universität
wien

MASTERARBEIT

Titel der Masterarbeit

“Observer Confidence in Subjective Quality Evaluation”

verfasst von

Werner Robitza, BSc

angestrebter akademischer Grad

Diplom-Ingenieur (Dipl.-Ing.)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 066 935
Studienrichtung lt. Studienblatt:	Masterstudium Medieninformatik
Betreut von:	Univ.-Prof. Dipl.-Ing. Dr. Helmut Hlavacs

Contents

List of Figures

List of Tables

1 Introduction

The twenty-first century has brought a rapid change in technologies. Whether it is the advent of high-speed Internet connections, the penetration of smart phones and tablet devices, or the now almost completed switch from analog to digital television broadcast: technology is evolving faster every year, and this evolution also changes the way we perceive our environment. The entertainment industry bases its income on products that try to please the human mind—often enough by providing a pleasing visual experience that sets off the current generation of products from the previous. Some examples of this include:

- **The switch from SDTV to HDTV:** Standard-definition television was something most people felt accustomed to. Terrestrial television standards such as PAL and NTSC had been introduced in the 1960s, popularized in the decades thereafter [37] and are still in use. Yet, with high definition TV, the market demanded more: soon, broadcasters would acquire and deliver content in HDTV, and people would buy HDTV equipment for their home.

  The primary goal of HDTV is to provide better visual fidelity by increasing the spatial (and temporal) resolution of the signal.
- **Digital Cinema:** Filming for cinema on analog equipment is still common today, but has often been superseded by digital acquisition, production and playback [52]. In cinemas that support digital playback, even analog movies are shown digitally. One of the reasons for this change is found in logistics, but the digital medium provides visual benefits: higher temporal and spatial resolution as well as no flickering or dust.
- **3D Vision:** Providing viewers with a third dimension is not new—in fact, 3D cinema technology was already established in the first half of the 20th century [29]. However, with today's digital technology, 3D vision promises to add value to existing services such as TV or cin-

ema. By adding the depth dimension, the visual experience is enhanced by a new feeling of immersion.

A great challenge with all such technologies is to *quantify* how users experience them. Whilst qualitative descriptions may provide sufficient answers to higher level questions about the service quality, these measurements can only be taken from a small number of users, or—with a larger number of users—are hard to evaluate. For example, a television provider could conduct a trial of a TV service with a predetermined number of users, collecting their opinion through diaries [17] or other repeated questionnaires. Without specific quantitative measurements of experienced quality, it is virtually impossible to answer the questions of *how much* a service can be improved, whether its improvement paid off (even in a literally financial sense), or in which conditions the service becomes unacceptable. The quality, as rated by the users, can become a direct Key Performance Indicator (KPI) for businesses.

Methods for measuring the service quality as well as the experienced quality from a user's perspective have existed even before the digital age. Documents released by the International Telecommunication Union (ITU) describe procedures to assess the subjective quality of television pictures [6] or multimedia applications in general [5]. They describe technical considerations (partly still based on analog viewing equipment) as well as rating methodologies and procedures for data analysis.

When reporting results from subjective quality tests, scores given to one stimulus are typically averaged across all observers. This is called the Mean Opinion Score (MOS). Observers may be rejected from further data analysis when their scores differ from the sample mean by a certain amount, i.e., when they are statistical outliers. The MOS then gives a simple estimate of the perceived quality of a stimulus. Averaging the results from multiple viewers ensures that there is one single value that can be reported and interpreted, which allows for easier understanding or processing in mathematical models, but also for human beings.

The meaning of the MOS value can be explained according to the scale it was measured with.[1] The score's variability, commonly expressed through the statistical confidence interval (CI), is reported along with it. The CI gives researchers an estimate of how close the measured MOS may be to the "true population" MOS, and it can be used to facilitate comparisons between different stimuli in order to decide whether they show significant statistical differences. Naturally, the CI becomes lower as we increase the number of participants in a subjective experiment, but in practice, there are often (financial or logistic) limitations as to how many observers can be recruited and tested.

In this thesis, I will focus on a specific aspect of subjective experiments: the confidence of the human observers themselves. In previous experiments I had carried out at the University of Vienna (described in [11, 14, 13, 30, 36, 41, 42, 44]), I noticed that some observers struggled to decide for a rating when forced to give one. In contrast to the confidence interval—a measure of "reliability" across all observers—the confidence of the subjects can however not be directly deduced from the acquired data. Whether users feel confident or not while they are taking a test may have an impact on their opinion scores. In fact, averaging across all observers hides their individuality and masks any personal factors that might influence their votes.

This work is based on the preliminary research done by Ulrich Engelke et al., who measured observer confidence in image quality experiments [16]. We follow the principal ideas and extend them to the domain of video quality, enriching the existing data by adding several thousand quality and confidence scores, comparing existing findings with ours.

The primary research questions treated in this work include the following:

- How can and should the confidence of human observers in subjective tests be measured?
- Which factors have an impact on the confidence?
- How can confidence ratings be used for data evaluation?

[1] For example, the scores might be collected on a numeric interval scale from 0 to 10, or an ordinal scale from "Bad" to "Excellent".

In Chapter 2, we will describe what the term “quality” means, look at the history and current developments in multimedia quality evaluation and describe the most important procedures for conducting subjective tests. Based on these methodologies, Chapter 3 will explain more about the motivation behind assessing observer confidence. What is confidence and how does it influence the behavior of users in tests? Why is measuring the confidence necessary? In this chapter we will also define specific hypotheses.

To answer the research questions stated above, seven experiment sessions were carried out at the University of Vienna and the Institut de Recherche en Communications et Cybernétique de Nantes in France (IRCCyN). These experiments, including the video material and test conditions, are described in detail in Chapter 4. The results of these experiments will be analyzed in Chapter 5 including recommendations on how to perform subjective experiments, considering the evaluated data. The thesis will be concluded in Chapter 6.

2 Multimedia Quality Evaluation

In this chapter, we will address the two main dimensions according to which one can quantify the quality of a multimedial presentation or service, Quality of Service (QoS) and Quality of Experience (QoE). The link between these concepts is often evaluated through means of subjective experiments, which are often carried out according to standardized procedures. We describe the meaning of QoS and QoE, list the most common evaluation methods and give an outlook on mixed method approaches that have recently been proposed as an alternative to many standardized procedures.

2.1 What is "Quality"?

The definition and understanding of the term "quality" is essential to the topics discussed within this thesis. On a more general level, an agreed-upon definition of quality is necessary in the domain of multimedia evaluation.

In the beginnings of analog broadcasting, it had already become clear that merely delivering the service was not enough to satisfy the audience. Due to technical constraints, perfect fidelity could never be guaranteed in the signal chain that went from the broadcaster to the consumer's living room. Factors such as the terrestrial transmission or the quality of the TV set itself would influence the way humans perceived the services. Even with the change from analog to digital, these issues ensue, albeit in different characteristics. With consumers often paying for today's entertainment services, they expect a certain minimum level of positive experience, of which quality is one of the contributing factors. This raises the question of how "quality" can be defined.

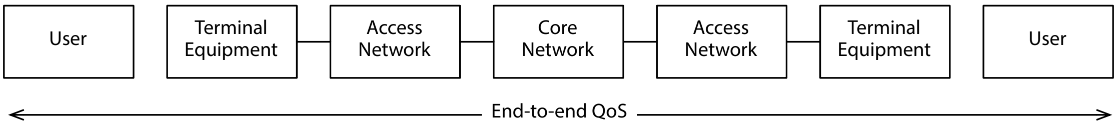


Figure 1: End-to-end Quality of Service as defined in [3].

2.1.1 Quality of Service

In the Recommendation E.800 [3], the ITU defines "quality" as follows:

> "The totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs."

To generalize this, in [3], the "Quality of Service" is defined as:

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

The needs of the user may be of any kind, however it is the responsibility of the telecommunications provider to map these descriptions to actual preference values. In other words: The characteristics mentioned here have to be measurable, which means that there must be a kind of objective metric behind them. "Objective" in this case refers to data that is not based on human interpretation (although the term "quality" itself is still not formally defined without taking into account the subjective nature of the human being).

To efficiently measure the quality of a service, one could use metrics deduced from network parameters, such as a signal to noise ratio (expressed as Peak Signal to Noise Ratio, PSNR), an average transmission bit rate (in Bits per second), the packet loss ratio in IP-based networks (in percent), etc. These measurements are often permanently and continuously available to the service providers and result in objective, quantifiable data. We summarize them as Quality of Service (QoS) measurements.

These metrics—in a first instance—only apply to the service itself, e.g., a TV broadcast, a Voice-over-IP telephone network, etc. To be precise, the metrics often only apply to individual components of a service (see Figure 1). They can be efficiently monitored and provide an estimate of the actual service quality when interpreted correctly. For example, a mobile television service provider might monitor the average throughput of their streams. The monitoring system might alert the provider when the throughput, i.e., the average data sent per time, drops below a certain level that would leave the customer unsatisfied. How would the provider find out how the satisfaction of their typical customer and the throughput are related? The user satisfaction is part of their experience of the service—thus, the provider also needs to be able to measure the experienced quality.

2.1.2 Quality of Experience

Beyond the objective QoS measurements there is the concept of Quality of Experience (QoE). Bringing together the definition of "quality" from the previous Section with the notion of (user) "experience" gives us a rough idea of what is meant by QoE: the totality of factors influencing the experienced quality from a user's—and not a system's—perspective.

Today, there is still no standardized definition of QoE. The term is missing both a globally accepted formal definition as well as a framework in which it can be situated. The ITU [4] defines QoE as:

> " The overall acceptability of an application or service, as perceived subjectively by the end user. "

However, the notion of acceptability only seems to address one dimension of QoE, disregarding other factors. Recent discussions within the domain of multimedia quality evaluation have brought up new definitions that might lead to a more commonly accepted standard, which define the term QoE as follows [7]:

> “ Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state. ”

As mentioned before, there is no inherent ground truth as to what the values acquired in objective QoS measurements mean for the customer experience. For example, a network operator will not be capable of translating the average transmission bit rate to a quantifiable index of customer experience, unless there is specific data that allows such a mapping, factoring in all other components in the signal chain—from the source video to the users' viewing context. To gather this data, ratings have to be acquired from real human observers. This is usually done in psychophysical experiments (also called “subjective” experiments in the following).

2.2 Subjective Experiments

Subjective experiments in multimedia quality evaluation are carried out to provide real-life feedback from a certain number of human observers to a selection of stimuli that are usually prepared in advance of the experiment. The primary goal is to acquire data about the users' (subjective—hence the name) experience with those stimuli. The possible methodologies are countless and depend on the goals of the assessment, but usually follow standardized forms (see also Section 2.3). Psychophysical experiments traditionally aim at gathering data in a controlled environment, so as to produce repeatable results and eliminate the number of confounding variables between and during individual test sessions.

2.2.1 Experiment Participants

Participants in subjective experiments are also referred to as "observers", "assessors", or "listeners", depending on the context.[2]

In general, the audience of such tests should be representative of the audience the test material will see in realistic conditions. Trial studies or surveys may help to narrow down the potential target audience to a specific population. For example, in a study targeting short video clips or television consumed on mobile devices, observers might be selected to be between 23 and 35 years of age [10].

Test persons are usually required to be naive in the sense of being non-experts in the domain of the test setting, so as not to introduce bias. In some circumstances, e.g., when time or monetary resources are limited, expert viewers could replace a set of naive observers [30, 5]. With naive observers, however, it is advisable to select participants who have not conducted other experiments in the field of quality assessment recently.

The number of observers needed for an experiment again depends on the overall goals and available resources. From a statistical standpoint, choosing to include more observers will result in sample data that is more representative of the reality. On the other hand, the time needed to conduct a study often linearly increases with the number of observers. Personnel trained with the conduction of experiments is rarely available for longer periods of time—an experiment therefore does not always benefit from including more observers than necessary to prove the research question that was set earlier. ITU recommendation BT.500-13 [6] for example requires 15 observers but allows for fewer in case of exploratory studies. In [5], an absolute minimum of four observers is specified, but it is also mentioned that there is "rarely any point in going beyond 40." According to [53], the inclusion of 15 observers already results in considerably low standard deviation. Finally, studies have shown that 24 subjects may lead to inter-experiment dataset correlations of 0.96 and higher [33].

[2] The general terms "observers" or "assessors" can be applied to multi-modal experiments that involve both audio and video stimuli or multimedia content.

2.2.2 Test Material Selection

Similar to human observers, the test material shown to them should constitute a representative sample of the reality. The experiment designers will select individual video or audio clips from a pool of sources, which are usually freely available movies, television shows, or dedicated resources published by the research community.[3] The length of individual samples will change depending on the purpose of the experiment. For example, a study on long-term quality effects in television might only show a single stimulus with a duration of 30 minutes. Likewise, an experiment focusing on Internet-based video clips would only show excerpts of a few minutes. Finally, if the goal of the study is to focus on technical aspects (i.e., with the intent of minimizing the influence of the narrative content of the source), short clips of ten seconds or even still images could be used.

On the one hand, the source material selection can aim at providing a broad range of content and characteristics. This allows to predict the influence of the conditions under study on various content types. On the other hand, content may be chosen to be relevant to the system under test. For many types of studies, the content can be grouped into "genres" such as animated movies, documentations, action movies or sports. However, it is often sought to describe the difference between sources in a quantitative manner, referring to specific, objectively measurable characteristics. When we look at video, two dimensions of information are coded: the spatial information (SI), which denotes what can be seen at a specific point in time, and the temporal information (TI), which describes the motion of the spatial information over the course of individual pictures. In general it can be said that the more temporal and spatial information has to be coded, the higher the necessary transmission bandwidth will be. Vice-versa, under a constrained bandwidth scenario, higher SI and TI will lead to decreased quality. This underlines the importance of SI and TI calculation for video quality measurements.

[3] Such resources along with test results can be found on `http://dbq.multimediatech.cz/`, published through the Qualinet project, or at `http://stefan.winkler.net/resources.html`.

In [5], the calculation of SI and TI is explained in detail. SI is calculated based on the Sobel filter applied to the luminance plane of every frame (F_n). The SI is the maximum standard deviation σ over all frames of the source:

$$SI = \max_{time}\{\sigma_{space}\,[Sobel(F_n)]\} \tag{2.1}$$

The TI is based on the location difference $M(i,j)$ of a pixel at position i,j which is defined as the difference of $F_n(i,j) - F_{n-1}(i,j)$. Again, the maximum of all standard deviations of M, over all frames, is taken as the TI value for the source:

$$TI = \max_{time}\{\sigma_{space}\,[M(i,j)]\} \tag{2.2}$$

For visual inspection, SI and TI values of each sequence can be plotted against each other. The experimenter may then choose a set of sources that match criteria such as high temporal motion with low spatial detail, or a broad range of SI/TI variance, et cetera. An example of SI and TI values for a given set of SRC can be seen in Figure 2.[4]

2.2.3 Generation of Test Conditions

Confounding variables in experiments and statistical models are variables that correlate with both dependent and independent variables and may therefore lead to wrong interpretations of the obtained results [18]. In order to minimize the confounding variables stemming from different source contents, typically only few clips are selected and then manipulated according to the specific criteria of the experiment. These manipulations generally alter the visual or auditory quality of the source. They can be seen as treatments of the original content. Observers are then shown the processed clips and—depending on the procedure—the originals.

[4] The SRC were used in [15] and are available from `http://www.irccyn.ec-nantes.fr/spip.php?article491`.

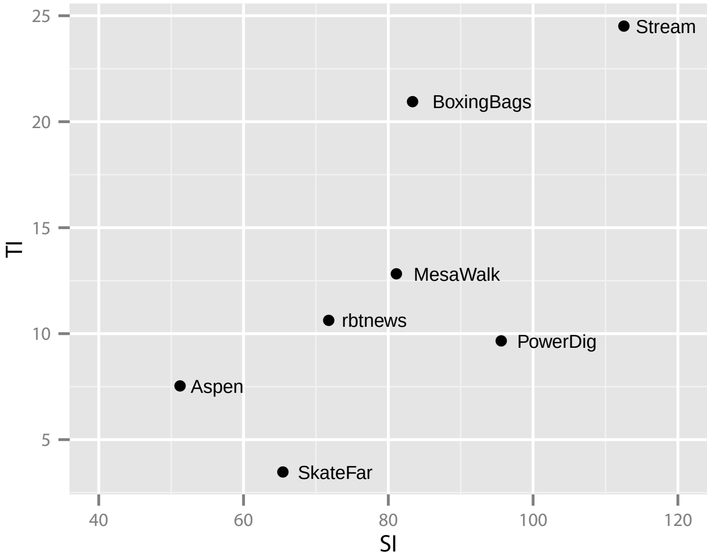


Figure 2: SI and TI values for a set of SRC.

In the domain of QoE, the following terms are commonly used:

— **SRC:** *source* stimulus. Refers to the unaltered, original material, with a specific duration.

