understand the effects of audiovisual integration.

7.4. Conclusion

To conclude, Open Profiling of Quality is a validated tool for mixed-methods evaluations of subjective, audiovisual quality. It has been developed in constructive research in which special care was taken to make the method applicable for evaluations with naïve test participants. Beyond the development of Open Profiling of Quality, the work in this thesis has also shown shortcomings in the methods and proposes adaptations and extensions to address these shortcomings. The Extended-

OPQ approach and the development of a comparison model are two additional important outcomes. Eventually, the proposed methods complement the User-centered Quality of Experience evaluation framework in which the work was embedded.

8. Bibliography

[1] E. Kasanen, K. Lukka, and A. Siitonen, "The Constructive Approach in Management Accounting Research," *Journal of Management Accounting Research*, vol. 5, no. Fall, pp. 243–264, 1993.

[2] S. Haslam and C. McGarty, *Research methods and statistics in psychology.* London, UK: Sage Publications, 2003.

[3] A. Bryman, *Social Research Methods*, 3rd ed. Oxford, UK: Oxford University Press, 2008.

[4] H. Coolican, *Research methods and statistics in psychology.* London: J. W. Arrowsmith Ltd, 2004, vol. 4.

[5] ITU Recommendation ITU-T P.910, *Subjective Video Quality Assessment Methods for Multimedia Applications*, ITU Telecom. Standardization Sector of ITU, Geneva, Switzerland, 1999.

[6] W. Shadish, T. Cook, and D. Campbell, *Experimental and quasi-experimental designs.* Boston, MA: Houghton Mifflin, 2002.

[7] T. D. Cook and D. T. Campbell, *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin Company, 1979.

[8] A. Oulasvirta, "Field experiments in hci: Promises and challenges," in *Future Interaction Design II*, H. Isomäki and P. Saariluoma, Eds. Springer London, 2009, pp. 1–30, doi:10.1007/978-1-84800-385-9_5.

[9] S. Jumisko-Pyykkö, "User-centered quality of experience and its evaluation methods for mobile television," Ph.D. dissertation, Tampere University of Technology, 2011.

[10] Oxford Dictionaries, *Oxford Dictionary of English*, 3rd ed. Oxford University Press, 8 2010.

[11] ISO EN 9000, *Quality management systems – Fundamentals and vocabulary*, International Organization for Standardization, 2005.

[12] ISO EN 9001, *Quality management systems – Requirements*, International Organization for Standardization, 2008.

[13] R. S. Heller, C. D. Martin, N. Haneef, and S. Gievska-Krliu, "Using a theoretical multimedia taxonomy framework," *Journal on Educational Resources in Computing (JERIC)*, vol. 1, p. Article 6, 2001, doi:10.1145/376697.376701.

[14] S. Jumisko-Pyykkö, U. Reiter, and C. Weigel, "Produced quality is not perceived quality - a qualitative approach to overall audiovisual quality," in *Proc. of the 3DTV Conference (3DTV-CON)*, Turkey, 2007.

[15] K. Nahrstedt and R. Steinmetz, "Resource management in networked multimedia systems," *Computer*, vol. 28, no. 5, pp. 52–63, May 1995.

[16] G. Wikstrand, “Improving user comprehension and entertainment in wireless streaming media: Introducing cognitive quality of service,” Department of Computer Science, Umea University, Umea, Sweden, Tech. Rep., 2003.

[17] E. B. Goldstein, *Sensation and perception*, 7th ed. Thomson Wadsworth, Belmont, CA, USA, 2007.

[18] U. Reiter, “Bimodal Audiovisual Perception in Interactive Application Systems of Moderate Complexity,” Ph.D. dissertation, Ilmenau University of Technology, Ilmenau, Germany, 2009.

[19] U. Neisser, *Cognition and reality: principles and implications of cognitive psychology.* San Francisco: W. H. Freeman, 1976.

[20] J. J. Gibson, *The Ecological Approach to Visual Perception.* LAWRENCE ERLBAUM ASSOCIATES, 1979.

[21] S. T. Fiske and S. E. Taylor, *Social Cognition. Singapore.* McGrow-Hill Book Co., 1991.

[22] U. Neisser, “Multiple systems: „a new approach to cognitive theory“,” *European Journal of Cognitive Psychology*, vol. 6, no. 3, pp. 225–241, 1994.

[23] D. S. Hands, “A basic multimedia quality model,” *IEEE Transactions on Multimedia*, vol. 6, no. 6, pp. 806–816, Dec. 2004.

[24] S. Soto-Faraco and A. Kingstone, “Multisensory integration of dynamic information,” in *Handbook of Multisensory Processes*, G. Calvert, C. Spence, and B. Stein, Eds. Cambridge, USA: MIT Press, 2004.

[25] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 746–748, 1976.

[26] M. Coen, “Multimodal Integration – A Biological View,” in *Proc. of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, Seattle, Washington, USA, 2001, pp. 4–10.

[27] R. Guski, *Wahrnehmen – ein Lehrbuch.* Stuttgart, Germany: Kohlhammer, 1996.

[28] R. Welch and D. Warren, “Immediate perceptual response to intersensory discrepancy,” *Psychological Bulletin*, vol. 88, no. 3, pp. 638 – 667, 1980.

[29] A. Peregudov, E. Grinenko, K. Glasman, and A. Belozertsev, “An audiovisual quality model of compressed television materials for portable and mobile multimedia applications,” in *Proc. of the 14th IEEE International Symposium on Consumer Electronics (ISCE2010)*, Braunschweig, Germany, 2010, pp. 1 –6, doi:10.1109/ISCE.2010.5523737.

[30] ITU-T Recommendation P.10 Amendment 1, *Vocabulary for performance and quality of service Amendment 1: New Appendix I - Definition of Quality of Experience (QoE)*, International Telecommunication Union, Geneva, Switzerland, 2008.

[31] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang, “Quality of experience in distributed interactive multimedia environments: toward a theoretical framework,” in *Proc. of the Seventeen ACM International Conference on Multimedia (MULTIMEDIA ’09).* New York, NY, USA: ACM, 2009, pp. 481–490.

[32] R. B. Johnson and A. J. Onwuegbuzie, “Mixed methods research: A research paradigm whose time has come,” *Educational Researcher*, vol. 33, no. 7, pp. 14–26, 2004.

[33] R. B. Johnson, A. J. Onwuegbuzie, and L. A. Turner, "Toward a Definition of Mixed Methods Research," *Journal of Mixed Methods Research*, vol. 1, no. 2, pp. 112–133, 2007, doi:10.1177/1558689806298224.

[34] J. W. Creswell and V. L. Plano Clark, *Designing and Conducting Mixed Methods Research.* Thousand Oaks, California, USA: Sage Publications, 2006.

[35] A. Tashakkori and C. Teddlie, "Quality of inferences in mixed methods research: Calling for an integrative framework," in *Advances in Mixed Methods Research*, M. M. Bergman, Ed. London: Sage, 2008.

[36] N. K. Denzin, *The research act: An introduction to sociological methods.* New York: McGraw-Hill, 1978.

[37] S. Jumisko-Pyykkö, J. Häkkinen, and G. Nyman, "Experienced quality factors qualitative evaluation approach to audiovisual quality," in *Proc. of the IS&T/SPIE 19th Annual Symposium of Electronic Imaging*, 2008.