— **HRC:** *hypothetical reference circuit.* Defined as "a video system under test such as a codec or digital video transmission system," [34] it refers to all possible parameters of the system, from an encoding stage to the physical transmission.

— **PVS:** *processed video sequence.* The result of applying the HRC treatment on a given SRC.

When designing an experiment, the experimenter will usually set an upper bound of the total experiment duration and then choose the number of PVS that can be shown in this period. For example, if six SRC clips of ten seconds length are chosen, a session of 25 minutes $(25 \cdot 60 = 1.500)$ could contain 150 PVS and therefore allow for a maximum of 25 HRCs to be tested. In practice, however, the researcher needs to accommodate for the time participants take to give their rating to a stimulus, as well as an additional pause of several seconds between stimuli presentations. This reduces the possible number of PVS. In our previous example, this could mean that 15 seconds are planned per PVS presentation, leading to only 100 possible presentations.

2.2.4 Context and Environment

The subjective experience of audiovisual content should never be interpreted without the context it is encountered in. Context in this case refers to the totality of circumstances of the experienced event, from the observer's social context, their preconceived ideas, but also the locality as well as the devices and means of presentation. Typically, one would differentiate "real-life" context from laboratory context, the main distinctive features of those lying in the physical location and—as a consequence—the number of confounding variables. As mentioned in the beginning of this section, reliability is an important factor for the credibility and success of experiments. In this case, we focus on inter-test reliability: it refers to the possibility for other researchers to successfully reproduce an experiment without introducing errors [54]. This, on a large scale, depends on the context of the experiment being well-defined and documented.

Historically, a significant part of research documentation focused on the technical parameters of the laboratory environment, namely the room luminance, the color and luminance ratio of the background behind the presentation device as well as the distance of the observer to the device, and several more. For the viewing device itself, there may be restrictions on the peak luminance of the display and its contrast ratio. However, with the increased diversity in display technology (e.g., the change from CRT to TFT and LCD displays), or the adoption of new consumption contexts altogether (e.g., mobile television), the historical definitions could be seen as too restrictive for research focusing on a context leaned towards the "real-life" or observers who consider themselves "early adopters". The exploration of new contexts is an ongoing effort in the current field of QoE evaluation [26]. It is therefore not surprising that on a global scale the definition of a typical laboratory environment has not changed, while the plurality of possible viewing devices and environments has given researchers more and more leeway to define their own context.

Concerns may be raised about the reliability of experiments not conducted in strictly controlled environments. Even more so, uncontrolled experiments are deliberately distanced from laboratory studies. So-called "living labs" take this idea as far as the actual home of the user, where researchers might not even be physically present during all experimental phases [46]. In such a case, a trade-off is made between the intrinsic reliability of the experimental method versus the benefit of gathering data from a "real" rather than a constructed context. Even in the domain of subjective QoE testing, recent studies have found little to no influence of the experimental context. Instead, the critical control variable for subjective experiments remains the number of participants [33]. It could be argued that in the foreseeable future, the dependency of subjective experiments on methods that came into practice several decades ago will steadily diminish, at least with respect to the context settings.

2.2.5 Instructions and Training

Before giving their ratings, each observer should be properly instructed by means of a written or oral introduction to the purpose of the study and the course of action. A written instruction is preferred in order to prevent ambiguities or bias from different experimenters working with the participants.

An important part of each test session is formed by the so-called "training session", in which observers are asked to perform ratings for a set of selected stimuli. Their ratings are typically not recorded or discarded later. In a training session, observers should be made familiar with the types of degradations they are about to experience. A properly selected set of training material should result in the rating scale being used in its full extent. Put differently, if users cannot see the characteristics of the stimuli before their ratings are taken, they might shift their interpretation of the scale throughout the experiment session. It is therefore advisable to show stimuli that exhibit the widest range of quality—for example a reference sequence and a video treated with the "worst" condition, as judged by an expert.

While not very common, a training session can also be guided in the form of the experimenter talking with the observers during or after their rating. Questions pertaining to the usage of the scale or the experiment procedure in general could be answered. This should help clarify issues before participants start with the actual rating session (which usually cannot be interrupted).

2.3 Subjective Quality Evaluation Procedures

At the point where the goal of a study is defined, a suitable procedure for presenting the stimuli and polling for ratings needs to be chosen. Even with a set of generated PVS, the procedure itself depends on a number of aspects, which include (but are not limited to) the following questions:

- How long should the experiment session last?
- Where should the experiment take place?
- How many times should each PVS be viewed?
- Are ratings taken once per stimulus or continuously?
- Do ratings rate one stimulus absolutely, or do they only compare two (or more) stimuli?
- Are ratings taken qualitatively or quantitatively?
- If quantitatively, which scale are ratings taken on?

The answers to these questions define the procedure, i.e. the experimental methodology. As mentioned previously, a requirement for published research is that it is reproducible in a reliable fashion. The methodology should therefore be described as detailed as possible. Naturally, the need for such a definition resulted in the written formalization of methodologies by bodies such as the ITU or EBU, which researchers can choose to follow and extend or modify. As this thesis focuses on video quality in particular, the remainder of this section will introduce the most common methodologies used today, grouped by whether they assess the quality of one stimulus or multiple stimuli at the same time.

It should be noted that some methods are intended for use with general multimedia content (e.g. those from [5], whereas others were developed in the domain of TV broadcasting (e.g. [6]). The main differences between these fields lie in the variability of multimedia content with regard to spatial and temporal resolution, the different viewing contexts, and the codecs used. Care should be taken to choose a method that suits the intended scenario of the study.

2.3.1 Single Stimulus Methodologies

Single stimulus methodologies present one PVS at a time and ask for a rating with respect to this stimulus only. They are suited best for evaluating content without focusing on the degradation itself, but the overall impression of quality as viewed by an end user. They are therefore also suitable for scenarios where an original source might not be available, or where the quality of the original signal is already low.

2.3.1.1 SSCQE — Single Stimulus Continuous Quality Evaluation

SSCQE is defined in [6] and is a method for quality evaluation of long duration content. It was intended for television scenarios. Observers see the material in its entirety without a reference (i.e., the SRC) and are continuously polled for ratings. The ratings are taken on a slider, a device that is mounted on the desk with a linear fader of 10cm traveling range (see Figure 3). Samples should be taken at a minimum rate of 2 Hz. Depending on the context of the study, other rating devices may be suitable. For example, gloves with sensors have been proposed and are under evaluation for mobile rating contexts [13].

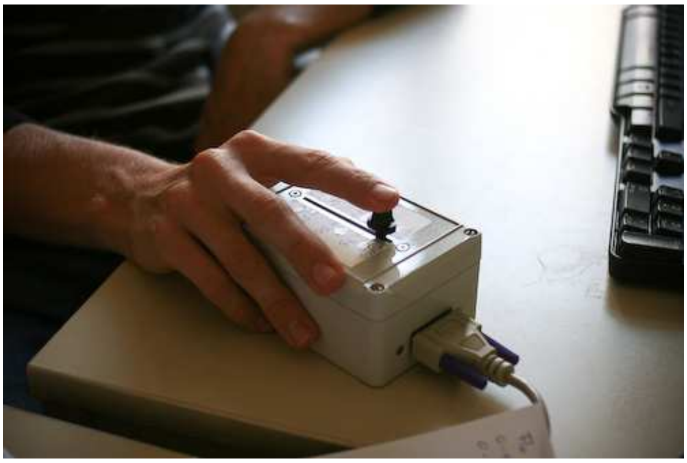


Figure 3: Linear slider to be used in SSCQE experiments.

Test material can be prepared in such a fashion that multiple HRCs are included in the presentation. For example, the bitrate parameter may change every minute, although the presented content is still from the same SRC. The session is therefore logically split up into parts that are unnoticeable by the observer.

The analysis of SSCQE ratings is amongst the more complicated procedures, since the rating response of an observer cannot be directly correlated to a specific moment in the PVS. A rating delay that is unique to each observer has to be taken into account. Continuous scores have to be aggregated in certain periods. It is recommended that a memory effect is assumed, meaning that human observers will "remember" the quality of a previously shown point in time. This effect should be modeled as an exponential decay function. Also, anchoring the ratings and compensating for an overall systematic shift in quality perception is necessary, since observers may be too optimistic or pessimistic, or change their perception (or interpretation of the scale) throughout the presentation.

2.3.1.2 ACR — Absolute Category Rating

ACR, described in [5], focuses on short duration content of about 10 seconds length.[5] Observers see one PVS after another, and are given about 10 seconds time to rate the overall quality of the previously seen sequence on the following scale (numbers in parentheses are the ordinal values):

- Excellent (5)
- Good (4)
- Fair (3)
- Poor (2)
- Bad (1)

ACR also allows for other scales to be used, e.g. with nine or eleven points such as defined in Annex B of [5]. ACR also could be performed with quasi-continuous scales, where the above-mentioned five labels are included as reference marks. However, it has recently been found that more fine-grained scales do not necessarily increase discriminative power [22]. Quite the contrary, most observers would still align their ratings at the given guides. As a result, the difference between ratings taken on different scales was found to be insignificant.

Ratings obtained from the ACR method are usually averaged per PVS into the Mean Opinion Score (MOS) and presented along with the confidence interval, although the original five-point scale itself is only ordinal. The benefits and drawbacks of this process will be analyzed in Chapter 3.

Since the classic ACR method does not include the original SRC for each PVS, ratings may be skewed. The reason for this is that observers are asked to absolutely judge the quality while they have never seen the range of quality, or at least the highest possible quality for each SRC. Therefore, the ACR-HR (Hidden Reference) method includes every SRC as a PVS (without any HRC applied

[5] Equivalents of this method are the SS (Single Stimulus) methods from [6]

to it) in the presentation, without explicitly marking it as such.[6] This means that the observers do not know that they are seeing a reference condition.

In ACR-HR, a DMOS (Difference MOS) is calculated per PVS. It is defined as the score of the PVS minus the score given to the hidden reference, plus the number of points on the scale. For example, if the PVS was rated as “Poor”, and the HR as “Good”, then the DMOS equals $2 - 4 + 5 = 3$. Care must be taken with DMOS values greater than 5, i.e. where the PVS was rated better than the HR. A crushing function may be used to deemphasize the weight of these ratings.

2.3.2 Double or Multiple Stimulus Methodologies

In contrast to single-stimulus, double-stimulus methodologies aim at identifying differences rather than absolute quality. They may be suitable for testing the preservation of fidelity in transmission chains. Multiple-stimulus methodologies typically require more effort to implement and may not be as commonly used for smaller studies.

2.3.2.1 DSCQS — Double Stimulus Continuous Quality Scale

In DSCQS [6], PVS of about 10 seconds length are chosen. A session may last up to 30 minutes. The observer sees the reference and the PVS in alternation. Reference and PVS are randomly labeled as *A* and *B*. Observers may choose to switch between both presentations and view them up to three times or more until they settled on a quality score for both, which they indicate on a continuous scale where the intervals are labeled with the ACR categories (“Excellent” to “Bad”), resulting in six guide marks.

[6] The only requirement is that the SRC itself must be considered as “good” or “excellent” by an expert viewer already. This makes the ACR-HR method unsuitable for content that is already impaired.

It is important to interpret the scores as a difference between reference and condition only—the absolute values or textual labels from the guide marks shall not be used for analysis.

2.3.2.2 DCR — Degradation Category Rating / DSIS — Double Stimulus Impairment Scale

DCR (from [5]) and DSIS (from [6]) both refer to the same method. It is similar to the DSCQS (PVS length, session duration), however the order of presentation is fixed and known to the observer: the reference SRC is shown before the PVS. Ratings should express the degradation of the quality when comparing original and impaired version, on the following scale:

- imperceptible (5)
- perceptible, but not annoying (4)
- slightly annoying (3)
- annoying (2)
- very annoying (1)

The presentation of both clips may be repeated once, and the observer may already rate the quality during this iteration. For data analysis, the values of the impairment scale are averaged per PVS into a MOS. This is, inherently, a DMOS as introduced in the ACR method. DCR/DSIS data allows to develop a "failure characteristic" for transmission systems, which gives the probability that a certain content will be perceived with a certain level of quality.

DCR/DSIS was shown to be error-prone with regard to contextual effects [1]. Contextual effects occur when the rating of a stimulus (or a pair of stimuli) is influenced by the previously shown stimuli. In comparison, DSCQS does not show contextual effects.

2.3.2.3 PC — Pair Comparison

The PC method [5] differs from the other double-stimulus procedures in that it does not compare a reference against the impaired sequences generated from the reference, but only the different HRCs against each other. Given a set of HRCs $H_1, H_2, \ldots, H_n$, all possible $n \cdot (n-1)$ combinations are generated, and then shown in both possible orders. No order is shown twice, and by default, the sequences are shown after each other. In a variation of PC, both sequences may be shown simultaneously. Observers then only indicate their preference of one HRC over the other, not through a specific score.

As indicated by the possible number of combinations, the procedure does not allow for as many HRCs to be tested as others. Its advantage, however, is that it juxtaposes different conditions directly, whereas other double-stimulus methods only compare against a reference, and direct HRC comparisons are not possible.

2.3.2.4 SAMVIQ — Subjective Assessment Methodology for Video Quality

SAMVIQ [2, 28] is a more recent methodology stemming from the EBU. It is a multi-stimulus method, in the sense that the observer can choose the order of the individual PVS presentations. In contrast to other methods, it also lets the observer continue at their chosen pace, which reduces the risk of attention loss during an experiment session. In SAMVIQ, SRCs are chosen with a length of around 10 to 15 seconds.

In addition to the impaired PVS, SAMVIQ includes both explicit and hidden references for each SRC. The rationale behind this is to compensate for the bias that viewers exhibit when they rate explicitly marked references. About a third of users scores an explicit reference higher than the (visually identical) hidden reference. As a consequence, choosing not to include references (or the lack of SRCs in general), leads to higher variability in the resulting data.

The observer can view each sequence as many times and as long as they want, and rate them accordingly—the only exception being that the first time a PVS is viewed it has to be watched in its entirety. Only when all sequences of a SRC are rated, the observer may continue rating the next set of PVS. The rating scale is a continuous scale with five marked intervals using the ACR labels from “Excellent” to “Bad”, resulting in six guide points. For data evaluation, the scores are averaged into a MOS, as known from other methods.

2.3.2.5 SDSCE — Simultaneous Double Stimulus for Continuous Evaluation

A variation of SSCQE, the SDSCE method allows the observers to view the reference source alongside the impaired version, both at the same time. This method is aimed at testing the fidelity of a stream rather than the absolute quality. It also helps in avoiding systematic shifts over time, since observers may align their ratings continuously with the original.

2.3.3 Alternative Methodologies

In the recent years, the advent of new technologies required researchers extend or adapt the existing subjective test methodologies for new viewing contexts. To give an example, while the QoE of a 3DTV service *can* be measured by following a typical ACR procedure, the technical specifications of existing recommendations were written without accounting for 3D hardware. The current standards therefore have to be revised, or new methodologies will have to be agreed upon by the community and the standards bodies.

One issue however remains with all of the above-mentioned methodologies: they aggregate the individual scores per PVS into a MOS. The MOS as a simple arithmetic mean makes it easy to compare results, but at the same time masks the individual ratings which could very well also be taken into account on a per-user basis. This is especially important for PVS where the inter-subject agreement is low, which would indicate problematic content, and require more insight into the user's decision making process.

Recent propositions suggest taking a bottom-up perspective, where users could take specific perceptual attributes to describe the quality. Some of these methods stem from the sensory sciences, which were intended for evaluation of food products. One method is the so-called "Napping" procedure [39, 40], which was successfully ported to the audiovisual QoE domain [48], and is also available as a ready-to-use tool for quick experiments on tablet devices [43]. It allows users to sort PVS according to their perceived *similarity* in the quality domain rather than quality itself. Later, observers are asked to label PVS with keywords that describe certain quality features they identified, such as blurriness, flickering, jerky motion, or noise. Another method is the Repertory Grid Technique [49]. Here, triples of stimuli are sorted by the participant in such a way that two stimuli are similar, but different from the third. This composition is called a "construct". Both of these methods make heavy use of factorial analysis.

In general, in the recent years, there has been a strong focus on descriptive (qualitative) analysis of QoE instead of pure quantitative experiments, which focus less on the user than on the system under test. Mixed-method approaches where qualitative and quantitative data is gathered are becoming more common. A framework for user-centered QoE has been proposed in [25]. The principal issue that remains is the lack of standardization for reporting qualitative or mixed-method results, which would not only make inter-laboratory tests and comparisons feasible, but with the definition of a new measure to replace the MOS, render the design of objective quality metrics more efficient.

3 Observer Confidence

As seen in Chapter 2, subjective experiment methodologies follow specific protocols that describe the ways in which observers are asked for their ratings. In this chapter I will focus on the implications of these protocols on the assessors' decision making process, and the possible influence on the gathered results. This chapter will motivate and outline the experiments further described in Chapter 4.

3.1 Questionnaires for QoE Measurement

In a general fashion, the data acquisition method of subjective QoE experiments can be viewed as a questionnaire. The Oxford Dictionary of English defines a questionnaire as a "set of printed or written questions with a choice of answers, devised for the purposes of a survey or statistical study." [32] Before computer-based testing with graphical user interfaces, responses to QoE tests would in fact be recorded on a sheet of paper, each question referring to one or more videos the observer had just seen.

3.1.1 Measuring QoE Means Measuring Opinion

QoE experiments inherently ask for subjective opinion—perhaps an opinion towards a certain stimulus' visual or auditory excellence. They are generally not conceived as tests with a *given* truth that an observer with the right knowledge should find.[7] In fact, requiring non-experts to participate

[7] A notable exception to this are content recognition tasks in which users are asked to identify objects in a video stream, such as surveillance footage. Here, the ground truth is known at the time of asking. The focus of the experiment still remains on video quality though, not on the physical or mental ability of the observer to successfully detect an object.

in theses tests achieves quite the opposite: the true experience of the observer should manifest in their score given to a stimulus. While researchers will never have direct access to the attitude of the observer towards the stimulus, the questionnaire scores will become meaningful when looking at the results of multiple assessors—the questionnaire becomes a measure of behavior too [9]. It is only when aggregating results that drawing conclusions is possible. Herein lies a problem, namely when the internal validity of responses cannot be proven.

The ACR protocol (see Section 2.3.1.2) is exemplary of the way current subjective video quality evaluation methods elicit data from observers. After each stimulus, a response is recorded. This response—in form of a vote for a specific option on a given scale—is stored along with all other responses. In this regard, the response is forced, and might not represent the observer's true perception when it was given under the pressure of having to complete the experiment session. Yet, none of the common methodologies allow for ratings to be explicitly marked in any fashion that would later allow the experimenter to judge them differently from others. It would be desirable to detect responses that could be invalid, in the sense of having knowingly being made in order to proceed to the next item without expressing an opinion in the rating. This data could then be removed from analysis altogether, or—if the data is still deemed valuable—interpreted with regard to the nature of it being possibly invalid.

One way to prune (known) invalid data is to allow participants to skip questions altogether. In fact, a large number of skipped items may be a sign for the chosen rating scale to be inefficient [31]. This, however, has practical disadvantages. A dataset with missing items can be compensated for by including more observers. For social studies with dozens, hundreds, or even thousands of participants, a few skipped items seldom have a statistically significant effect. In QoE testing, where a number of 15 observers or fewer are common, this effect would be much more noticeable.

At this point, it becomes important to stress the difference between the concepts of reliability and validity. Reliability refers to a consistency during and even across tests, showing that with a certain methodology, reproducible results are obtained. More specifically, reliability can be measured for each test subject individually ("within subjects"), or between all observers. Per [5]:

> " Intra-individual reliability refers to the agreement between a certain subject's repeated ratings of the same test condition. Inter-individual ("between subjects") reliability refers to the agreement between different subjects' ratings of the same test condition. "

Validity however is an orthogonal concept: it only relates the elicited scores to the true opinion of the observer during rating, and thus is the "agreement between the mean value of ratings obtained in a test and the true value which the test purports to measure" [5]. Possible lack of validity as a key issue of tests where ratings are forced.

3.1.2 Cleaning Experimental Data

Standardized QoE testing protocols define methods to cleanse the dataset of unwanted results. They primarily focus on the intra-rater reliability, i.e., the reliability that an observer gave ratings that they would be able to reproduce to a certain degree when re-tested. It also refers to the reliability throughout a test session, e.g., a consistent use of the rating scale.

Current standards identify the two main reliability issues for continuous evaluation protocols such as SSCQE [5, 6].

- **Systematic shifts**: This refers to an offset of an observer's overall voting curve when compared to the average. This shift is "systematic" because it is possibly due to a misunderstanding (or misinterpretation) of the rating scale, or a bad construction of the test protocol itself.
- **Local inversions**: While the voting curve may on average correspond to the typical response, there could be intervals in which ratings are inverted or not following the expected trend, possibly due to observers not paying attention.

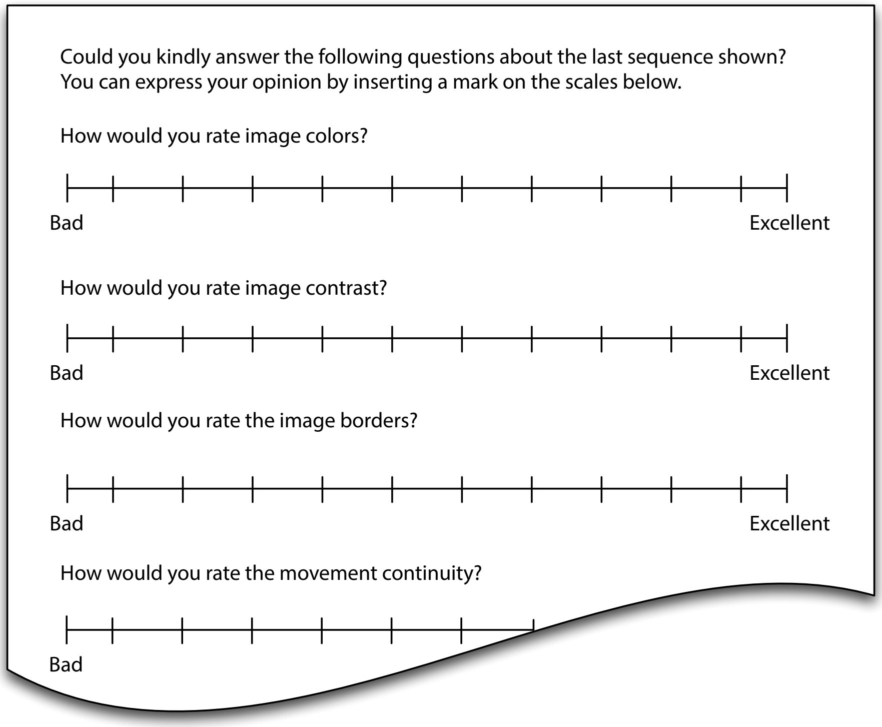
For protocols where a reference is hidden in the set of PVS, additional reliability checks can be included. Intuitively, one would discard votings where the PVS treated with an HRC is rated *better* than the original SRC, however in [5] those are considered valid. During analysis they may still be weighed differently (see Section 2.3.1.2).

For non-continous rating protocols (ACR, DCR/DSIS, DSCQS, ...), a recommended screening method consists of incrementing two counters, P_i and Q_i, per observer i when the rating per PVS is greater or lower than the average rating plus or minus twice the standard deviation of the average rating. Let n be the number of PVS shown to an observer. If the ratio $\left|\frac{P_i - Q_i}{P_i + Q_i}\right|$ is below 0.3 and the ratio $\frac{P_i + Q_i}{n}$ is above 0.05, the observer may be removed from the dataset entirely.

Another means of removing potentially unreliable observer data consists in testing them for visual acuity. The Ishihara test for color-blindness [23] and Snellen charts are mandatory as a pre-test before the actual rating procedure. Observers with deficient color perception or a non-normal vision (even when corrected) are frequently rejected.

3.1.3 Enriching Experimental Data

One could raise the question of whether removing users altogether is beneficial to acquiring representative results from a study. In fact, under the assumption of valid responses, observers that were deemed non-reliable could have merely perceived the stimuli differently. It could be argued that their opinions should be included in the analysis of the data. Of course, when the ratings of all observers are averaged into a Mean Opinion Score (and the associated statistical confidence interval), much of the information that could be gathered about a stimulus is lost. Studies from the field of Human-Computer Interaction show that only a small fraction of possible stimuli attributes can be explained by an average model (see [27], cited after [21]).



Could you kindly answer the following questions about the last sequence shown?
You can express your opinion by inserting a mark on the scales below.

How would you rate image colors?

Bad Excellent

How would you rate image contrast?

Bad Excellent

How would you rate the image borders?

Bad Excellent

How would you rate the movement continuity?

Bad

Figure 4: Exemplary questionnaire for rating quality dimensions, after ITU-T Rec. P.910.

The typical rating scales, whether ordinal or continuous, were necessarily designed to be quick to use, so as to allow a large number of PVS to be tested within a short time frame, and easy to analyze. Finally, they are also easy to report to non-experts in the domain. It is, however, expected that some studies will gather data that shows minimal to no difference in terms of overall quality between stimuli. Especially with votes that indicate medium or bad quality, additional data may be required to allow researchers to identify which feature of the stimulus resulted in the observer's voting decision. In [5], provisions for this are taken, yet they are still rarely used in today's QoE experiments. An example for such a questionnaire can be seen in Figure 4.

Clearly, there is a trade-off to be made between an experiment procedure that specifically asks for distinct quality features at the expense of requiring more time, and a procedure that gives more concise data with the drawback being that the data has less intrinsic value.

3.2 Measuring Confidence

Based on our experience with conducting subjective experiments in the video QoE domain, we found that non-expert experiment participants would sometimes struggle to decide on a rating for a specific stimulus. This would not occur throughout an entire session, but would still be noticeable when observing the assessors. In the written instructions that are given to participants before an experiment with a one-dimensional rating scale, we usually inform participants that they should vote quickly, without giving a lot of thought to what they had just seen, so as not to "over-think" in the process. Yet, the opposite seems to happen on occasion. We therefore hypothesize that users have troubles picking a rating on a presented scale when they are forced to. Also, we expect them to show a delay in rating when they cannot decide for a rating.

Since experiment sessions are unsupervised, it would be preferable to have the assessors explicitly mark the ratings where they struggle to settle for a final score. A simple way to do so is to introduce a self-reporting questionnaire that asks them for the level of confidence they felt while rating. An observer could then choose to give their own rating a different meaning—while at the same time helping facilitate the analysis of data. We therefore designed our protocol to include one additional question per stimulus: "How confident did you feel while rating?"

3.2.1 Social Desirability

A general problem with questionnaires is that respondents may try to "please" the interviewing person in order to give a good impression of themselves. Having to report scores that might shine a negative light on them puts them in a dilemma [9]. This effect is much more pronounced for knowledge questions than statements of attitude. While QoE evaluation is no test of knowledge, failure to give a successful or representative rating could lead to insecurity. Insecurity on the other hand could be interpreted as a sign of weakness, resulting in the aforementioned dilemma for the observer when they are asked to rate their own confidence.

To combat this effect, participants must be informed that negative answers to either question—quality ratings and confidence—will not be punished in any way. Quite the contrary, in our experiments, we would inform observers that a lack of confidence was to be expected for some stimuli and thus encourage them to give valid votes. Another form of encouragement is the retribution of assessors for their successful participation. Retribution can be given in form of credits for academic courses, small goods or vouchers, or even cash. It is important to remind the participants that the retribution will not change depending on their confidence scores.

3.2.2 Using Confidence Scores

How can the obtained confidence scores be used during data analysis? To answer this, consider the typical reporting of ratings from the protocols mentioned in Section 2.3. A list of PVS or HRCs will be associated with the arithmetic mean of the ratings given, the statistical confidence interval and the standard deviation. Values from rating scales may be continuous, e.g. from 0 to 100, or ordinal, from "Bad" to "Excellent". In the latter case, the rating scales are interpreted as interval scales—where each item has the same distance to the next one—although they are in fact just ordinal. While it the comparison of means calculated from ordinal scales is controversial [24], it is common practice for QoE evaluation.

High confidence intervals on MOS can be correlated with a disagreement between observers.[8] They are an indicator of large variations within the collected data and therefore undesirable. Commonly, the disagreement is explained by perception differences of assessors. How can we reduce this variation? For instance, one observer might have given overall preference to a certain genre. This can be alleviated by reporting MOS for each HRC (so-called "condition MOS"), averaging over all SRC

[8] In the scope of this explanation, the concept of inter-rater agreement, commonly calculated through Cohen's and Fleiss' Kappa statistics, is not considered.

shown. Another source of variation would be a misinterpretation of the scale. Such a shift can be corrected for by normalizing the ratings of each observer. Finally, an assessor could prefer some kinds of distortions or artifacts over others. For example, one person could be susceptible to jerky motion more than another viewer. While this variation cannot be eliminated, it could be the basis of the most important secondary findings of a QoE study, apart from the MOS, when focusing on an individual user's perception. Methodologies for such an extended analysis were already mentioned in Section 2.3. Finally, a source of variation would be invalid votes, i.e. where the observer's confidence was low.

Where do we expect confidence to be low? As hypothesized in [16], one would assume that observers have more troubles rating stimuli of mediocre quality content, whereas judging (apparently) perfect quality stimuli should be easier. Knowing whether this is the case for video material—just like for still images—is therefore one of the key aspects of the studies presented in this thesis.

By reliably eliminating low confidence votes from the reports, it would be possible reduce the confidence interval. Likewise, a second type of "confidence interval", based on the confidence data, could be introduced.

3.2.3 Hypotheses

In the remainder of this thesis we want to treat the following hypotheses:

— **H1:** Observers experience a lack of confidence when rating in subjective quality experiments.

— **H2:** Observers rate extremely high quality and low quality content with higher confidence.

— **H3:** The level of confidence can be efficiently measured through means of a self-reporting questionnaire.

— **H4:** The content of a stimulus or the type of distortions have an impact on the confidence.

— **H5:** The time taken to rate is an indicator of lack of confidence.

— **H6:** The self-esteem of the observers influences their confidence ratings.

4 Experiment Design

In this chapter, I will describe the experiment sessions that were conducted in order to collect the data used for analysis and discussion in the following chapters. In total, seven experiment sessions were carried out during the year 2012, six of which at the University of Vienna, Austria, and one at the IRCCyN. All experiments focused on evaluating a set of video clips (without audio) and followed a standard protocol (i.e., ACR or DCR, as explained in Chapter 2). I was personally involved in the design and conduction of every experiment except for *DNA Watermarking*.

In addition to asking for quality scores, we let participants report their level of confidence on a separate scale. Confidence levels were taken after each quality rating, thus there are as many confidence scores per PVS as quality scores. As described in Chapter 3, we wanted to de-emphasize the importance of quality scores alone. Rating the confidence should not be seen as a side-task by the observers. Thus, both tasks were given the same importance when instructing participants. We were explicit about reminding them that a lack of confidence was to be expected for some videos. A negative confidence score would not be negatively judged by the experimenters, and would have no impact on the final evaluation, or even the monetary retribution of the participants.

4.1 Experiment 1: DNA Watermarking (WM)

4.1.1 Motivation

Watermarking is a technique that allows content creators to protect images or video material by altering the data that is sent. This digital signature is hidden among the data. It is imperceptible by the human eye and cannot be easily removed: it would require extensive modification of the video.

The watermarking procedure used here involves generating two strands of video from the original source. These strands are created from the uncompressed sources of the video material, which were modified so as to introduce the watermarks. Then, a fingerprint of binary symbols is randomly generated for each newly distributed video. For the final video that is sent to the user, one Group of Pictures (GOP) from each strand is chosen in sequence. For example, let the fingerprint be `100110…`, then the GOPs in the final stream will be $G1_1, G2_0, G3_0, G4_1, G5_1, G6_0, \ldots$, where in Gn_s, n is the GOP number and s denotes the strand from where it is taken.

The existence of a certain watermark in a video stream allows identification of illegally distributed content, but it comes at a price: the size needed to store two or more strands for each source video will grow. Not all content providers can triple their available storage space without increased costs.[9] To compensate for this, the average bitrate of the strands could be lowered. This results in degradation of overall video quality.

The experiment, which is described in more detail in [15], evaluated two scenarios:

- The storage space cannot be increased. How much degradation in quality can be expected when watermarking the content?
- The storage space can be increased. What amount of additional space is necessary to preserve the original quality of the video, when embedding watermarks?

4.1.2 Source Material, Treatment and Conditions

Seven video clips were used as source material. Each SRC had dimensions of 640 × 480 pixels (VGA) and a duration of 12 seconds. They were shot at 30 Hz frame rate and stored in the YUV 4:2:0 color space.

[9] Three times as much space is needed when the original video is to be kept alongside two newly generated strands with the same quality.

In order to test the two scenarios mentioned above, 12 HRCs were set up. They differ in resolution and bitrate allocation for each strand (see [15] for details). In two HRCs, no watermarking was applied for reference purposes. The x264 encoder was used to generate the PVS, with a fixed GOP size of 12. Overall, this resulted in 84 PVS for this experiment.

4.1.3 Test Protocol and Observers

The test followed the ACR protocol [5], allowing the users to rate the quality of the content on a five-level scale. 11 non-expert observers took part, mostly students aged between 18 and 25 (mean 22.5, $\sigma = 2.34$). There was one female among the participants. They saw the PVS on a 24 inch consumer LCD computer screen, in their original VGA resolution, with a viewing distance of 5H. The experiment session lasted 25–30 minutes in average.