[38] J. Radun, T. Leisti, J. Häkkinen, H. Ojanen, J. L. Olives, T. Vuori, and G. Nyman, "Content and quality: Interpretation-based estimation of image quality." *ACM Transactions on Applied Perception (TAP)*, vol. 4, no. 4, pp. 1–15, January 2008.

[39] ITU Recommendation ITU-R BT.500-11, *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU Telecom. Standardization Sector of ITU, Geneva, Switzerland, 2002.

[40] H. T. Lawless and H. Heymann, *Sensory evaluation of food: principles and practices*, 1st ed. New York: Chapman & Hall, 1999.

[41] P. Engeldrum, *Psychometric scaling: a toolkit for imaging systems development.* Winchester, Mass: Imcotek Press, 2000.

[42] F. Kozamernik, P. Sunna, E. Wyckens, and D. I. Pettersen, "Samviq: Subjective quality of internet: Video codecs - phase2 evaluations using samviq," European Broadcasting Union, Tech. Rep., 2005.

[43] ITU Recommendation ITU-R BT.1438, *Subjective Assessment of Stereoscopic Television Pictures*, ITU Telecom. Standardization Sector of ITU, Geneva, Switzerland, 2000.

[44] ITU Recommendation ITU-T P.911, *Subjective Audiovisual Quality Assessment Methods for Multimedia Applications*, ITU Telecom. Standardization Sector of ITU, Geneva, Switzerland, 2000.

[45] ISO EN 8586-2, *Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 2: Experts*, International Organization for Standardization, 1994.

[46] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application.* Chichester, England: Wiley, April 2006.

[47] ITU Recommendation ITU-T P.831, *Subjective Performance Evaluation of Network Echo Cancellers*, ITU Telecom. Standardization Sector of ITU, Geneva, Switzerland, 1998.

[48] EBU BPN 056, "Subjective assessment methodology for video quality," European Broadcasting Union, Project Group B/VIM, Tech. Rep., 2003.

[49] J. L. Blin, "New Quality Evaluation Method Suited to Multimedia Context," in *Proc. of the Second International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Scottsdale, Arizona, USA, January 2006.

[50] M. D. Brotherton, Q. Huynh-Thu, D. S. Hands, and K. Brunnström, "Subjective multimedia quality assessment," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E89-A, no. 11, pp. 2920–2932, 2006.

[51] D. Rouse, R. Pepion, P. Le Callet, and S. Hemami, "Tradeoffs in Subjective Testing Methods for Image and Video Quality Assessment," in *Proc. of Human Vision and Electronic Imaging XV*, B. E. Rogowitz and T. N. Pappas, Eds., vol. 7527, no. 1. SPIE, 2010, doi:10.1117/12.845389.

[52] S. R. Gulliver, T. Serif, and G. Ghinea, "Pervasive and standalone computing: the perceptual effects of variable multimedia quality," *International Journal of Human-Computer Studies*, vol. 60, no. 5-6, pp. 640–665, 2004.

[53] S. R. Gulliver and G. Ghinea, "Stars in their eyes: what eye-tracking reveals about multimedia perceptual quality," *IEEE Transaction on System, Man and Cybernetics, Part A.*, vol. 34, no. 4, pp. 472 – 482, 2004.

[54] G. Ghinea and J. P. Thomas, "Qos impact on user perception and understanding of multimedia video clips," in *Proc. of the sixth ACM international conference on Multimedia (MULTIMEDIA '98).* New York, NY, USA: ACM, 1998, pp. 49–54.

[55] S. R. Gulliver and G. Ghinea, "Defining user perception of distributed multimedia quality," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 2, no. 4, pp. 241 – 257, 2006.

[56] J. D. McCarthy, M. A. Sasse, and D. Miras, "Sharp or smooth?: Comparing the effect of quantization vs. frame rate for streamed video," in *Proc. of ACM Conference on Human Factors in Computing Systems (CHI 2004)*, E. Dykstra-Erickson and M. Tscheligi, Eds. Vienna, Austria: ACM Press, 2004, pp. 535 – 542, doi:10.1145/985692.985760.

[57] S. Jumisko-Pyykkö, V. Kumar Malamal Vadakital, and M. M. Hannuksela, "Acceptance threshold: Bidimensional research method for user-oriented quality evaluation studies," *International Journal of Digital Multimedia Broadcasting*, vol. 2008, p. Article ID 712380, 2008, doi:10.1155/2008/712380.

[58] S. Jumisko-Pyykkö and T. Utriainen, "A Hybrid Method for Quality Evaluation in the Context of Use for Mobile (3D) Television," *Multimedia Tools and Applications*, vol. 'Online First', pp. 1–41, 2010, doi:10.1007/s11042-010-0573-4.

[59] S. Jumisko-Pyykkö and M. M. Hannuksela, "Does context matter in quality evaluation of mobile television?" in *Proc. of the 10th international conference on Human computer interaction with mobile devices and services (MobileHCI '08).* New York, NY, USA: ACM, 2008, pp. 63–72.

[60] S. Jumisko-Pyykkö and T. Utriainen, "User-centered quality of experience of mobile 3dtv: how to evaluate quality in the context of use?" in *Proc. of 'Multimedia on Mobile Devices', a part of the Electronic Imaging Symposium*, R. Creutzburg and D. Akopian, Eds., vol. Proc. SPIE, Vol. 7542, 75420W, San Jose, California, USA, January 2010.

[61] S. Jumisko-Pyykkö and T. Vainio, “Framing the Context of Use for Mobile HCI,” *International Journal of Mobile Human Computer Interaction (IJMHCI)*, vol. 2, no. 4, pp. 1 – 28, 2010.

[62] A. Dey, G. Abowd, and D. Salber, “A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications,” *Human-Computer Interaction*, vol. 16, no. 2, pp. 97 – 166, December 2001, doi:10.1207/S15327051HCI16234_02.

[63] H. Stone and J. L. Sidel, *Sensory evaluation practices*, 3rd ed. San Diego: Academic Press, 2004.

[64] H. Knoche, J. D. McCarthy, and M. A. Sasse, “Can small be beautiful?: assessing image resolution requirements for mobile tv,” in *Proc. of the 13th annual ACM international conference on Multimedia (MULTIMEDIA ’05).* New York, NY, USA: ACM, 2005, pp. 829–838.

[65] H. Knoche, “Quality of experience in digital mobile multimedia services,” Ph.D. dissertation, University College London, London, UK, 2010.

[66] A. Strauss and J. Corbin, *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Thousand Oaks, CA: Sage, 1998, vol. 2.

[67] G. Nyman, J. Radun, T. Leisti, J. Oja, H. Ojanen, J. Olives, T. Vuori, and J. Häkkinen, “What do users really perceive: probing the subjective image quality,” in *Proc. SPIE*, vol. Vol. 6059, 605902, 2006.

[68] J. R. Piggott, S. J. Simpson, and S. A. R. Williams, “Sensory analysis,” *International Journal of Food Science & Technology*, vol. 33, no. 1, pp. 7–12, 1998, doi:10.1046/j.1365-2621.1998.00154.x.

[69] S. Bech, R. Hamberg, M. Nijenhuis, C. Teunissen, H. de Jong, P. Houben, and S. Pramanik, “Rapid perceptual image description (RaPID) method,” in *Proc. of the SPIE Human Vision and Electronic Imaging*, B. E. Rogowitz and J. P. Allebach, Eds., vol. 2657, no. 1. SPIE, 1996, doi:10.1117/12.238728.