4.2 Experiment 2–3: Foreground-Background Separation (`FB ACR / FB IS`)

4.2.1 Motivation

In mobile TV applications, or mobile video consumption in general, the content is often presented on screens that are smaller in size, as compared to a static consumption scenario, e.g., in the living room or cinema. Current mobile data transmission techniques like 3G networks also impose restrictions on the available streaming bandwidth. Packet losses or sudden network availability changes may require the video player to reduce the dimensions of the shown video from HD to vertical resolutions of 480 pixels or less. While technically, this can be achieved without interruption of the video stream (e.g., through adaptive streaming technologies), the visual quality of the material will be degraded, if only due to the reduced size.

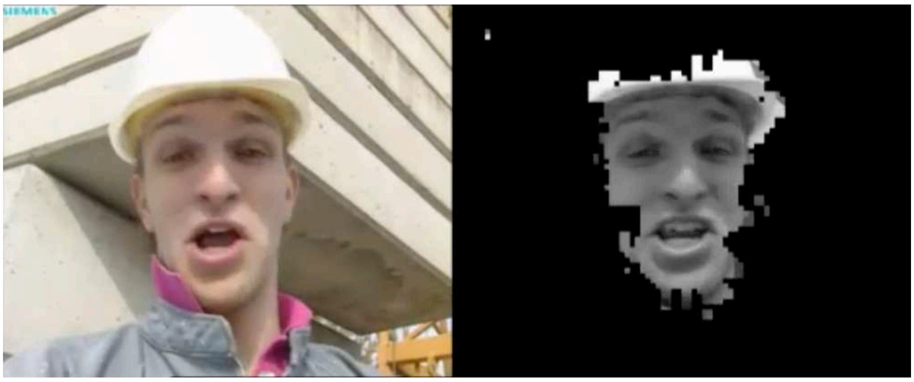


Figure 5: Foreground–Background separation algorithm results on the *Foreman* video. Reference frame is on the left; detected foreground macroblocks are shown on the right.

In consumption scenarios where only limited bandwidth is available, the source material could be encoded in such a way that the more important parts of the video are allocated more bits than the rest. In practice, this can be achieved by separating the foreground from the background, e.g., a person standing in front of a static scene. This would result in better quality for foreground objects, while distortions and compressions artifacts would be more common in the background [12].

We implemented an algorithm that extracts foreground objects from their background by analyzing the motion vectors generated by the JM H.264 reference encoder.[10] This results in two video planes (foreground, background) that now can be encoded with different quality settings in a second pass. An example with the *Foreman* video[11] can be seen in Figure 5.

4.2.2 Source Material, Treatment and Conditions

11 SRC videos in VGA resolution (640 × 480) were used for this test, taken both from VQEG sources[12] (*Mobile Calendar*) as well as video sequences produced at the University of Vienna [44] (*Handball*, *Ice Hockey*, *Weather Forecast*, *Running Dog*, *Two People Walking 1*, *Two People Walking*

[10] `http://iphome.hhi.de/suehring/tml/`

[11] available from `http://trace.eas.asu.edu/yuv/`

[12] `ftp://vqeg.its.bldrdoc.gov/MM/vga/`

HRC	Foreground	Background
1	26	26
2	26	32
3	26	38
4	26	44
5	32	32
6	32	38
7	32	44
8	38	38
9	38	44
10	44	44

Table 4.1: Foreground and background QP levels for each HRC.

2, *Birthday Party*, *House Zoom In*, *Surfers* and *Run*). They were rendered at 25 frames per second and each had a duration of 10 seconds.

The HRCs consisted of a combination of fixed quantization parameters (QP) for foreground and background, respectively. The quantization parameter affects the amount of spatial detail being preserved when encoding the video. Lower values result in better visual quality and vice-versa. We generated a total of 10 HRCs, with the QP settings listed in Table 4.1. HRC 1 corresponds to a normal encoding process at high quality, without affecting the background. Likewise, HRC 10 results in a low quality sequence for both planes. Since each clip was treated with every HRC, a total of 110 PVS were shown to each observer.

4.2.3 Test Protocol and Observers

The foreground-background experiment followed two different protocols, as it was carried out in two sessions. We thus label them as experiments 2 and 3.

A total of 14 participants were tested. Half of the tests followed the ACR protocol. The other half used a variation of the DCR method [5], where the PVS were shown sequentially (as in ACR), but with the Impairment Scale (IS) featured in DCR (“Imperceptible”, “Perceptible, but not annoying”, "Slightly annoying", "Annoying", "Very Annoying"). We chose the IS in order to obtain potentially more meaningful results with regard to the visible impairments in the processed sequences rather than the overall absolute quality.

12 males and two females were tested. Six males and one female were assigned for the ACR session, the remainder to the IS session. The mean age was 25.1 ($\sigma = 5.27$) and 27.9 ($\sigma = 11.13$), respectively. The tests were carried out on a 22 inch consumer LCD monitor, at 5H viewing distance. The videos were shown in their original resolution on a medium grey background. Each experiment lasted about 25–30 minutes in average.

4.3 Experiment 4–6: SVC Compression (SVC1 / SVC2 / SVC3)

4.3.1 Motivation

Scalable Video Coding (SVC) is an extension of the H.264/MPEG-4 Part 10 video coding standard [47]. It provides facilities for streaming video over channels where loss may be expected. SVC allows the simultaneous transmission of temporally, spatially or quality-scaled streams together with the original. It also exploits temporal and spatial redundancies between those streams to increase compression efficiency, e.g., by allowing hierarchical prediction of pictures.

The original experiments can be found in [35]. From these, we took the SRCs and HRCs to conduct the sessions described in this Section. The experiments aimed at finding out the influence of various combinations of QP levels assigned to the individual scalability layers. One issue with the original experiments was that the number of possible combinations of SRCs and HRCs would have required each observer to conduct not fewer than twelve test sessions of 30 minutes each—this would be hardly practical. The results in [35] were therefore based on the ratings given to a subset of all generated PVS. Still, four sessions were needed for each observer, which in practice can be too many.

Going a step further, we developed a method to select a limited number of PVS from a large pool, so that the most representative stimuli could be shown to observers in one session only [36]. This process is called "instance selection". Based on the three instance selection techniques described in [36], we therefore conducted three experiments featuring the PVS selected from our algorithms. The three algorithms used were called "resample", "reservoir" and "spread" and in the following are labeled as SVC1 through SVC3. It should be stressed again that with this setup, not every observer saw every PVS.

4.3.2 Source Material, Treatment and Conditions

We used 11 SRC clips in VGA (640 × 480) resolution with a duration of 10 seconds each: *ShadowBoxing*, *BoxingBags*, *Stream*, *Aspen*, *MesaWalk*, *rbtnews*, *PowerDig*, *SkateFar*, *Family*, *HighWay* and *HalfTimeWide*. The SRC are available from the IRCCyN/IVC SVC4QoE Replace Slice Video VGA database.[13] 24 HRCs were generated by applying SVC coding and AVC coding to the original sources.

- **SVC HRCs:** The SVC streams contained one base layer at QVGA resolution (320 × 240) and an enhancement layer in VGA. Quality levels were introduced by setting the QP of both layers to the values 26, 32, 38 or 44 in various combinations, resulting in 16 SVC HRCs.
- **AVC HRCs:** For each QP value, we included four HRCs where the QVGA base layer would be upscaled to VGA, as well as four HRCs in native VGA resolution.

For the purpose of our experiments, a test session should only last about 25 minutes[14]. We also estimated that each vote takes about 5 seconds. Including an additional waiting time of 3 seconds between each PVS, this sums up to 18 seconds per PVS. Therefore, one can show 83 PVS per session, while our experiment design had generated 275 PVS.[15]

[13] http://www.irccyn.ec-nantes.fr/spip.php?article769

[14] This is the effective time for presenting the content. Additional time for instructions, training and debriefing is not factored in here.

[15] This number includes the reference stream, thus $(24 + 1) \times 11 = 275$.

4.3.3 Test Protocol and Observers

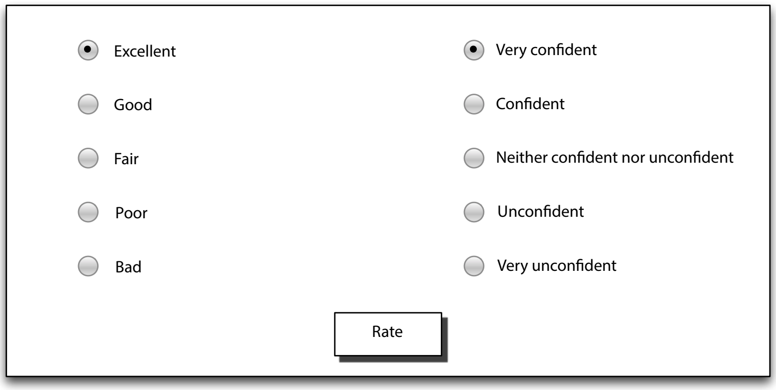
For the three techniques each, 10, 11 and 10 observers were shown the videos following the ACR protocol. On average, the participants were aged 22.1 ($\sigma = 2.62$), 24.5 ($\sigma = 4.50$) and 25.1 ($\sigma = 8.22$), respectively. The material was shown on a 22 inch consumer LCD display at 5H viewing distance.

4.4 Experiment 7: SVC Error Concealment (SVC EC)

4.4.1 Motivation

Based on the Scalable Video Coding experiments described in the previous section, we chose to focus on the effects of error concealment techniques for SVC transmission. Transmission errors can be expected in a lossy transmission environment. These would lead to reduced information in the stream. If the amount of transmitted data is not enough to show the enhancement layer, it would require the decoder two switch from one of the enhancement layers to the base layer—provided a good enough transmission rate. This comes at a cost of visual degradation of the complete stream shown to the viewer, even if only parts of the transmitted pictures might have been affected by the loss. Instead of showing the base layer entirely, the decoder could also try and merely conceal impairments in the higher SVC layers by taking information from the base layer.

Contrary to the others, this experiment was conducted at the IRCCyN in France. We used a Samsung Galaxy Note 10.1 tablet device to present the videos in order to find out whether there are systematic shifts in terms of video quality recognition and confidence values (compared to the "regular" viewing and testing contexts that involve a PC or TV screen). We also adapted the confidence scale for this experiment to a five-point scale from "Very confident", "Confident", "Neither confi-



● Excellent
○ Good
○ Fair
○ Poor
○ Bad

● Very confident
○ Confident
○ Neither confident nor unconfident
○ Unconfident
○ Very unconfident

Figure 6: Interface used for ACR and confidence rating.

dent nor unconfident", "Unconfident" to "Very unconfident" to align it with the five-point ACR scale. A graphic representation of the voting interface is shown in Figure 6.

4.4.2 Source Material, Treatment and Conditions

The test material for this experiment is based on the IRCCyN/IVC SVC4QoE Replace Slice Video VGA database[16], with 9 SRC videos in VGA resolution (640 × 480) encoded at 30 Hz framerate: *Aspen*, *BoxingBags*, *HalfTimeWide*, *Highway*, *MesaWalk*, *Powerdig*, *rbtnews*, *ShadowBoxing* and *SkateFar*. Each video had a total duration of 10 seconds.

To generate the HRCs, a transmission was simulated by removing slices from the source bitstreams based on a loss simulator. The loss was applied in such a fashion that only one slice out of four in a picture was removed, and that visually important regions of a scene would be affected. Hence, there was no fully randomized error pattern. The duration of the error was one second. Then, the base layer was encoded at either 15 or 30 Hz framerate, with bitrates of 120 and 200 kBit/s, respectively.

In order to reconstruct the stream from the distorted transmission bitstream, three techniques were applied:

[16] `http://www.irccyn.ec-nantes.fr/spip.php?article768`

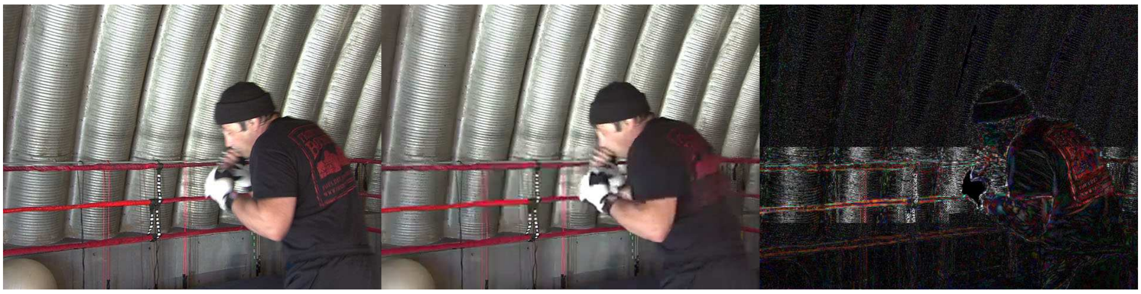


Figure 7: Patched slice for error concealment. From left to right: reference frame, patched frame, difference between frames.

- **Upscaled:** The base SVC layer with a resolution of QVGA (320 × 240) is taken and upscaled to VGA with a Lanczos filter. In the case of a 15 Hz base layer, frames are doubled to achieve the same framerate as the enhancement layer. This concealment is applied to the entire video sequence after detecting an error.
- **Patched:** The missing slice in the enhancement layer is "patched" by taking the equivalent base layer slice and inserting its pixels into the decoded picture. This has the effect of inserting low visual quality content into an otherwise high quality frame. An example of patching with SRC 8 / HRC 6 is shown in Figure 7.
- **Switched:** All frames that are missing a slice are replaced by an upscaled version of the base layer frame. The remainder of the video shows the enhancement layer.

Two additional HRCs were introduced for comparison: a non-damaged SVC transmission at 600 kB/s, and a damaged H.264/MPEG-4 AVC transmission. In the latter, simple buffer repetition was used as error concealment. The full list of HRCs, including the base layer settings as well as the transmission technique, are given in Table 4.2. Here, HRC 0 is the reference stream. A total of 135 PVS were shown.

HRC	Hz	kB/s	Concealment
0	30	n.a.	none
1	30	600	none
2	15	120	upscaled
3	15	120	patched
4	15	120	switched
5	30	120	upscaled
6	30	120	patched
7	30	120	switched
8	15	200	upscaled
9	15	200	patched
10	15	200	switched
11	30	200	upscaled
12	30	200	patched
13	30	200	switched
14	30	600	none

Table 4.2: HRCs generated for the SVC Error Concealment experiment.

4.4.3 Test Protocol and Observers

27 users participated in the SVC Error Concealment experiment, 12 male and 15 female. Their average age was 32, ranging from 19 to 49 ($\sigma = 11.13$). The test followed the ACR procedure, including additional questionnaires (see also Figure 8). We implemented the presentation of the stimuli with a custom interface for the Android operating system. The tablet allowed the viewers to conduct the experiment at approximately 4H viewing distance, however they could slightly move the device while sitting and watching the videos.

Since presenting 135 PVS, including voting, would have required an extended session of 40 minutes or longer, observers had to take a small break of five minutes at the middle of the test set.

4.4.4 Reaction Time Measurement

Similar to the previous research done by Engelke et al. [16], we recorded the timestamps of various actions during the rating session, for each PVS shown:

- The time when the voting screen appeared (t_s)
- The time when the quality rating was chosen (t_q), relative to t_s
- The time when the confidence rating was chosen (t_c), relative to t_s
- The time when the confidence rating was chosen (t_{c2}), relative to t_c
- The time when the overall rating was finished (t_r), relative to t_s

In order to prevent a possible bias, observers were not told about the measurements. Due to the implementation of the rating interface with the Android operating system, we were able to record the timestamps with a precision of less than a few milliseconds compared to the stopwatch-based method in [16], which we believe to introduce measurement errors caused by the experimenter's reaction time.

4.4.5 Survey Assessment

For the SVC Error Concealment experiment, we introduced additional surveys beyond the default questionnaires usually found in recommendations (e.g., [5] or [6]). Specifically, we assessed the self-esteem of the observers directly before and after the rating session, as well as a pre- and post-test survey related to the participant's previous experience, task load and confusion. The entire process can be seen in Figure 8.

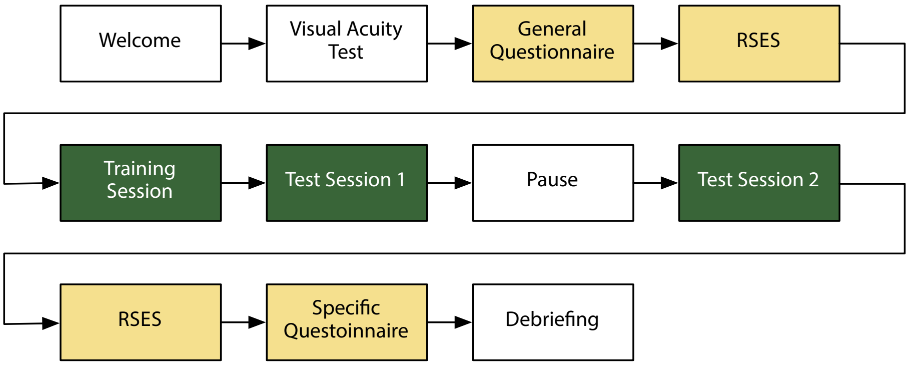


Figure 8: Test process for the SVC Error Concealment experiment.

4.4.5.1 Rosenberg Self-Esteem Scale

The Rosenberg Self-Esteem Scale is a measure of self-esteem, consisting of ten questions to be answered on a four-point Likert scale [45]. The questions consist of five positive and five negative statements about the subject's personality. The subject is then asked to agree or disagree with those. Full agreement with a positive statement is awarded 4 points, and full agreement with a negative statement -4. The sum of points constitutes the RSES score, with a maximum of 40. The questions asked are the following:

- *On the whole, I am satisfied with myself.*
- *At times, I think I am no good at all.*
- *I feel that I have a number of good qualities.*
- *I am able to do things as well as most other people.*
- *I feel I do not have much to be proud of.*
- *I certainly feel useless at times.*
- *I feel that I'm a person of worth, at least on an equal plane with others.*
- *I wish I could have more respect for myself.*
- *All in all, I am inclined to feel that I am a failure.*
- *I take a positive attitude toward myself.*

Since our experiment was carried out in Nantes (France), we translated the RSES to French [50]. We expected the RSES to be a very repsonsive indicator, i.e., very likely to give different results when used right before and after treatment and thus indicate an effect thereof.

4.4.5.2 Pre- and Post-Test Survey

Before the rating procedure and the RSES test, we asked participants the following questions:

- *Have you been in other experiments before?*
- *If yes, how many experiments?*

After the rating procedure and the subsequent RSES test, we gave observers another set of questions to answer with Likert-style responses ("strongly agree", "agree", "neither agree nor disagree", "disagree", "strongly disagree"):

- *I felt confident about my ratings during the whole test.*
- *I don't think my ratings are representative compared to others.*
- *These tasks were mentally demanding.*
- *I sometimes did not know what quality to choose.*
- *After rating a few videos, it was easier for me to rate the following videos.*
- *These tasks were stressful.*
- *I sometimes did not know what level of confidence to choose.*
- *I think I accomplished what I was asked to do.*
- *These tasks were irritating.*

Above items include questions from the NASA TLX questionnaire[17] which focuses on perceived task load.

[17] http://humansystems.arc.nasa.gov/groups/tlx/

4.5 Experiment Overview

This section features a tabular overview of the experiment variables (see Table 4.3) and observer demographics (Table 4.4).

Experiment	Session	Observers	SRC	HRC	PVS	Method
DNA Watermarking		11	7	12	84	ACR
Foreground-Background Sep.	1	7	11	10	110	ACR
	2	7	11	10	110	IS
SVC Compression	1	10	11	25	83	ACR
	2	11	11	25	83	ACR
	3	10	11	25	83	ACR
SVC Error Concealment		27	9	15	135	ACR

Table 4.3: Experiment variables overview. HRCs include reference stream.

Experiment	Session	⌀Age	Min Age	Max Age	σ Age	♀	♂
DNA Watermarking		22.5	18	25	2.34	1	10
Foreground–Background Sep.	1	25.1	20	36	5.27	1	6
	2	27.9	18	48	11.13	1	6
SVC Compression	1	22.2	18	28	2.62	4	8
	2	24.5	21	37	4.50	3	9
	3	25.1	20	51	8.22	5	8
SVC Error Concealment		32.0	19	49	11.13	12	15

Table 4.4: Detailed demographics overview.

5 Experiment Results

In this chapter, the results of the experiments described in the previous chapter will be presented. A brief overview of the collected data will be given. We will then highlight the results of individual experiments with regard to criteria commonly evaluated in the domain of QoE. Specific hypotheses defined in Chapter 3 will be tested. Also, the findings will be compared to those of [16].

5.1 Scale Usage

First, we want to take a look at the global data obtained in the experiment sessions. By giving a holistic view on the quality and confidence scores, we can observe general effects and check whether our method elicited meaningful data. Inspecting overall distributions is commonly being done for QoE experiments to get a first impression of the results.

5.1.1 Quality

One of the first steps in evaluating subjective QoE experiment data is to check the distribution of the quality ratings. A well-designed and well-prepared experiment should result in a (close to) uniform distribution of all scores, or exhibit a low statistical skewness. The reason for this is that each quality level should be equally represented. We can visualize the rating distribution with a density plot of the scores. Figure 9 shows such a plot for all experiments.

As clearly visible, the DNA Watermarking experiment suffers from a skewed distribution of quality scores, with few PVS in the low quality ranges. The Foreground-Background experiment PVS are

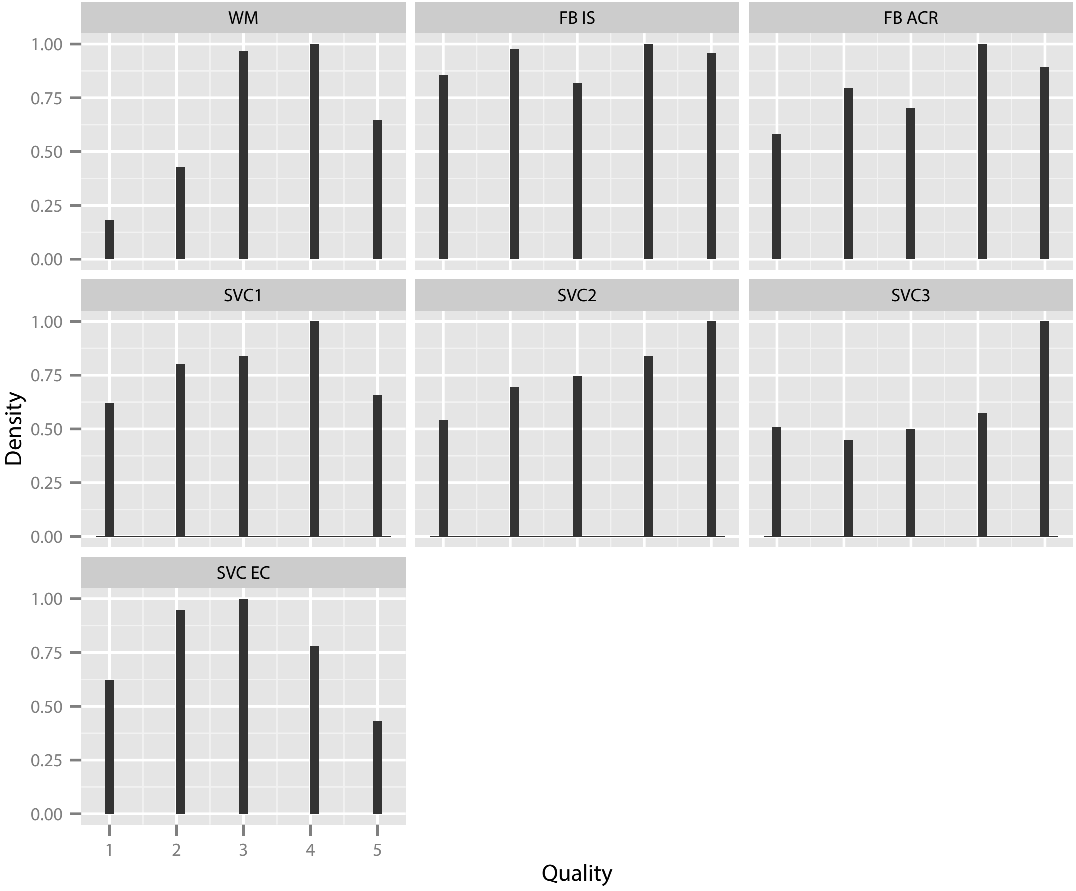


Figure 9: Density plots for quality scores across all experiments.

well distributed. The SVC3 experiment appeared to include too many high quality PVS, but this was to be expected from the algorithm that was used to determine the PVS (see [35] for an explanation). Lastly, the SVC EC experiment shows a slightly left-skewed distribution with fewer extreme values than others. This over-representation of medium quality content however does not affect the evaluation.

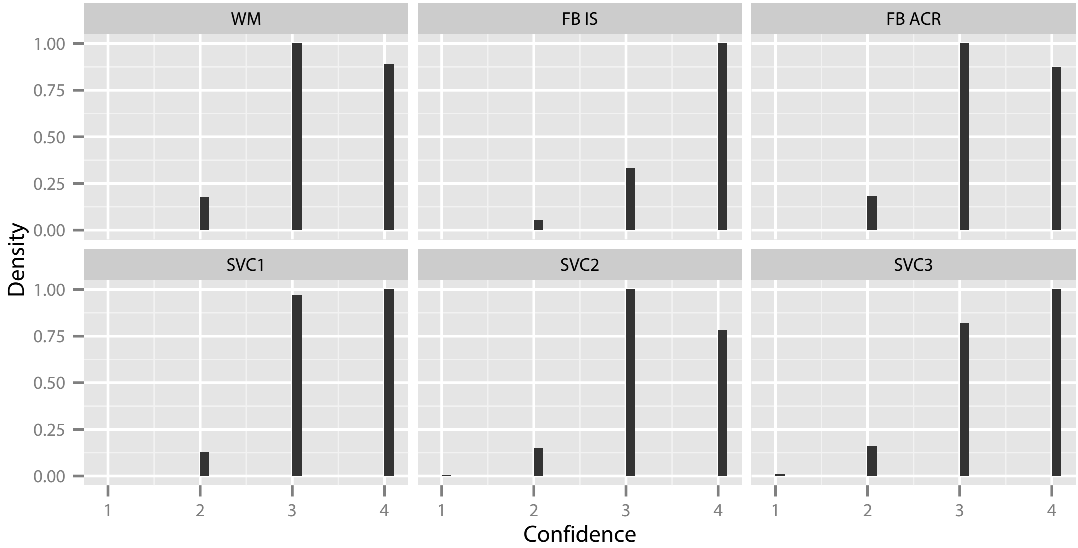


Figure 10: Density plots for confidence scores in experiments with four-value scale.

5.1.2 Confidence

For the confidence voting, a four-value scale was chosen for the experiments conducted in Vienna, and a five-value scale was used for the SVC Error Concealment experiment in order to align it with the ACR scale. Looking at the global voting data, we want to answer the question of how confident the assessors generally felt. The plots for each experiment with the four-value scale are shown in Figure 10.

As can be seen, most users felt confident or very confident for the ratings they gave. In fact, the first option ("very unconfident") was only chosen 11 times out of 4938 votes in total (0.22%). The "confident" option dominates, with the exception being the Foreground-Background experiment that used the Impairment Scale and the third SVC experiments. We will discuss this finding later. The data is summarized in Table 5.1.

Confidence	WM	FB IS	FB ACR	SVC1	SVC2	SVC3
very unconfident	1	1	1	0	3	5
unconfident	78	30	68	51	71	61
confident	447	184	374	380	465	308
very confident	398	555	327	391	363	376

Table 5.1: Confidence scores in experiments with four-value scale.

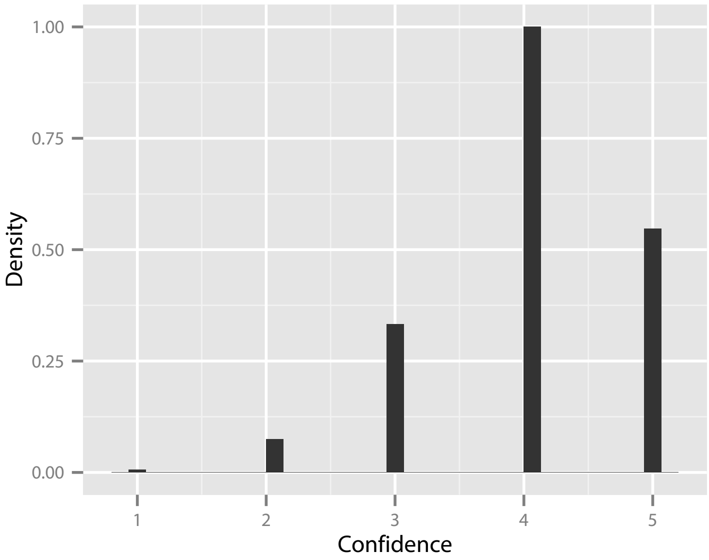


Figure 11: Density for confidence scores in SVC EC experiment.

For the five-value confidence scale used in the SVC Error Concealment experiment, the obtained data appears similar, with the option chosen most of all being "confident". Like in the previous experiments, the first option ("very unconfident") was only chosen 11 times out of 3510 votes (0.31%). The density plot for the SVC EC experiment is shown in Figure 11 and the raw data in Table 5.2.

5.1.2.1 Suitability of the Rating Scale

Visually comparing the two scales in their usage we can observe that the inclusion of a middle option did not necessarily correspond to an "undecided" item commonly used in Likert scales which users

Confidence	Count	Percentage
very unconfident	11	0.31
unconfident	133	3.79
neither confident nor unconfident	596	16.98
confident	1790	51.00
very confident	980	27.92

Table 5.2: Confidence scores for the SVC EC experiment

would choose if they could not settle for anything else. Quite the contrary, the overall shape of the distribution remains similar. This is an indicator that confidence is internally treated by participants as a continuous measure rather than an absolute categorical value that could be easily assessed with an ordinal scale such as the one used in our experiments. We therefore cannot fully accept our hypothesis H3, in which we stated that it would be possible to efficiently measure confidence through a questionnaire. A psychometric response scale such as the Visual Analog Scale (VAS) could potentially be more efficient when collecting and evaluating data [38].

5.1.2.2 Comparison With Previous Experiments

In [16], Engelke et al. report data for their subjective image quality experiment. Assessors showed an average confidence of 4.252 ($n = 1200$, $\sigma = 0.773$) for distorted images and 4.600 ($\sigma = 0.636$) for reference images. Since it also used a five-point ordinal scale, we compare these values with the results from our SVC EC experiment, only looking at distorted HRCs. On average, the confidence was 3.977 ($n = 2808$, $\sigma = 0.768$), which is significantly different from the data obtained by Engelke et al. ($p < 0.0001$ for a two-tailed t-test). All other things being equal with regard to the test instructions, this significant difference can only be explained by two factors:

1. the different wording of the rating scale (“very confident” and “confident” in our tests versus “high” and “low”)

2. the usage of video material instead of still images, which could cause additional confusion during voting

Further experiments are needed to explore these possible reasons. We can, however, conclude that there is indeed a lack of confidence experienced by observers as hypothesized in H1.

5.2 Correlation Between Quality and Confidence

One of the main aspects of this work is to highlight the influences of observer confidence on the quality ratings and vice-versa. In this section, we focus on these two measurements. In the following, we use the abbreviation MCS (Mean Confidence Score) to denote the average confidence ratings given to a PVS. It is therefore calculated like the MOS.

5.2.1 Average Quality vs. Confidence Ratings

We hypothesized (H2) that users would be less confident while rating content of mediocre quality, i.e., anything between the extreme values of a rating scale such as “Bad” and “Excellent”. The rationale for this is that perfect fidelity of a visual stimulus is considered “normal”. When the stimulus is unimpaired, the human visual system (HVS) will not be able to identify any unnatural distortions. Lack of distortions or artifacts can therefore easily be translated into the highest possible value on the rating scale. Likewise, we expect a certain amount of distortion to trigger a reaction that results in the HVS quickly marking a stimulus as unacceptable in terms of quality. While *acceptability* does not necessarily correspond to the same dimension as *quality*, we expected observers to be able to rate bad quality content without a lack of confidence.