[70] H. Stone, J. Sidel, S. Oliver, A. Woolsey, and R. Singleton, “Sensory evaluation by quantitative descriptive analysis,” *Food technology*, vol. 28, pp. 24–34, 1974.

[71] C. Civille and H. Lawless, “The Importance of Language in Describing Perceptions,” *Journal of Sensory Studies*, vol. 1, pp. 203 – 215, 1986.

[72] J. Piggott, “Selection of terms for descriptive analysis,” in *Sensory science theory and applications in foods*, ser. IFT basic symposium series, H. Lawless and B. Klein, Eds. New York, USA: Dekker, 1991.

[73] H. Stone, J. Sidel, S. Oliver, A. Woolsey, and R. Singleton, “Sensory evaluation by quantitative descriptive analysis,” in *Descriptive Sensory Analysis in Practice*, M. Gacula Jr., Ed. Trumbull, Conneticut, USA: Food & Nutrition Press, Inc., 1997.

[74] A. A. Williams and S. P. Langron, “The use of free-choice profiling for the evaluation of commercial ports,” *Journal of the Science of Food and Agriculture*, vol. 35, no. 5, pp. 558–568, 1984.

[75] F. Jack and J. Piggott, “Free choice profiling in consumer research,” *Food quality and preference*, vol. 153, no. 3, pp. 129–134, 1991-1992.

[76] Gower, “Generalized procrustes analysis,” *Psychometrika*, vol. 40, no. 1, pp. 33–51, March 1975.

[77] J. A. McEwan, J. S. Colwill, and D. M. H. Thomson, "The application of two free-choice profile methods to investigate the sensory characteristics of chocolate," *Journal of Sensory Studies, Food & Nutrition Press*, vol. 3, pp. 271–286, 1989.

[78] G. A. Kelly, *The psychology of personal constructs*. Norton, New York, 1955.

[79] D. Thomson and J. McEwan, "An application of the repertory grid method to investigate consumer perceptions of foods," *Appetite*, vol. 10, no. 3, pp. 181 – 193, 1988, doi:10.1016/0195-6663(88)90011-6.

[80] N. Gains and D. Thomson, "Contextual evaluation of canned lagers using repertory grid method," *International Journal of Food Science & Technology*, vol. 25, no. 6, pp. 699–705, 1990, doi:10.1111/j.1365-2621.1990.tb01131.x.

[81] N. Gains, "The repertory grid approach," in *Measurement of Food Preferences*, H. MacFie and D. Thomson, Eds. Blackie Academic and Professional, 1994, pp. 51 – 76.

[82] J. Piggott and M. Watson, "A comparison of free-choice profiling and the repertory grid method in the flavor profiling of cider," *Journal of Sensory Studies*, vol. 7, no. 2, pp. 133 – 145, 1992.

[83] C. M. Delahunty, A. McCord, E. E. O'Neill, and P. A. Morrissey, "Sensory characterisation of cooked hams by untrained consumers using freechoice profiling," *Food Quality and Preference*, vol. 8, no. 5-6, pp. 381–388, September-November 1997.

[84] A. A. Williams and G. M. Arnold, "Comparison of the aromas of six coffees characterized by conventional profiling, free-choice profiling and similarity scaling methods," *Journal of the Science of Food and Agriculture*, vol. 36, no. 3, pp. 204 –214, 1985, doi:10.1002/jsfa.2740360311.

[85] N. Gains and D. M. H. Thomson, "Sensory profiling of canned lager beers using consumers in their own homes," *Food Quality and Preference*, vol. 2, no. 1, pp. 39 – 47, 1990, doi:10.1016/0950-3293(90)90029-T.

[86] P. N. Jones, H. J. H. McFie, and S. L. Beilken, "Use of preference mapping to relate consumer preference to the sensory properties of a processed meat product (tinned cat food)." *Journal of the Science of Food and Agriculture*, vol. 47, no. 1, pp. 113–123, 1989.

[87] C. Gomez, F. Fiorenza, L. Izquierdo, and E. Costell, "Perception of mealiness in apples: a comparison of consumers and trained assessors," *Zeitschrift für Lebensmitteluntersuchung und -Forschung A*, vol. 207, pp. 304–310, 1998, doi:10.1007/s002170050337.

[88] F. Husson, S. L. Dien, and J. Pagès, "Which value can be granted to sensory profiles given by consumers? methodology and results," *Food Quality and Preference*, vol. 12, no. 5-7, pp. 291 – 296, 2001, doi:10.1016/S0950-3293(01)00014-3.

[89] E. Cristovam, A. Paterson, and J. R. Piggott, "Differentiation of port wines by appearance using a sensory panel: comparing free choice and conventional profiling," *European Food Research and Technology*, vol. 211, pp. 65–71, 2000, doi:10.1007/s002170050590.

[90] H. Abdi and D. Valentin, "Some new and easy ways to describe, compare, and evaluate products and assessors," in *New trends in sensory evaluation of food and non-food products*, D. Valentin, D. Nguyen, and L. Pelletier, Eds. Ho Chi Minh, Vietnam: Vietnam National University-Ho chi Minh City Publishing House, 2007, pp. 5–18.

[91] P. Faye, D. Brémaud, M. D. Daubin, P. Courcoux, A. Giboreau, and H. Nicod, “Perceptive free sorting and verbalization tasks with naïve subjects: an alternative to descriptive mappings,” *Food Quality and Preference*, vol. 15, no. 7-8, pp. 781–791, 2004.

[92] D. Picard, C. Dacremont, D. Valentin, and A. Giboreau, “Perceptual dimensions of tactile textures,” *Acta Psychologica*, vol. 114, no. 2, pp. 165–184, October 2003.

[93] R. Cartier, A. Rytz, A. Lecomte, F. Poblete, J. Krystlik, E. Belin, and N. Martin, “Sorting procedure as an alternative to quantitative descriptive analysis to obtain a product sensory map,” *Food Quality and Preference*, vol. 17, no. 7-8, pp. 562 – 571, 2006, doi:10.1016/j.foodqual.2006.03.020.

[94] E. Risvik, J. A. McEwan, J. S. Colwill, R. Rogers, and D. H. Lyon, “Projective mapping: A tool for sensory analysis and consumer research,” *Food Quality and Preference*, vol. 5, no. 4, pp. 263 – 269, 1994, doi:10.1016/0950-3293(94)90051-5.

[95] E. Risvik, J. A. McEwan, and M. Rødbotten, “Evaluation of sensory profiling and projective mapping data,” *Food Quality and Preference*, vol. 8, no. 1, pp. 63 – 71, 1997, doi:10.1016/S0950-3293(96)00016-X.

[96] J. Pagès, “Collection and analysis of perceived product inter-distances using multiple factor analysis: Application to the study of 10 white wines from the Loire Valley,” *Food Quality and Preference*, vol. 16, no. 7, pp. 642 – 649, 2005, doi:10.1016/j.foodqual.2005.01.006.

[97] J. Pagès, M. Cadoret, and S. Le, “The sorted napping: A new holistic approach in sensory evaluation,” *Journal of Sensory Studies*, vol. 25, no. 5, pp. 637–658, 2010, doi:10.1111/j.1745-459X.2010.00292.x.

[98] H. Abdi, D. Valentin, C. Chrea, and S. Chollet, “Analyzing assessors and products in sorting tasks: DISTATIS, theory and applications,” *Food Quality and Preference*, vol. 18, pp. 627–640, 2007.