To test our hypothesis H2, we look at the MOS and MCS per PVS, broken down by experiment session. Figure 12 shows these results. The SVC EC responses, where a five-point confidence scale

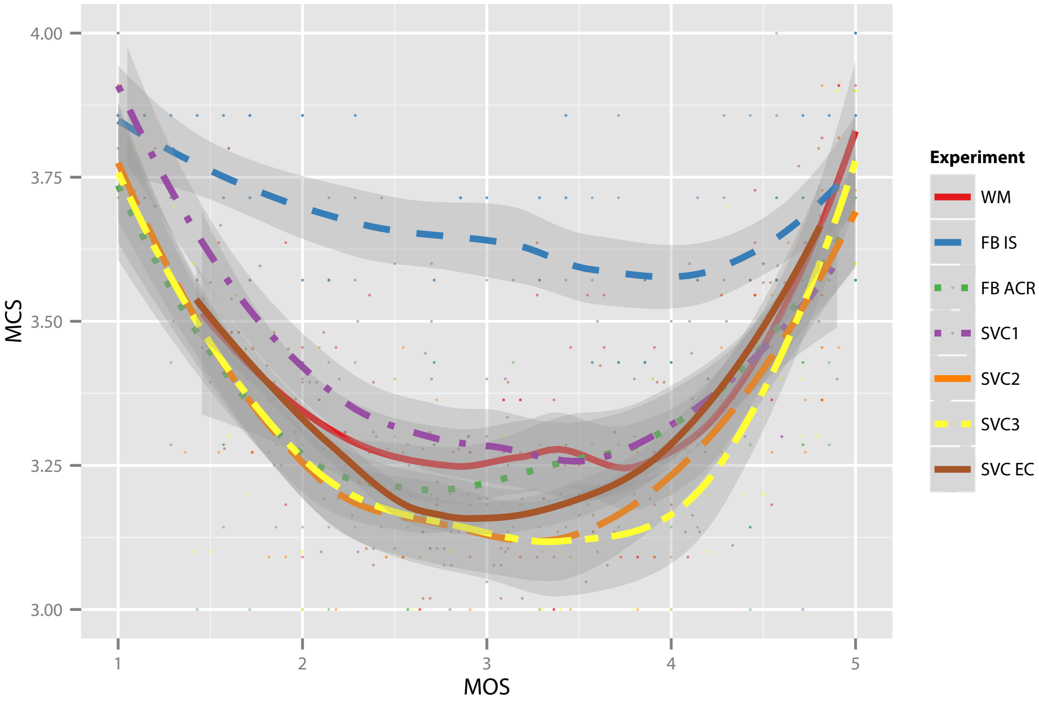


Figure 12: MOS vs. MCS for all experiments with LOESS fit.

was used, were re-scaled to a four-point scale. As expected, the confidence drops when reaching the middle of the quality scale.

We can also express the MCS as a function of mean absolute difference from the middle element of the quality scale. Let us define the MMOS of a PVS as:

$$MMOS = \frac{\sum_{n=1}^{N} |3 - Q_n|}{N}$$

where N is the number of observers and Q_n is the individual quality rating given to the PVS. For example, if the quality was rated as “Good”, this is equivalent to $|3 - 4| = 1$.

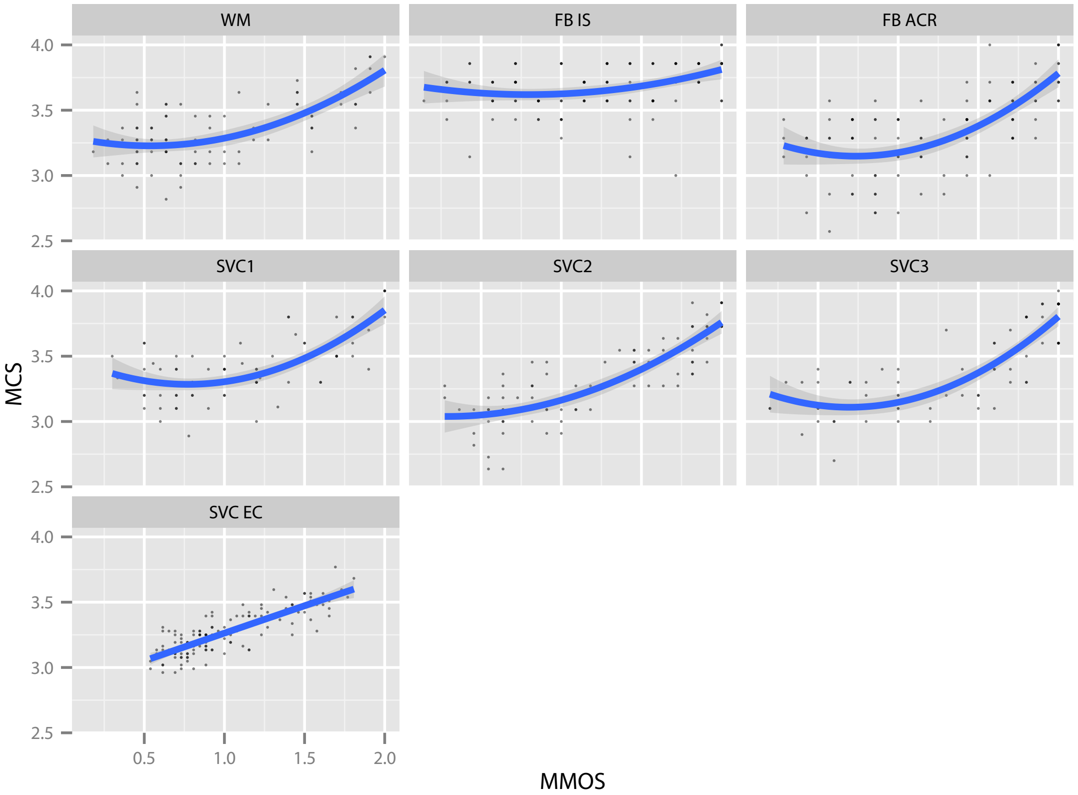


Figure 13: MMOS vs. MCS for all experiments with second-order polynomial fit.

There is indeed a significant correlation between MCS and the MMOS for all experiments ($p < 0.001$). Results for individual experiments are shown in Figure 13. We can observe that for all experiments except SVC EC, a second-order polynomial fit explains the drop in confidence very well.

Table 5.3 summarizes the individual model coefficients as well as an overall model calculated on all experiment results for $y = p_2 x + p_1 x + p_0$.

Experiment	p_2	p_1	p_0	R^2
`WM`	0.54	1.43	3.34	0.49
`FB IS`	0.36	0.57	3.68	0.13
`FB ACR`	0.92	1.99	3.33	0.47
`SVC1`	0.69	1.37	3.41	0.53
`SVC2`	0.49	2.19	3.32	0.69
`SVC3`	0.89	2.09	3.41	0.71
`SVC EC`	-0.01	1.60	3.27	0.69
All	1.69	4.44	3.39	0.42

Table 5.3: Model parameters: MCS as a function of MMOS.

5.2.2 Influence of Rating Scales

The Foreground-Background experiment constitutes a special case: half of the experiment sessions did not use the ACR scale ("Bad" to "Excellent"), but the Impairment Scale (see Section 2.3.2.2). All other experiment variables remained the same. As explained previously, the confidence scores appear much higher for the IS than for the ACR. This is visible in Figure 14. In fact, the mean confidence for the IS experiment is significantly greater than for the ACR experiment ($p < 0.001$).

The difference clearly shows how the wording of the scale influences the (experienced) confidence of the observers during rating. In turn, the confidence has an effect on the accuracy and validity of their votes. The explanation is simple: while the ACR scale requires the participant to translate the subjective impression of the experienced quality into a specific absolute label, the Impairment Scale asks the observer for their *direct* experience: if they saw distortions ("perceptible" vs. "imperceptible") and—if distortions were present—how annoying those were. The IS therefore allows for an easier judgement. Put differently, it is easier for observers to tell that there are annoying distortions rather than deciding that the PVS is "poor" and not "fair".

It is generally not recommended to use the IS in an experimental protocol that does not allow the viewer to see a reference and a distorted stimulus, i.e., a double-stimulus methodology. In fact, the ACR scale was specifically constructed for single-stimulus methodologies. Does this necessar-

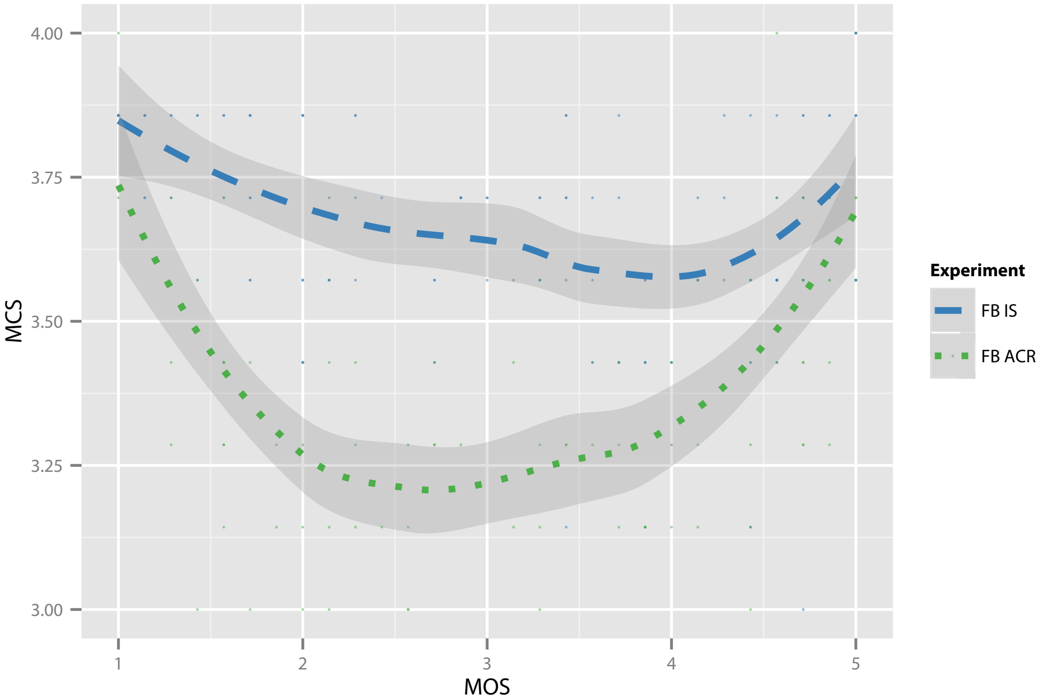


Figure 14: MOS vs. MCS for the two Foreground-Background experiments.

ily prohibit us from using the IS? When comparing the MOS for the FB ACR and FB IS HRCs, Figure 15 clearly shows how the scales can be considered almost equivalent in their usage, as they result in nearly the same condition MOS. Before scaling the individual votes of each user to standard scores—which is what the figure shows—only two HRCs (7 and 9) showed notable differences. As seen in Table 4.1, these HRCs combine high foreground QP values (i.e., low quality) with the highest QP for the background, as opposed to other HRCs where the quality difference is not visible (1, 5, 8) or very extreme (4).

To summarize, it becomes obvious that the choice of absolute scales with categoric labels leads to insecure responses by observers. More experiments are needed to compare the confidence ratings of the IS compared against continuous scales, but we suggest that if the ACR scale is chosen, more

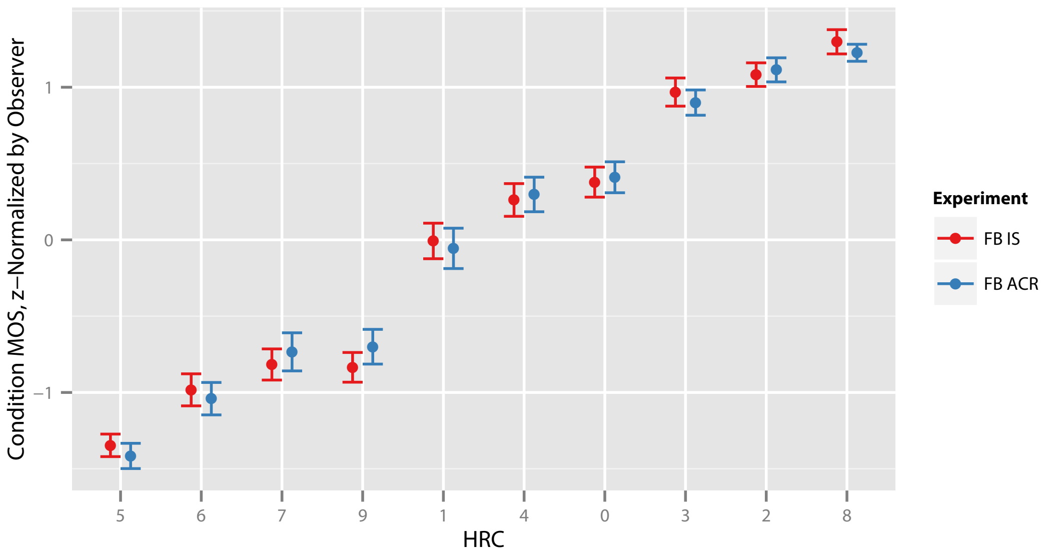


Figure 15: Condition MOS for FB ACR and FB IS experiments, with 95% CI.

care is taken to 1) instruct the observers about the meaning of the scale and 2) consider personal factors during data analysis. For the second aspect, we will show results later in this chapter.

5.3 Influence of Stimuli and Treatments on Confidence

In the previous section, we showed that a lack of confidence strongly correlates with ratings that indicate medium quality content. For a successful evaluation of subjective experiment data, however, it is important to find the specific causes for a drop in experienced quality. In practice, many QoE experiments only study the influence of very specific types of distortion. With a limited number of treatments it is therefore relatively easy to model the effect of a treatment on the perceived quality. On the other hand, larger scale tests such as the Video Quality Experts Group's Multimedia Phase I Validation Test [51] incorporated over 5000 PVS with both transmission and codec compression distortions. In such a case, explaining a certain MOS through the characteristics of an HRC becomes harder.

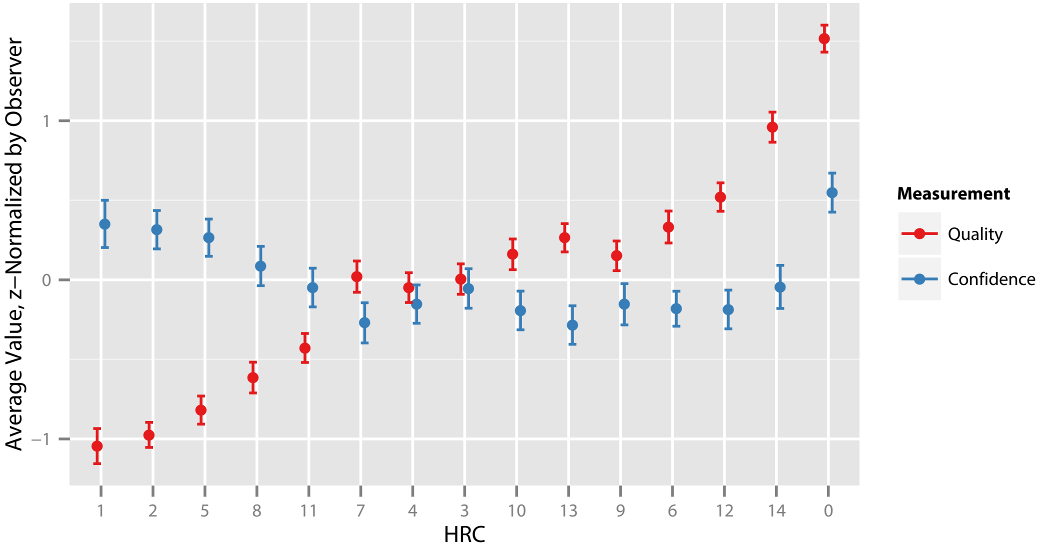


Figure 16: Condition MOS and MCS for SVC EC experiment, with 95% CI.

The first question we wanted to answer was whether specific SRC contents would have an effect on the confidence. In order to answer this, we calculated the correlation coefficients between SI, TI and the MCS for each SRC. We did not find any significant correlation. For example, for the SVC EC experiment, Pearson's correlation coefficient for SI vs. MCS is 0.05 ($p = 0.828$) and for TI vs. MCS is -0.30 ($p = 0.255$). Concluding from this, the overall spatial activity does not seem to influence the confidence at all, and the impact of TI is marginal.

When we repeat this analysis for HRCs, we have to consider the previously found results that show how overall quality affects the confidence. Different HRCs implicitly result in different quality scores since they were created to elicit them—but does the confidence vary depending on the kind of distortion? Figure 16 shows that this does not appear to be the case. The average confidence only slightly changes with no significant differences except for HRC 0, which is the unimpaired reference. To remove possible influences of different scale usage, the figure shows MOS and MCS for each observer scaled to standard scores.

	0%	25%	50%	75%	100%
t_q	0.43	0.81	1.26	2.11	73.85
t_c	0.72	1.96	2.68	3.94	76.35
t_r	1.53	2.60	3.38	4.83	77.27

Table 5.4: Recorded time intervals for SVC EC experiment and their quantiles.

Since neither specific SRCs or HRCs seem to have a notable global effect on the average confidence, it can be concluded that lack of confidence is very specific to certain PVS, only for certain users. We can therefore reject hypothesis H4. The results in the previous section showed that there is a measurable relationship between reported quality and confidence, but with the rating scale (and therefore the test procedure) taking a higher impact than the actual distortions.

5.4 Rating Time and Confidence

For the SVC EC experiment, specific time stamps were recorded (see Section 4.4.4) in order to measure the possible interdependencies between quality, confidence, and reaction behavior.

5.4.1 Distribution for Rating Times

First, we look at the distributions of the rating times t_q, t_c and t_r. These denote the time to rate quality, confidence, and the total time to finish the rating procedure. Each distribution is asymmetric to the right, meaning that the majority of recorded intervals are short, with only few outliers. Figure 17 shows the distributions and Table 5.4 the quantiles of the recorded samples. Notable outliers exist for t_q, where in six instances it took observers more than 20 seconds to give a rating. This delay also leads to the longer tails in the distributions for t_c and t_r, since those are interval measures based on t_q.