[99] B. Escofier and J. Pagès, “Multiple factor analysis (afmult package),” *Computational Statistics & Data Analysis*, vol. 18, no. 1, pp. 121 – 140, 1994.

[100] P. Faye, D. Bremaud, E. Teillet, P. Courcoux, A. Giboreau, and H. Nicod, “An alternative to external preference mapping based on consumer perceptive mapping,” *Food Quality and Preference*, vol. 17, no. 7-8, pp. 604–614, 2006.

[101] H. Lawless and H. Heymann, *Sensory Evaluation of Food: Principles and Practices*, 2nd ed. New York: Springer, 2010.

[102] M. Santosa, H. Abdi, and J.-X. Guinard, “A modified sorting task to investigate consumer perceptions of extra virgin olive oils,” *Food Quality and Preference*, vol. 21, no. 7, pp. 881 – 892, 2010, doi: 10.1016/j.foodqual.2010.05.011.

[103] S. V. Kirkmeyer and B. J. Tepper, “Understanding Creaminess Perception of Dairy Products Using Free-Choice Profiling and Genetic Responsivity to 6-n-Propylthiouracil,” *Chemical Senses*, vol. 28, no. 6, pp. 527–536, 2003, doi:10.1093/chemse/28.6.527.

[104] M. Lelièvre, S. Chollet, H. Abdi, and D. Valentin, “Beer-Trained and Untrained Assessors Rely More on Vision than on Taste When They Categorize Beers,” *Chemosensory Perception*, vol. 2, pp. 143–153, 2009, doi:10.1007/s12078-009-9050-8.

[105] I. Jaime, D. J. Mela, and N. Bratchell, “A Study of Texture-Flavor Interactions using Free-Choice Profiling,” *Journal of Sensory Studies*, vol. 8, no. 3, pp. 177–188, 1993, doi:10.1111/j.1745-459X.1993.tb00212.x.

[106] T. Worch, S. Lê, and P. Punter, “How reliable are the consumers? comparison of sensory profiles from consumers and experts,” *Food Quality and Preference*, vol. 21, no. 3, pp. 309 – 318, 2010, doi: 10.1016/j.foodqual.2009.06.001.

[107] K. A. Moussaoui and P. Varela, “Exploring consumer product profiling techniques and their linkage to a quantitative descriptive analysis,” *Food Quality and Preference*, vol. 21, no. 8, pp. 1088 – 1099, 2010, doi:10.1016/j.foodqual.2010.09.005.

[108] M. Lelièvre, S. Chollet, H. Abdi, and D. Valentin, “What is the validity of the sorting task for describing beers? a study using trained and untrained assessors,” *Food Quality and Preference*, vol. 19, no. 8, pp. 697 – 703, 2008, doi:10.1016/j.foodqual.2008.05.001.

[109] S. L. Dien and J. Pagès, “Hierarchical multiple factor analysis: application to the comparison of sensory profiles,” *Food Quality and Preference*, vol. 14, no. 5-6, pp. 397 – 403, 2003, the Sixth Sense - 6th Sensometrics Meeting.

[110] C. Narain, A. Paterson, and E. Reid, “Free choice and conventional profiling of commercial black filter coffees to explore consumer perceptions of character,” *Food Quality and Preference*, vol. 15, no. 1, pp. 31 – 41, 2004, doi:10.1016/S0950-3293(03)00020-X.

[111] J. Pagès and F. Husson, “Inter-laboratory comparison of sensory profiles: methodology and results,” *Food Quality and Preference*, vol. 12, no. 5-7, pp. 297 – 309, 2001.

[112] S. Lê, J. Pagès, and F. Husson, “Methodology for the comparison of sensory profiles provided by several panels: Application to a cross-cultural study,” *Food Quality and Preference*, vol. 19, no. 2, pp. 179 – 184, 2008, doi:10.1016/j.foodqual.2007.04.008.

[113] P. M. Brockhoff and I. M. Skovgaard, “Modelling individual differences between assessors in sensory evaluations,” *Food Quality and Preference*, vol. 5, no. 3, pp. 215 – 224, 1994, doi:10.1016/0950-3293(94)90037-X.

[114] T. L. Childers, M. J. Houston, and S. E. Heckler, “Measurement of individual differences in visual versus verbal information processing,” *Journal of Consumer Research*, vol. 12, no. 2, pp. 125–134, 1985.

[115] N. Zacharov and K. Koivuniemi, “Audio descriptive analysis & mapping of spatial sound displays,” in *Proc. of the 7th International Conference on Auditory Display*, J. Hiipakka, N. Zacharov, and T. Takala, Eds. Laboratory of Acoustics and Audio Signal Processing and the Telecommunications Software and Multimedia Laboratory, Helsinki University of Technology, 2001, pp. 95–104.

[116] G. Lorho, “Individual vocabulary profiling of spatial enhancement systems for stereo headphone reproduction,” in *Proc. of Audio Engineering Society 119th Convention*, vol. Convention Paper 6629, New York (NY), USA, 2005.

[117] ——, “Perceptual evaluation of mobile multimedia loudspeakers,” in *Proc. of Audio Engineering Society 122th Convention*, 2007.

[118] ——, “Perceived quality evaluation: An application to sound reproduction over headphones,” Ph.D. dissertation, Aalto University, School of Science and Technology, Espoo, Finland, 2010.

[119] J. Berg and F. Rumsey, “Spatial Attribute Identification and Scaling by Repertory Grid Technique and other methods,” in *Proc. of the AES 16th International Conference on Spatial Sound Reproduction*, 1999, pp. 51–66.

[120] J. Delarue and J. M. Sieffermann, “Sensory mapping using flash profile. Comparison with a conventional descriptive method for the evaluation of the fiavour of fruit dairy products,” *Food Quality and Preference*, vol. 15, no. 4, pp. 383–392, June 2004.

[121] J. Radun, T. Leisti, T. Virtanen, J. Häkkinen, T. Vuori, and G. Nyman, “Evaluating the multivariate visual quality performance of image-processing components,” *ACM Transactions on Applied Perception*, vol. 7, pp. 1–16, June 2008, doi:10.1145/1773965.1773967.

[122] J. Häkkinen, T. Kawai, J. Takatalo, T. Leisti, J. Radun, A. Hirsaho, and G. Nyman, “Measuring stereoscopic image quality experience with interpretation based quality methodology,” in *Image Quality and System Performance V*, vol. 6808, 68081B, Januar 2008.

[123] T. Shibata, S. Kurihara, T. Kawai, T. Takahashi, T. Shimizu, R. Kawada, A. Ito, J. Häkkinen, J. Takatalo, and G. Nyman, “Evaluation of stereoscopic image quality for mobile devices using interpretation based quality methodology,” in *Stereoscopic Displays and Applications XX*, vol. 7237 72371E, February 2009.

[124] Thesaurus, “Thesaurus online dictionary,” Website, 2010, available online at http://www.thesaurus.com; visited on 09.07.2010.

[125] Oxford University Press, “The Oxford Dictionary of English, Revised Edition,” Website, 2005, available online at http://www.thesaurus.com; visited on March 15, 2011.

[126] ISO EN 8586-1, *Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 1: Selected assessors*, International Organization for Standardization, 1993.

[127] ISO EN 7029, *Statistical distribution of hearing threshold as a function of age*, International Standardization Organization, 2000.

[128] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Erlbaum Associates, 1988.

[129] J. A. McEwan, “Preference mapping for product optimization,” in *Multivariate analysis of data in sensory science*, T. Naes and E. Risvik, Eds. Amsterdam, The Netherlands: Elsevier, 1996, pp. 185–220.

[130] M. S. Lewis-Beck, A. Bryman, and T. F. Liao (eds.), *The SAGE encyclopedia of social science research methods.* SAGE Publications, Inc., 2004.

[131] J. M. Murray, C. M. Delahunty, and I. A. Baxter, “Descriptive sensory analysis: past, present and future,” *Food Research International*, vol. 34, no. 6, pp. 461–471, 2001.

[132] D. C. Oreskovich, B. P. Klein, and J. W. Sutherland, “Procrustes analysis and its applications to free choice and other sensory profiling,” in *Sensory science theory and applications in foods*, H. T. Lawless and B. P. Klein, Eds. New York: Marcel Dekker, 1991, pp. 353–394.

[133] D. Dijksterhuis, "Procrustes analysis in sensory research," in *Multivariate analysis of data in sensory science*, T. Naes and E. Risvik, Eds. Amsterdam, The Netherlands: Elsevier, 1996, pp. 185–220.

[134] J. Pagès, "Analyse factorielle multiple et analyse procustéenne," *Revue Statistique appliquée*, vol. 53, no. 4, pp. 61–86, 2005.

[135] G. B. Dijksterhuis and J. C. Gower, "The interpretation of generalized procrustes analysis and allied methods," *Food Quality and Preference*, vol. 3, no. 2, pp. 67 – 87, 1991-1992, doi:10.1016/0950-3293(91)90027-C.

[136] J. Kunert and E. Qannari, "A simple alternative to generalized procrustes analysis: application to sensory profiling data," *Journal of sensory studies*, vol. 14, no. 2, pp. 197–208, 1999.

[137] H. Abdi and D. Valentin, "Multiple factor analysis," in *Encyclopedia of Measurement and Statistics*, N. Salkind, Ed. Thousand Oaks (CA): Sage, 2007, pp. 657 – 663.

[138] J. Pagès, "Multiple factor analysis: Main features and application to sensory data," *Revista Colombiana de Estadística*, vol. 27, no. 1, pp. 1–26, 2004.

[139] J. Pagès and M. Tenenhaus, "Multiple factor analysis combined with pls path modelling. application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgements," *Chemometrics and Intelligent Laboratory Systems*, vol. 58, no. 2, pp. 261 – 273, 2001.

[140] T. Lokki and K. Puolamäki, "Canonical analysis of individual vocabulary profiling data," in *Proc. of the International Workshop on Quality of Multimedia Experience (QoMEX2010)*, Trondheim, Norway, June 2010.

[141] L. Perrin, R. Symoneaux, I. Maître, C. Asselin, F. Jourjon, and J. Pagès, "Comparison of three sensory methods for use with the napping procedure: Case of ten wines from loire valley," *Food Quality and Preference*, vol. 19, no. 1, pp. 1 – 11, 2008, doi:10.1016/j.foodqual.2007.06.005.

[142] G. B. Dijksterhuis and W. J. Heiser, "The role of permutation tests in exploratory multivariate data analysis," *Food Quality and Preference*, vol. 6, no. 4, pp. 263 – 270, 1995, doi:10.1016/0950-3293(95)00025-9.

[143] I. Wakeling, M. Raats, and H. MacFie, "A new significance test for consensus in generalized procrustes analysis," *Journal of Sensory Studies*, vol. 7, no. 2, pp. 91–96, 1992, doi:10.1111/j.1745-459X.1992.tb00526.x.

[144] B. King and P. Arents, "A statistical test of consensus obtained from generalised Procrustes analysis of sensory data," *Journal of Sensory Studies*, vol. 6, pp. 37–48, 1991.

[145] M. Tenenhaus, J. Pagès, L. Ambroisine, and C. Guinot, "Pls methodology to study relationships between hedonic judgements and product characteristics," *Food Quality and Preference*, vol. 16, no. 4, pp. 315 – 325, 2005.

[146] H. Abdi, "Partial least squares regression and projection on latent structure regression (PLS Regression)," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 1, pp. 97–106, January/February 2010.

[147] P. Schlich, “Preference mapping: Relating consumer preferences to sensory or instrumental measurements,” in *Bioflavour 95*, P. Etievant and P. Schreiner, Eds. Versailles: INRA Editions, 1995.

[148] H. Martens and T. Naes, *Multivariate Calibration.* London: John Wiley & Sons, 1994.

[149] W. A. IJsselsteijn, D. G. Bouwhuis, J. Freeman, and H. de Ridder, “Presence as an experiential metric for 3-d display evaluation,” *SID Symposium Digest of Technical Papers*, vol. 33, no. 1, pp. 252–255, 2002. [Online]. Available: http://link.aip.org/link/?SYM/33/252/1

[150] L. Onural and H. M. Ozaktas, “Three-dimensional television: from science-fiction to reality,” in *Three-Dimensional Television: Capture, Transmission, Display*, H. M. Ozaktas and L. Onural, Eds. Berlin, Germany: Springer, 2007.

[151] A. Boev, D. Hollosi, A. Gotchev, and K. Egiazarian, “Classification and simulation of stereoscopic artifacts in mobile 3dtv content,” in *Proc. of SPIE Stereoscopic Displays and Applications XX*, A. J. Woods, N. S. Holliman, and J. O. Merritt, Eds., vol. 7237, no. 1. SPIE, 2009, p. 72371F, doi:10.1117/12.807185.

[152] P. J. H. Seuntïens, “Visual experience of 3d tv,” Ph.D. dissertation, Technische Universiteit Eindhoven, 2006.

[153] R. G. Kaptein, A. Kuijsters, M. T. M. Lambooij, W. A. IJsselsteijn, and I. Heynderickx, “Performance evaluation of 3D-TV systems,” in *Proc. of SPIE Image Quality and System Performance V*, S. P. Farnand and F. Gaykema, Eds., vol. 6808, no. 1. San Jose, CA, USA: SPIE, 2008, p. 680819. [Online]. Available: http://link.aip.org/link/?PSI/6808/680819/1

[154] L. Stelmach, W. J. Tam, D. Meegan, and A. Vincent, “Stereo image quality: Effects of mixed spatiotemporal resolution,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, pp. 188–193, 2000.

[155] L. B. Stelmach, W. J. Tam, D. V. Meegan, A. Vincent, and P. Corriveau, “Human perception of mismatched stereoscopic 3d inputs,” in *Proc. of the International Conference on Image Processing (ICIP)*, vol. 1, 2000, pp. 5–8.

[156] W. Tam, L. Stelmach, and P. Corriveau, “Psychovisual aspects of viewing stereoscopic video sequence,” in *Stereoscopic Displays and Virtual Reality Systems V*, M. T. Bolas, S. S. Fisher, and J. O. Merritt, Eds., vol. 3295. San Jose, CA, USA: SPIE, 1998, pp. 226–235.

[157] A. Kuijsters, W. A. Ijsselsteijn, M. T. M. Lambooij, and I. E. J. Heynderickx, “Influence of chroma variations on naturalness and image quality of stereoscopic images,” in *Proc. of Human Vision and Electronic Imaging XIV*, B. E. Rogowitz and T. N. Pappas, Eds., vol. 7240, no. 1. San Jose, CA, USA: SPIE, 2009, p. 72401E, doi:10.1117/12.817749.