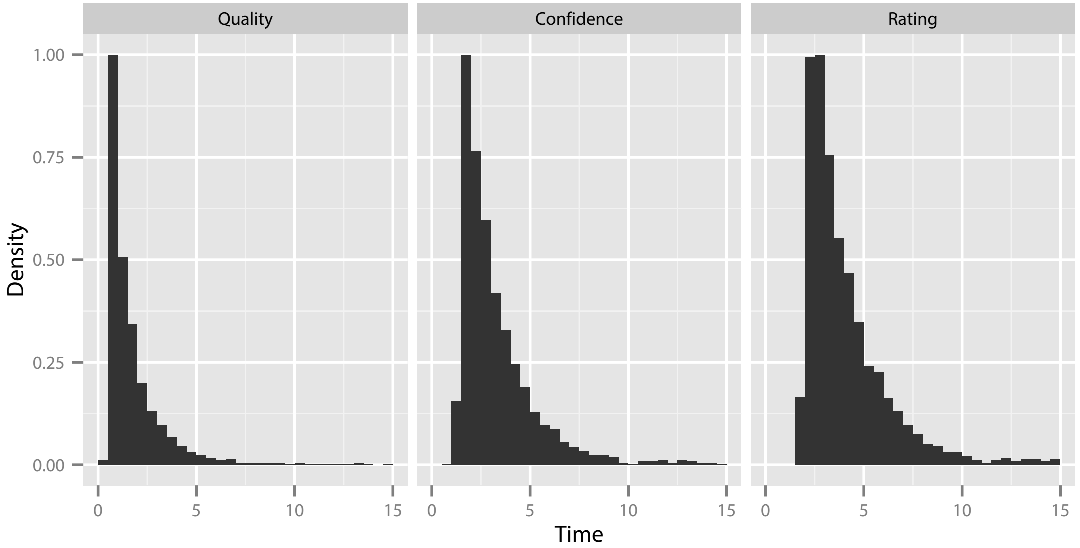


Figure 17: Rating times distribution for quality, confidence, and overall rating.

To further inspect the user behavior, we examined the average time to rate the quality (t_q) for each specific quality and confidence score. Figure 18 shows these results. On the x-axis we differentiate the quality scores 1–5 (“Bad” through “Excellent”) and confidence scores (“very unconfident” to “very confidenct”). The y-axis shows the average t_q for the respective quality or confidence score. We can observe a very large average t_q of 4.89 seconds (±3.31 at 95% CI) for ratings where the observers were “very unconfident”. However, since there are only 11 ratings at this data point, a very large CI is expected. As hypothesized in H5, the average t_q drops with increasing confidence.

The above results are different from those reported by Engelke et al. [16]. The authors—to their surprise—found a *decreasing* rating time for confidence scores 1 and 2, but they also point out that there was only one rating recorded for the lowest confidence. Compared to our findings, where the CI is still large, but at a much higher average t_q, the data point reported in the previous literature could therefore be interpreted as an outlier.

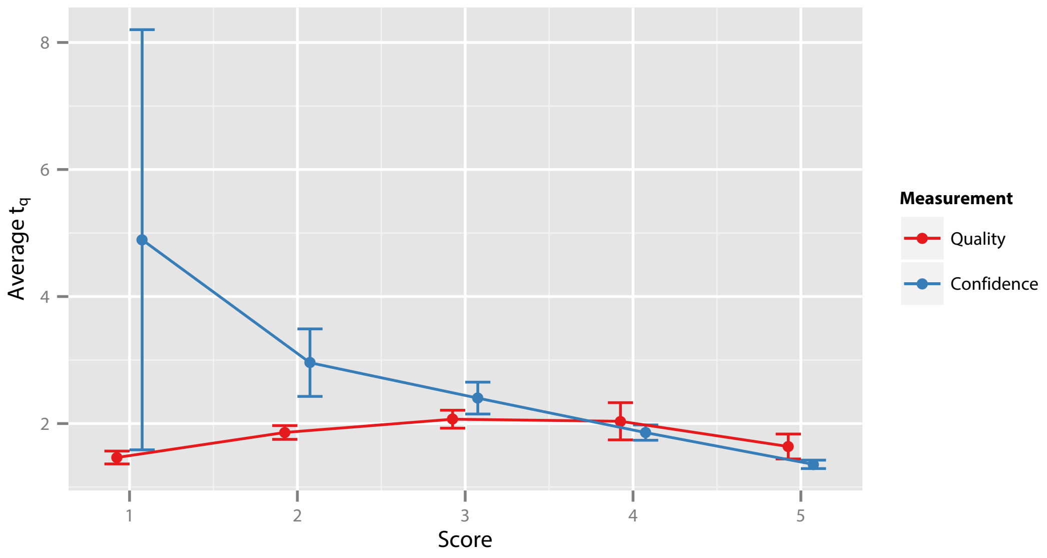


Figure 18: Average quality rating time t_q for specific quality or confidence scores, with 95% CI.

Since about 98% of the quality rating times (t_q) are less than or equal to 10 seconds, we suggest that ratings where $t_q > 10$ are to be pruned from the dataset. This also aligns with the 10 seconds frequently reported in context of the so-called “recency effect”, which explains how observers may “forgive” bad quality content after seeing better quality [5, 19]. It is suggested that for stimuli of 10 seconds length, the recency effect is not present. Taking into account the results shown in Figure 18 we can therefore conclude that for the remaining data, the rating time could be a very good indicator for the observer confidence when it is not feasible to explicitly measure it.

5.4.2 Correlation Measures and Modeling

Rating times can be measured without additional effort. Typically, PVS are presented using computer software, which may implement a precise timer to record the user events. If rating time appears to be a good indicator of confidence, it would be easy to factor in response times during data evaluation. To model the relationship between confidence and rating times, there are two approaches, both of which will be explained in this section.

5.4.2.1 Mean Confidence Score vs. Average t_q

Similar to what Engelke et al. showed, we can summarize the data for each PVS, modeling the MCS (dependent) and the average t_q (independent) using polynomial regression. First, we want to remove extreme outliers from the dataset. When observers took longer than 10 seconds to rate, their scores are not considered. Here, we can only observe a weak correlation of -0.153, which is not statistically significant ($p = 0.0765$), compared to a correlation of -0.697 reported in [16].

This large difference can only be explained by large variations in the observer behavior regarding rating times. However, we are uncertain as to why these variations occurred in our experiment. Therefore, predicting the confidence for a given PVS is not feasible using the voting times of all observers.

5.4.2.2 Modeling Confidence from Individual Rating Time

A second approach would consist in not summarizing the data, but modeling confidence as an ordinal dependent variable and each individual t_q as the independent. Ordinal logistic regression cannot be used in this case, since the confidence scores do not appear to fulfill the proportional odds criterion [8]. We therefore performed a multinomial logistic regression [20] using the `nnet` package in *R*. To visualize the results, we show the probabilities of a certain confidence being chosen at specified values of t_q in Figure 19. Such a model can be used in practice to determine the likelihood of a certain confidence response: the visualization makes it apparent that for quick ratings (below 2.5 seconds), the probability of the observer being confident or very confident is high. Rating times

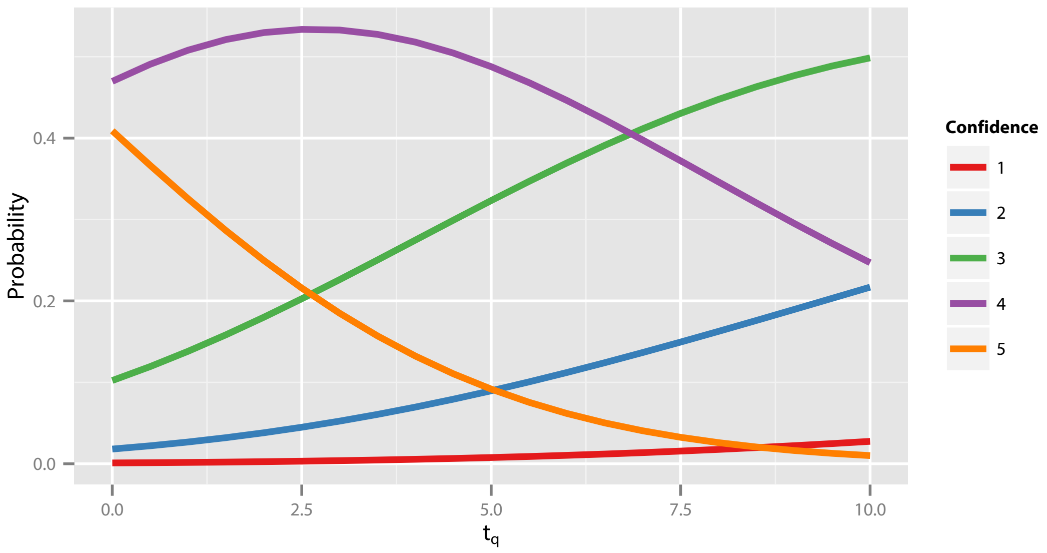


Figure 19: Predicted confidence scores from t_q using multinomial logistic regression.

above 2.5 seconds increase the likelihood of an insecure choice. For $t_q > 6$, there is a relatively high probability that the observer was lacking confidence during voting.

5.4.2.3 Influence of Test Duration and PVS Position

As QoE test sessions can take up to 30 minutes, or even up to an hour (with the inclusion of breaks), we expected participants to change their behavior during time. To prove this, we calculated the average t_q for each PVS in the SVC EC experiment, sorted by the playback order in the test. The results can be seen in Figure 20. In fact, there is a significant drop in rating time towards the end of the test session, with a Pearson correlation of -0.729 ($p < 0.001$). The linear model to predict rating time t_q as a function of position p is:

$$t_q = -0.0059p + 2.12$$

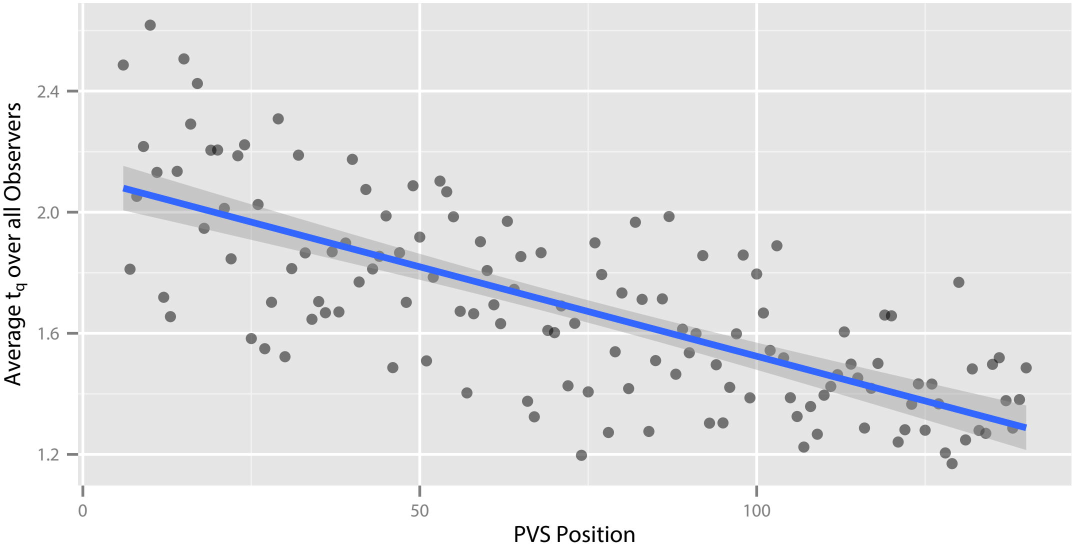


Figure 20: Quality rating times as a function of test progress, with linear fit.

with $R^2 = 0.533$ and $df = 133$.

Since the evaluation of quality is not a knowledge task, we do not expect observers to "learn" how to rate QoE after time. Since they are instructed to view each PVS carefully, a reduced rating time towards the end of an experiment possibly hints at more erratic votes and not reduced confidence, as we would expect given the relationship we identified previously and also taking into account the survey data summarized in the following.

5.5 User-focused Evaluation

The following section contains the evaluation of data that pertains to individual users, the surveys they took, their qualitative responses, and their self-confidence reports.

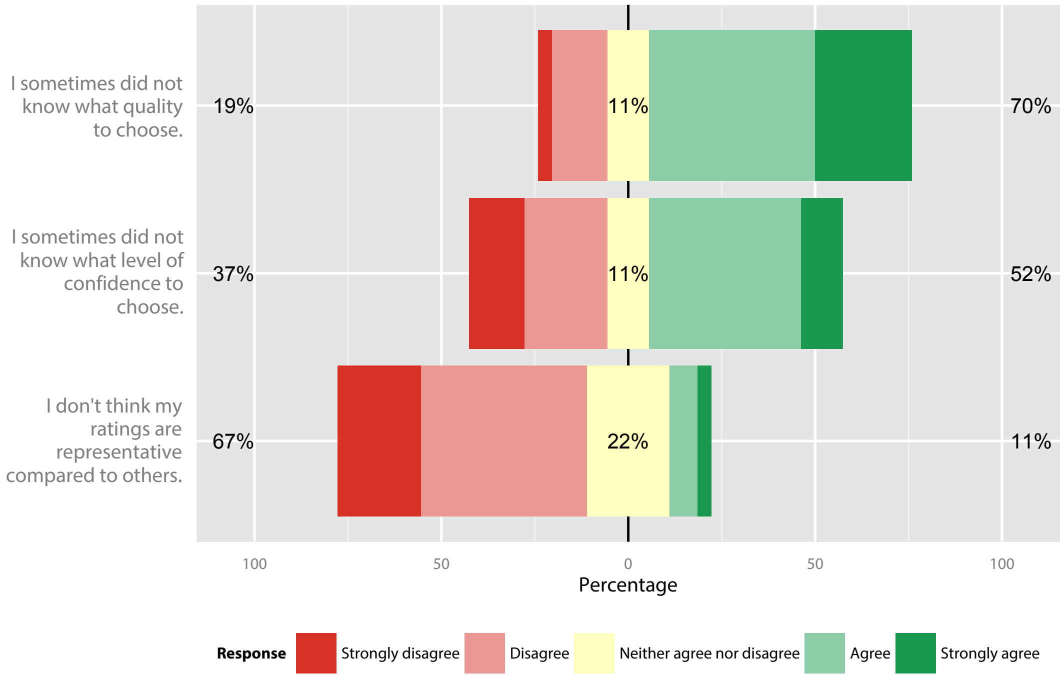


Figure 21: Responses to the post-test survey, part 1.

5.5.1 Survey Analysis

In the SVC EC experiment, we asked observers to complete a pre- and post-test survey (see Section 4.4.5.2) consisting of two and nine questions, respectively. In this section we want to focus on the post-test survey and present its results.[18] The first questions were aimed at finding out whether users struggled to give quality and confidence ratings, and whether they thought that their ratings would be representative. Figure 21 shows part of the results from the questionnaire.

As can be seen, the majority of users reported that they had troubles choosing a quality level (44% agreed, 26% strongly agreed). This is an important indicator: it confirms our hypothesis that users struggle to settle for a score and therefore give (intrinsically) incorrect ratings as a result, since

[18] Note that the order of questions as presented in this section does not match the actual order in the questionnaire.