[158] M. Barkowsky, R. Cousseau, and P. Le Callet, “Influence of depth rendering on the quality of experience for an autostereoscopic display,” in *Proc. of the First International Workshop on Quality of Multimedia Experience (QoMEX 2009)*, 2009, pp. 192 –197, doi:10.1109/QOMEX.2009.5246954.

[159] M. Barkowsky, P. Campisi, P. Le Callet, and V. Rizzo, “Crosstalk measurement and mitigation for autostereoscopic displays,” in *Proc. of Three-Dimensional Image Processing (3DIP) and Applications*, A. M. Baskurt, Ed., vol. 7526, no. 1. San Jose, CA, USA: SPIE, 2010, p. 75260R, doi:10.1117/12.839184.

[160] M. Lambooij, M. Fortuin, W. A. Ijsselsteijn, and I. Heynderickx, "Measuring visual discomfort associated with 3d displays," in *Proc. of SPIE Stereoscopic Displays and Applications XX*, A. J. Woods, N. S. Holliman, and J. O. Merritt, Eds., vol. 7237. SPIE, 2009.

[161] M. Lambooij, W. A. IJsselsteijn, M. Fortuin, and I. Heynderickx, "Visual discomfort in stereoscopic displays: a review," *Journal of Imagng Science and Technology*, vol. 53, no. 3, pp. 1–14, 2009.

[162] J. Häkkinen, M. Pölönen, J. Takatalo, and G. Nyman, "Simulator sickness in virtual display gaming: a comparison of stereoscopic and non-stereoscopic situations," in *Proc. of the 8th Conference on Human-Computer interaction with Mobile Devices and Services (MobileHCI '06)*, vol. 159. ACM, 2006, pp. 227–230.

[163] R. Storms, "Auditory-visual cross-modal perception phenomena," Ph.D. dissertation, Naval Postgraduate School, Monterey California, 1998.

[164] U. Reiter and U. Kühhirt, "Object-based A/V application systems: IAVAS I3D status and overview," in *Proc. of IEEE International Symposium on Consumer Electronics (ISCE)*, Irving, Texas, USA, June 2007.

[165] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *The International Journal of Aviation Psychology*, vol. 3, no. 3, pp. 203–220, 1993.

[166] W. Ijsselsteijn, H. de Ridder, and J. Vliegen, "Subjective evaluation of stereoscopic images: effects of camera parameters and display duration," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 225–233, 2000.

[167] G. Tech, A. Smolic, H. Brust, P. Merkle, K. Dix, Y. Wang, K. Müller, and T. Wiegand, "Optimization and comparison of coding algorithms for mobile 3dtv," in *3DTV Conference: The True Vision Capture, Transmission and Display of 3D Video, 2009*, 6 2009, pp. 1–4.

[168] ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), *Advanced Video Coding for Generic Audiovisual Services*, ITU-T and ISO/IEC JTC 1, 2007.

[169] *Text of ISO/IEC 14496-10:200X/FDAM 1 Multiview Video Coding. Doc. N9978*, ISO/IEC JTC1/SC29/WG11, Hannover, Germany, 2008.

[170] H. Brust, A. Smolic, K. Mueller, G. Tech, and T. Wiegand, "Mixed resolution coding of stereoscopic video for mobile devices," in *The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON 2009)*. Potsdam, Germany: Institute of Electrical and Electronics Engineers (IEEE), 2009.

[171] *ISO/IEC CD 23002-3: Representation of auxiliary video and supplemental information. Doc. N8259*, ISO/IEC JTC1/SC29/WG11, Klagenfurt, Austria, 2007.

[172] N. Atzpadin, P. Kauff, and O. Schreer, "Stereo Analysis by Hybrid Recursive Matching for RealTime Immersive Video Conferencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 3, pp. 321–334, March 2004.

[173] P. Merkle, Y. Wang, K. Müller, A. Smolic, and T. Wiegand, "Video plus depth compression for mobile 3d services," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*, 4-6, 2009, pp. 1–4.

[174] G. Tech, H. Brust, K. Müller, A. Aksay, and D. Bugdayci, “D2.5 development and optimization of coding algorithms for mobile 3dtv,” Mobile3DTV, Project No. 216503, Tech. Rep., 2009.

[175] S. I. Uehara, T. Hiroya, H. Kusanagi, K. Shigemura, and H. Asada, “1-inch diagonal transflective 2d and 3d lcd with hddp arrangement,” in *Stereoscopic Displays and Applications XIX, Proc. of SPIE-IS&T Electronic Imaging 2008*, S. Displays and A. XIX, Eds., vol. 6803, San Jose, CA, USA, February 2008, conference Chairs: Andrew J. Woods, Nicolas S. Holliman, John O. Merritt.

[176] M. O. Bici, D. Bugdayci, G. B. Akar, and A. Gotchev, “Mobile 3D video broadcast,” in *Proc. of the International Conference on Image Processing (ICIP ’10)*, Hong Kong, China, 2010, pp. 2397 – 2400.

[177] G. Faria, J. Henriksson, E. Stare, and P. Talmola, “DVB-H: Digital broadcast services to handheld devices,” *Proceedings of the IEEE*, vol. 94, no. 1, pp. 194–209, 2006.

[178] ETSI TR 102 377 V1.3.1., *Digital Video Broadcasting (DVB): DVB-H implementation guidelines*, European Telecommunications Standards Institute ETSI, 2009.

[179] “FATCAPS: A Free, Linux-Based Open-Source DVB-H IP-Encapsulator,” available online: http://amuse.ftw.at/downloads/encapsulator.

[180] “DECAPS: DVB-H Decapsulator Software,” available online: http://sp.cs.tut.fi/mobile3dtv/download/.

[181] S. Lê, J. Josse, and F. Husson, “Factominer: an r package for multivariate analysis,” *Journal of Statistical Software*, vol. 25, pp. 1–18, 2008.

[182] R Development Core Team, “R: A Language and Environment for Statistical Computing,” R Foundation for Statistical Computing, Vienna, Austria.

[183] E. G. Carmines and R. A. Zeller, *Reliability and validity assessment, no. 17*, ser. Quantitative Applications in the Social Sciences. Thousand Oaks, CA, USA: Sage Publications, Inc., 1979.

[184] K. Eulenberg, “Untersuchung zum Einfluss von Kontext auf sensorische Profile,” Master’s thesis, Ilmenau University of Technology, Ilmenau, Germany, 2010, available only in German.

[185] ITU-T Study Group 12 - Performance, QoS and QoE, http://www.itu.int/net/ITU-T/info/sg12.aspx, visited: 21.05.2011.

[186] ITU-T Study Group 12, “Question 1312 - QoE, QoS and performance requirements and assessment methods for multimedia including IPTV,” Telecommunication Standardization Sector of the International Telecommunication Union, http://www.itu.int/ITU-T/studygroups/com12/sg12-q13.html, visited: 21.05.2011.

[187] M. McTigue, H. Koehler, and M. Silbernagel, “Comparison of four sensory evaluation methods assessing cooked dry bean flavour,” *Journal of Food Science*, vol. 54, no. 5, pp. 1278–1283, 1989.

[188] H. R. Hartson, T. S. Andre, and R. C. Williges, “Criteria for evaluating usability evaluation methods,” *International Journal of Human-Computer Interaction*, vol. 15, no. 1, pp. 145–181, 2003.

[189] J. T. Yokum and J. S. Armstrong, “Beyond accuracy: comparison of criteria used to select forcasting methods,” *International Journal of Forecasting*, vol. 11, no. 4, pp. 591–597, 1995.

[190] M. Cliff, K. Wall, B. Edwards, and M. King, "Development of a vocabulary for profiling apple juices," *Journal of Food Quality*, vol. 23, no. 1, pp. 77–86, 2000.

[191] M. Drake and C. V. Civille, "Flavor lexicons," *Comprehensive Reviews in Food Science and Food Safety*, vol. 2, no. 1, pp. 33–40, 2003.

[192] A. Noble, R. Arnold, and B. Masuda, "Progress towards a standardized system of wine aroma terminology," *American Journal of Enology and Viticulture*, vol. 35, no. 2, pp. 76–77, 1984.

[193] M. Meilgaard, C. Daigliesh, and J. Clapperton, "Beer flavour terminology," *Journal of the Institute of Brewing*, vol. 85, pp. 38–42, 1979.

[194] M. Brandt, E. Skinner, and J. Coleman, "Texture profile method," *Journal of Food Science*, vol. 28, pp. 404–409, 1963.

[195] M. Meilgaard, C. V. Civille, and B. T. Carr, *Sensory Evaluation Techniques.* Boca Raton, FL: CRC Press, 1991.

[196] E. Samoylenko, S. McAdams, and V. Nosulenko, "Systematic Analysis of Verbalizations Produced in Comparing Musical Timbres," *International Journal of Psychology*, vol. 31, no. 6, pp. 255–278, 1996, doi:10.1080/002075996401025.

[197] Y. Lincoln and E. Guba, *Naturalistic inquiry.* Newbury Park, CA, USA: SAGE Publications, Inc., 1985.

[198] P. Markopoulos and M. Bekker, "How to compare usability testing methods with children participants," in *Interaction Design and Children.* Shaker Publisher, 2002, pp. 153–159.

[199] E. D. Smilowitz, M. J. Darnell, and A. E. Benson, "Are we overlooking some usability testing methods? a comparison of lab, beta, and forum tests," in *Proc. of the Human Factors and Ergonomics Society 37th Annual Meeting*, 1993.

[200] B. M. Stecher, M. L.Rahn, A. Ruby, M. N. Alt, and A. Robyn, *Using Alternative Assessments in Vocational Education.* RAND, 1997, iSBN 0-8330-2489-2.

[201] H. M. Company, *The American Heritage Dictionary of the English Language*, 4th ed. Boston, MA, USA: Houghton Mifflin Company, 2006.

[202] J. Pagès, M. Cadoret, and S. Le, "The sorted napping: A new holistic approach in sensory evaluation," *Journal of Sensory Studies*, vol. 25, no. 5, pp. 637–658, 2010.

[203] E. Strauss, E. Sherman, and O. Spreen, *A compendium of neuropsychological tests*, 3rd ed. New York, USA: Oxford University Press, 2006.

[204] F. Wickelmaier and S. Choisel, "Selecting participants for listening tests of multichannel reproduced sound," in *Proc. of the Audio Engineering Society 118th Convention*, Barcelona, Spain, 2005, p. convention paper 6483.

[205] J. B. E. M. Steenkamp and H. C. M. Van-Trijp, "Free-choice profiling in cognitive food acceptance research," in *Food acceptability*, D. M. H. Thomson, Ed. Elsevier Applied Science, London, 1988, pp. 363–378.

[206] T. Dahl and T. Næs, "Outlier and group detection in sensory panels using hierarchical cluster analysis with the procrustes distance," *Food Quality and Preference*, vol. 15, no. 3, pp. 195–208, 2004.

[207] S. Werner, R. Sass, and A. Siegel, "Comparison of Recording Methods for Measurements of Individualized HRTFs," in *Proc. of the 26th VDT International Convention*, Leipzig, Germany, 2010.

[208] F. Klein, "Individualisierte Entzerrung von HRTFs," Master's thesis, Ilmenau University of Technology, Ilmenau, Germany, 2010.

9. Own References

[209] A. Gotchev, A. Smolic, S. Jumisko-Pyykkö, D. Strohmeier, G. Akar, P. Merkle, and N. Daskalov, "Mobile 3d television: Development of core technological elements and user-centered evaluation methods toward an optimized system," in *Proc. of 'Multimedia on Mobile Devices', a part of the Electronic Imaging Symposium*, San Jose, California, USA, January 2009.

[210] A. Gotchev, G. Akar, T. Capin, D. Strohmeier, and A. Boev, "3D Media for Mobile Devices," *Proceedings of the IEEE*, vol. 99, no. 4, pp. 708–741, 2011, Invited Paper.

[211] S. Jumisko-Pyykkö and D. Strohmeier, "Report on research methodologies for the experiments," Mobile3DTV, Project No. 216503, Tech. Rep., 2008.

[212] D. Strohmeier, "Wahrnehmungsuntersuchung von 2D vs. 3D Displays in A/V-Applikationen mittels einer kombinierten Analysemethodik," Master's thesis, Ilmenau University of Technology, Ilmenau, Germany, 2007, only available in German language.

[213] A. Gotchev, S. Jumisko-Pyykkö, A. Boev, and D. Strohmeier, "Mobile 3DTV system: Quality and user perspective," in *Proc. of 4th International Mobile Multimedia Communications Conference (Mobimedia)*, 2008.

[214] D. Strohmeier, S. Jumisko-Pyykkö, and K. Kunze, "Open Profiling of Quality: A Mixed Method Approach to Understanding Multimodal Quality Perception," *Advances in Multimedia*, vol. 2010, p. 28, 2010, doi:10.1155/2010/658980.

[215] S. Jumisko-Pyykkö and D. Strohmeier, "Report on research methodologies for the experiments," Projekt MOBILE3DTV, Tech. Rep. Project MOBILE3DTV, 2008.

[216] S. Jumisko-Pyykkö, M. Weitzel, and D. Strohmeier, "Designing for user experience: what to expect from mobile 3d tv and video?" in *UXTV '08: Proceeding of the 1st international conference on Designing interactive user experiences for TV and video.* New York, NY, USA: ACM, 2008, pp. 183–192.

[217] D. Strohmeier, S. Jumisko-Pyykkö, K. Kunze, and M. Bici, "The Extended-OPQ method for User-centered Quality of Experience evaluation: A study for mobile 3D video broadcasting over DVB-H," *EURASIP Journal on Image and Video Processing, special issue 'Quality of Multimedia Experience'*, vol. 2011, p. 24, 2011, doi:10.1155/2011/538294.

[218] D. Strohmeier, S. Jumisko-Pyykkö, and K. Eulenberg, "Open Profiling of Quality: Probing the Method in the Context of Use," in *Proc. of the Third International Workshop on Quality of Multimedia Experience (QoMEX2011)*, Mechelen, Belgium, 2011.

[219] A. Gotchev, D. Strohmeier, K. Mueller, G. Akar, and V. Petrov, "Source and Channel Coding Recipes for Mobile 3D Television," in *Proc. of the 17th International Conference on Digital Signal Processing (DSP2011), Invited Paper to special session 'Multiview and 3D Video Coding'*, Corfu, Greece, 2011.

[220] D. Strohmeier, S. Jumisko-Pyykkö, and U. Reiter, "Profiling experienced quality factors of audiovisual 3d perception," in *Proc. of the Second International Workshop on Quality of Multimedia Experience (QoMEX)*, Trondheim, Norway, June 2010, pp. 70–75, doi:10.1109/QOMEX.2010.5518028. Awarded with the T-Labs Best Paper Award.

[221] D. Strohmeier and S. Jumisko-Pyykkö, "How does my 3d video sound like? - Impact of loudspeaker setups on audiovisual quality on mid-sized autostereoscopic display," in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Istanbul, Turkey, 2008.

[222] S. Jumisko-Pyykkö, T. Utriainen, D. Strohmeier, A. Boev, and K. Kunze, "Simulator sickness - Five experiments using autostereoscopic mid-sized or small mobile screens," in *Proc. of the 4th 3DTV Conference (3DTV-CON)*, Tampere, Finland, 2010.

[223] D. Strohmeier and G. Tech, "Sharp, bright, three-dimensional: open profiling of quality for mobile 3DTV coding methods," in *Proc. of 'Multimedia on Mobile Devices', a part of the Electronic Imaging Symposium*, R. Creutzburg and D. Akopian, Eds., vol. 7542, no. 1. San Jose, CA, USA: SPIE, 2010, p. 75420T, doi:10.1117/12.848000.

[224] ——, "On comparing different codec profiles of coding methods for mobile 3D television and video," in *Proc. of the Second International Conference on 3D Systems and Applications*, Tokyo, Japan, May 2010.

[225] D. Strohmeier and S. Jumisko-Pyykkö, *Proposal on open profiling of quality as a mixed method evaluation approach for audiovisual quality assessment (COM 12 - C 181 - E)*, International Telecommunication Union, Geneve, Switzerland, 2011.

[226] S. Jumisko-Pyykkö, D. Strohmeier, T. Utriainen, and K. Kunze, "Descriptive quality of experience for mobile 3d video," in *Proc. of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, ser. NordiCHI'10. New York, NY, USA: ACM, 2010, pp. 266–275, doi:10.1145/1868914.1868947.

[227] K. Kunze, D. Strohmeier, and S. Jumisko-Pyykkö, "Comparison of two Mixed Methods Approaches for Multimodal Quality Evaluations: Open Profiling of Quality and Conventional Profiling," in *Proc. of the Third International Workshop on Quality of Multimedia Experience (QoMEX2011)*, Mechelen, Belgium, 2011.

[228] S. Schneider, F. Raschke, G. Gatzsche, and D. Strohmeier, "Free Choice Profiling and Natural Grouping as Methods for the Assessment of Emotions in Musical Audio Signals," in *Proc. of the Audio Engineering Society 126th Convention*, May 2009, Paper no. 7786. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=14982

List of Figures

List of Tables

List of Abbreviations

ACR	Absolute Category Rating
ANOVA	Analysis Of Variances
BIFS	Binary Format for Scenes
CP	Conventional Profiling
DCR	Degradation Category Rating
DSCQS	Double-Stimulus Continuous Quality-Scale
DSIS	Double-Stimulus Impairment Scale
EBU	European Broadcasting Union
EPM	External Preference Mapping
est.	estimated
est.	estimated
FCP	Free-Choice Profiling
GPA	Generalized Procrustes Analysis
HMFA	Hierarchical Multiple Factor Analysis
IBQ	Interpretation-based Quality approach
IPM	Internal Preference Mapping
ITU	International Telecommunication Union
IVP	Individual Vocabulary Profiling Method
MDS	Multidimensional Scaling
MFA	Multiple Factor Analysis
MFA	Multiple Factor Analysis
MOBILE3DTV	Mobile 3DTV Content Delivery Optimization over DVB-H System
MOS	Mean Opinion Score
MSS	Mean Satisfaction Score
OPQ	Open Profiling of Quality
PC	Pair Comparison method
PCA	Principal Component Analysis
PLS	Partial Least Square Regression
QDA	Quantitative Descriptive Analysis
QoE	Quality of Experience
QoP	Quality of Perception
QUAL	Qualitative research
QUAN	Quantitative research

RGM	Repertory Grid method
SAMVIQ	Subjective Assessment Methodology for Video Quality
SDSCE	Simultaneous Double Stimulus for Continuous Evaluation
SSCQE	Single Stimulus Continuous Quality Evaluation
SSMR	Single Stimulus with Multiple Repetition
UC-QoE	User-Centered Quality of Experience

A. Contribution of co-authors to the publications

The contribution of co-authors to the main publications (see section 1.4) is as follows:

Peer-reviewed journal publications

P1 The original idea for the paper was developed by the first two authors. The experiment 3 is written by Mr. Strohmeier. Abstract, introduction, discussion and conclusions were written by S. Jumisko-Pyykkö. In all other sections the work was shared between the authors.

P2 The paper was mainly written by the candidate. O. Bici wrote the sections about preparations of test stimuli. All authors commented on the paper before submission.

P3 The candidate proposed the structure of the paper and wrote the sections about user requirements and the subjective tests on mobile 3D television and video. He also commented on the final paper before submission.

Conference publications

P4 The idea for the paper was developed by the candidate who also wrote the results and discussion section. Research method section was written jointly by K. Eulenberg and the candidate. S. Jumisko-Pyykkö wrote the introduction. S. Jumisko-Pyykkö and the candidate had both significant impact on improving and finalizing the paper.

P5 The idea for the comparison model was developed by all authors. The candidate implemented the German version for the Conventional Profiling approach. The paper was written jointly by all three authors. The candidate and K. Kunze wrote the sections about comparison model, research methods and test results. S. Jumisko-Pyykkö wrote abstract, introduction, and conclusions. All authors had significant impact in finalizing the paper.

P6 The work on this paper was mainly shared between the candidate and S. Jumisko-Pyykkö. U. Reiter commented on the final paper and contributed to the research method section.

P7 The paper was written mainly by the candidate. G. Tech wrote the section about the selected encoding methods.

P8 The original idea for the paper was proposed by S. Jumisko-Pyykkö who also wrote abstract, introduction, discussion and conclusions. The candidate wrote the Related work section jointly with K. Kunze. The research method was co-authored by S. Jumisko-Pyykkö, Mr. Utriainen and the candidate. The results per experiments have contribution of all authors. The final model (DQoE - mobile 3D video) was jointly developed by S. Jumisko-Pyykkö and the candidate. All authors had a significant contribution to finalize the paper.

B. Thesen

1. Methods of sensory evaluation are descriptive research tools which can be used to extend quantitative evaluations of hedonic excellence with elicitation of individual quality factors.

2. Open Profiling of Quality is a validated mixed methods research approach applicable with naïve participants which combines standardized psychoperceptual evaluation and an adaptation of Free-Choice Profiling. The two data sets can be linked in methods of External Preference Mapping.

3. Mixed methods research approaches differ in strengths and weaknesses which need to be assessed and understood to be able to select the proper approach to specific research questions. The different approaches developed from methods of semi-structured interviews, sensory profiling, or perceptive mapping can complement each other within different phases of technical system optimization.

4. Holistic comparison of mixed methods research methods needs to go beyond constrasting outcomes and must include aspects of costs, duration or personnel effort for systematic comparisons.

5. The perception of mobile 3D (autostereoscopic) television shows a hierarchical dependency of the two most impacting quality factors video quality and depth perception. Added value provided by depth perception is only perceived when the video is free of visible artifacts.

6. Modern quality evaluation studies need validated tools to describe the test sample beyond visual acuity, 3D and color vision, and hearing threshold. Especially in audiovisual quality evaluations, results can be biased due to different processing styles of test participants towards audio and video. In 3DTV research, bias can occur due to different perception of large disparities which can either result in very good perception of depth or, negatively, in perception of crosstalk.