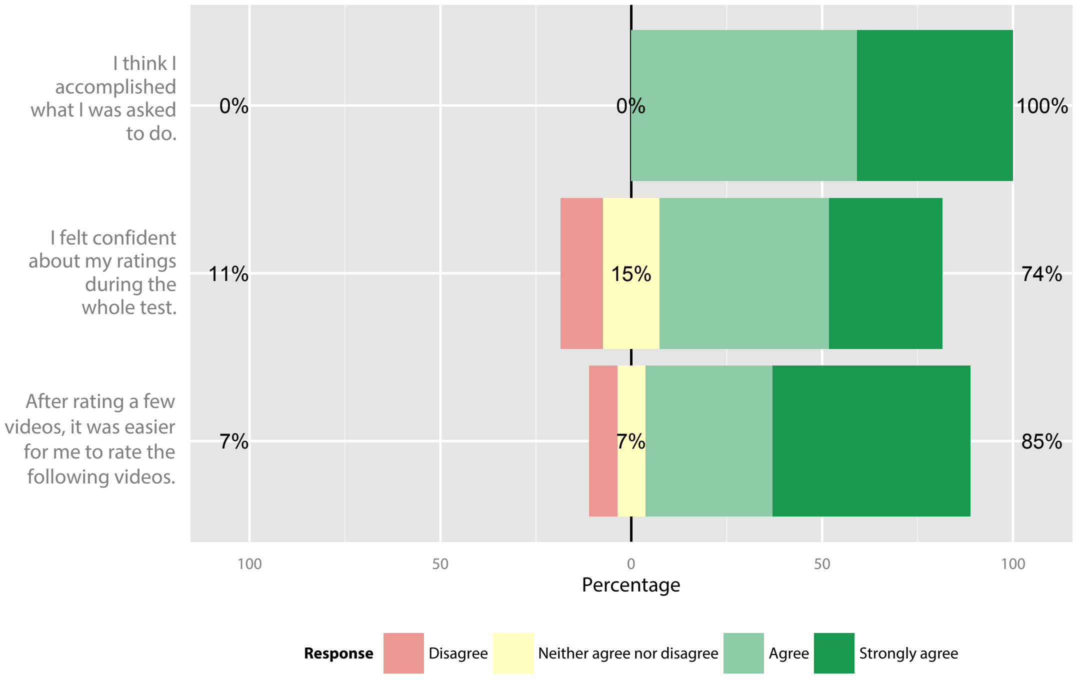


Figure 22: Responses to the post-test survey, part 2.

the methodology forces them to give a response. This also confirms hypothesis H1. It is therefore critical to give users a way to either skip these ratings altogether, or report the confidence of their rating. We expected a different picture for the self-assessment of the confidence. As a report of their own psychological state during rating, we anticipated that users could rate their confidence easily. The survey results however show that a slight majority in some cases did not know what confidence to choose. While the reasons for this are not entirely clear, an inappropriate scale could be the cause.

The responses from the third question of Figure 21 show that overall, users felt confident about their ratings being useful for the purpose of the study. Only 11% agreed that their ratings were not representative. During data analysis it would therefore be beneficial to check whether these users in fact gave scores that deviate from the average, or if their self-perception is merely skewed.

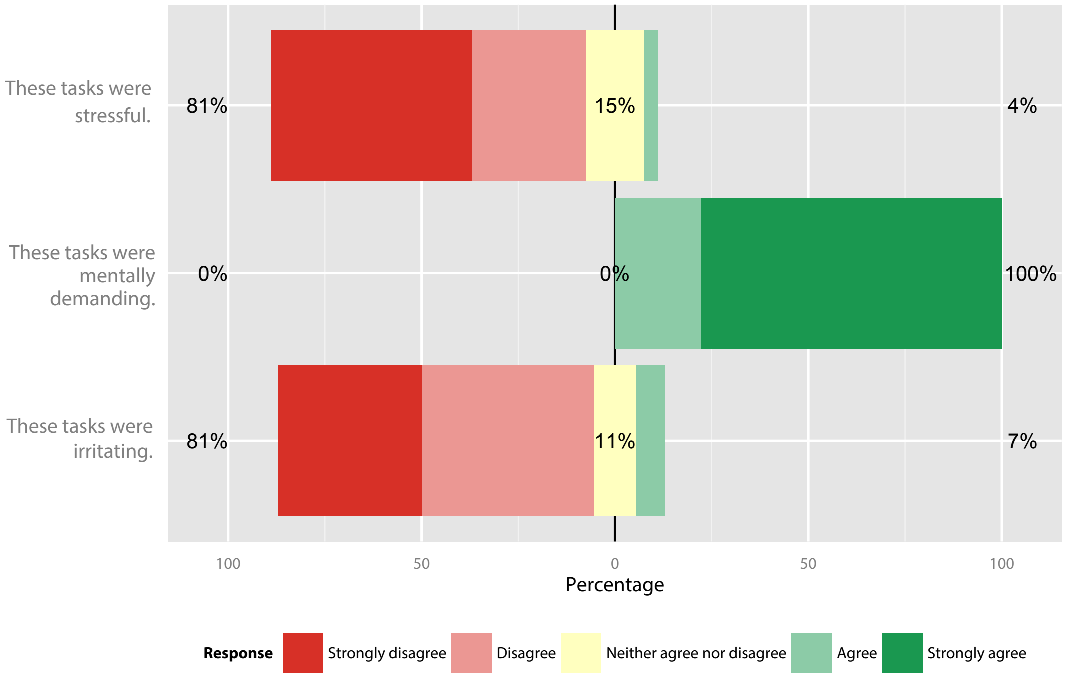


Figure 23: Responses to the post-test survey, part 3.

Another set of questions, presented in Figure 22, aimed at positive self-reflection: we asked participants if they thought they accomplished what they were asked to do, which all of them agreed or strongly agreed to. The second question shows that a majority of 74% *did* in fact feel confident throughout most of the test. While on a first glance this may seem to contradict the results from the previous question, in which participants noted that they sometimes did not feel confident during rating, they seem to take an overall positive stance on their performance. Put differently, a few insecure ratings do not seem to hamper the impression of their efforts.

Finally, we asked for the effect of (short-term) experience on their ability to give accurate ratings. 85% agreed or strongly agreed that after rating a few PVS, it became easier to rate the following. This underlines the necessity of properly preparing ("training") observers before recording their actual votes.

The last set of questions were taken from the NASA TLX questionnaire (Figure 23). It has a strong focus on evaluating the subjective cognitive load experienced by the users during the tests. We were specifically interested in whether users felt stressed or rushed, as this could explain inaccurate voting due to the test protocol forcing them to rate within a few seconds. This did not seem to be the case, with 81% agreeing that the tasks were not stressful. Likewise, when asked if they found the tasks irritating, 81% disagreed. However, all participants agreed that the tasks were mentally demanding (77% strongly agreed). This finding is also supported by interviews we conducted after QoE tests (not only the SVC EC experiment). For example, one observer said that "after a while the tests become really repetitive, and you want it to be over soon."

Concluding from this, we can see that there is no need to give users more time for their ratings than what the current protocols allow (e.g., less than 10 seconds as identified previously). The possible source of invalid votes therefore is not missing time or an overall stressful test setting, but the mental load forced on observers continuing over a timespan that they may interpret as too long. A test protocol that gives observers the chance to proceed at their own pace (like SAMVIQ) is preferable in this regard.

5.5.2 Self-Esteem And Confidence

We hypothesized that the self-esteem of a participant would have an influence on the confidence ratings they gave throughout the session (H6). For example, we assumed that an observer with a low self-esteem would, on average, give lower confidence ratings than observers with a high confidence. Knowing about this effect could make it easier to interpret the confidence scores taken. To test the hypothesis, we asked the observers to fill out the Rosenberg Self-Esteem Scale questionnaire before and after the actual test. As mentioned in Section 4.4.5.1, the RSES allows a maximum score of 40 (indicating high self esteem), with a typical population average of 30.

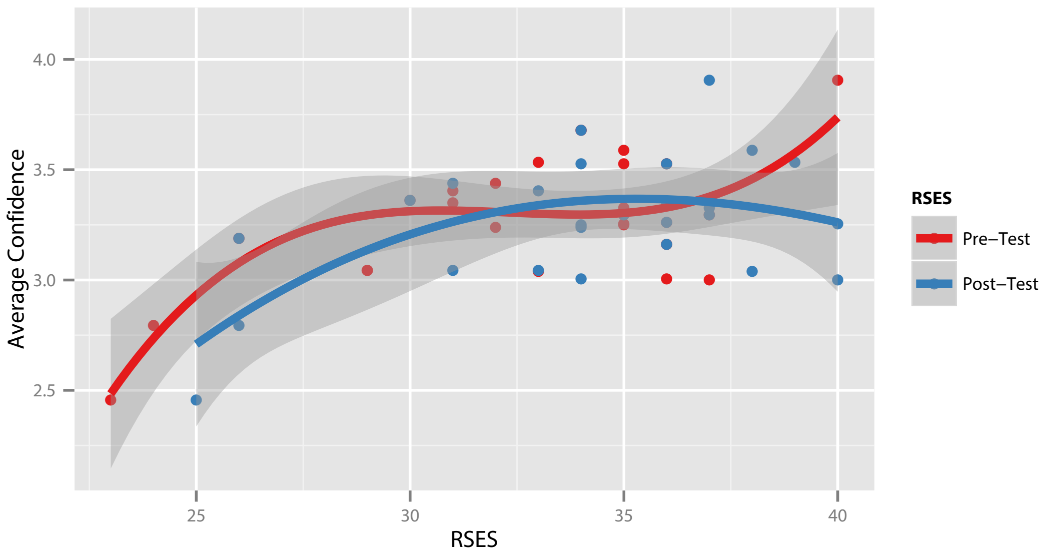


Figure 24: Pre- and post-test RSES results compared with average confidence and third-order polynomial fit.

In Figure 24 we can see the scores from the pre- and post-test RSES evaluation compared to each observers' average confidence. Each point indicates one observer ($n = 26$). To see whether there was a significant relationship, we calculated Pearson's correlation coefficient r. For the pre-test RSES, we measured a significant correlation of $r = 0.621$ ($p < 0.001$, with a standard error of 0.188 as confirmed by a bootstrap analysis with 1000 repetitions). The post-test RSES, while still significant, does not correlate as well with the average confidence ($r = 0.477, p = 0.014$, standard error 0.214), which indicates an influence of the quality test on the self-esteem, and diminishing the usefulness of the post-test RSES value. A third-order polynomial function predicts an observer's average confidence C depending on the pre-test RSES score r as:

$$C = 0.91r^3 - 0.30r^2 + 0.55r + 3.27$$

with $R^2 = 0.565$.

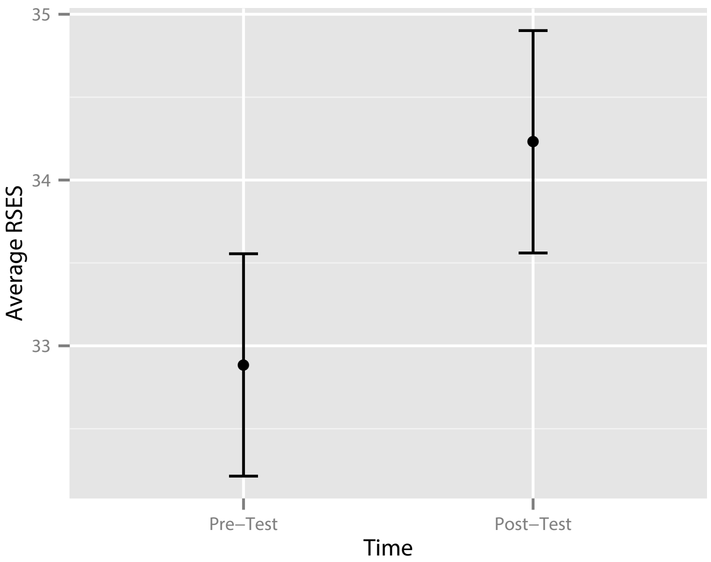


Figure 25: Pre- and post-test RSES results, averaged over all participants.

To prove that the RSES scores changed, we check the results from pre- and post-test: Figure 25 shows that there indeed is a significant difference: after the rating session, 50% of the subjects showed a higher self-esteem than before. 30% reported the same level of self-esteem. This could be explained by the fact that participants had the impression they succeeded in a task they thought was very mentally demanding, but not stressful.

We did not find an influence of gender on the RSES results: a t-test on the means of pre-test and post-test RSES scores per gender shows no significant difference (two-sided $p = 0.324$ and 0.792, respectively). Therefore, no further analysis is done in this regard.

Concluding from this section, we see that adequate psychological tests may provide a quick and easy way to determine voting behavior in advance, without the need to explicitly (and continuously) poll confidence levels during the quality test. A combination of a dedicated short psychological evaluation with a training set that polls for confidence scores could give much more insight into the data than just recording the quality scores alone.

6 Conclusions

The aim of this thesis was to describe the observer confidence in subjective QoE experiments—more specifically, video quality experiments. To motivate the research questions stated in this work, we first gave an overview of the history and current standards in QoE measurement in Chapter 2. Focus was laid on the ITU Recommendations BT.500-13 [6] and P.910 [5], which describe the most commonly used methodologies for the subjective testing of TV / multimedia quality. We highlighted the common protocols and rating scales, and identified some of the problems that these methodologies carry with them, namely that they force observers to give a rating in order to let them proceed to the next item.

In Chapter 3, the theoretical background of questionnaires was described, with regard to the way QoE experiments typically record quality scores. Based on feedback that we had received during previous QoE tests, we hypothesized that observers, when forced to rate quality, would sometimes struggle to decide for a rating, and as a result give invalid ratings that could lead to potential errors or misunderstandings during data evaluation.

We conducted seven experiment sessions, described in Chapter 4, in order to gather votes from observers rating their own confidence along the quality scores given to specific stimuli, using a four- or five-point Likert-type scale. The most important results of these experiment with regard to observer confidence were presented in Chapter 5. With the new dataset taken on video quality experiments we were able to compare our findings with those from Engelke et al. [16], who previously took confidence measurements on still images.

In the following, the most important findings will be summarized. We also give recommendations for

Confidence Rating Scale Similar to previous results shown in [16], the Likert-type rating scale consisting of four or five ordinal items (e.g. from "very unconfident" to "very confident" or "low" to "high") does not elicit uniformly distributed values, with peaks at the high confidence levels. While we do not expect there to be as many unconfident votes as there are confident ones, the expressiveness of the scale is limited. We therefore suggest to use a continuous Visual Analog Scale to capture confidence votes, recording the level of confidence from 0–100%.

Overall Confidence Across all our experiments, overall confidence was lower than shown in [16]. This difference stems from the fact that video material was shown, or that the wording of the scale between the experiments was different. We believe the former to be the cause. However, more experiments will have to be carried out in the domain of video QoE in order to prove this.

Influence of Quality on Confidence We hypothesized that observers would find it harder to rate medium quality content in comparison to extremely good or bad quality. This hypothesis could be proven, and we were able to model the mean confidence for a PVS as a function of the average distance from the middle of the scale, calculated over all observers.

Influence of Rating Scales A significant difference was found in the average confidence for the two Foreground-Background experiments whose experiment variables only differed in the scale being used (Impairment Scale rather than ACR). Observers therefore seem to find it easier to rate 1) the level of distortions they were able to see and 2) the acceptability as a binary response rather than being forced to translate their (internal) rating to a specific word on a constructed scale. Since the absolute results for both experiments aligned well, we therefore suggest to use the ACR scale with caution. Despite its broad usage and acceptance in the domain of QoE, a more carefully constructed scale could result in less observer variation.

Influence of Stimuli and Treatment No significant influence was found when correlating average confidence with the spatial information of source contents. Weak correlations existed between the temporal information and MCS. Analyzing average confidence for a set of HRCs showed no significant influence of a specific type of distortion—in fact, the resulting confidence scores seem to vary too much by observer. This means that a global analysis of confidence may not be expressive enough, and data should be evaluated on a by-observer basis.

Influence of Rating Times We measured the time needed to give a quality score and correlated it with the confidence the observers chose for that quality rating. As expected, average confidence drops with increasing time. A model for predicting the probability of a certain confidence level based on the elapsed time was constructed. We therefore suggest that quality ratings where observers took more than 10 seconds are to be removed from the dataset, and caution is taken for votes of over 6 seconds, where the probability of finding an unconfident response are higher than for a confident one.

Survey Analysis A post-test questionnaire for the SVC EC experiment revealed important insight into the observer behavior and their opinion of the general test procedures. A majority of users agreed that they sometimes did not know which quality rating to pick. This stresses the need for rating scales that intrinsically increase the confidence, or methodologies that allow users to skip ratings (or come back to them later, like SAMVIQ). Also, while the QoE tests were not perceived as stressful or irritating, observers felt that they were mentally demanding. This raises the question of whether conducting tests over longer time periods (perhaps even 15 minutes) procure meaningful votes towards the end. A solution for this issue is not easy to be found—it could, for example, consist in more pauses being allowed to be taken during an experiment.

Influence of Self-Esteem The Rosenberg Self-Esteem Scale as a measurement for perceived self-esteem is widely used in the field of psychology. We found a significant correlation between the observers' self-esteem (before the actual test session), and their average confidence ratings. Since the RSES is easy to administer, it gives helpful information about the observer behavior in advance of a test.

Based on the above findings, we recommend researchers to carefully evaluate the rating scale and overall methodology being used to assess QoE. The choice of rating scale has a measurable impact on the observers' overall self-perception during voting. An efficient and meaningful scale might not report quality on an absolute basis, but rather explain QoE in the dimensions of perceptible distortions and acceptability.

While taking confidence ratings throughout an entire test session might not be feasible, observers should be able to report their troubles at least during the training session before actual scores are collected. Such a training session could be extended over the typical small number of PVS and involve the experimenter discussing with participants before starting the main test procedure.

Computer-assisted testing makes it possible to automatically collect rating times for all PVS. We strongly recommend methodologies to include rating times in the data analysis and also suggest ratings to be removed if they fall outside predefined time spans.

Finally, the experiment results show that purely quantitative methods, like they have been used for decades, may not be perfectly suitable for new technologies or multimedia consumption contexts. The MOS as a single reporting number may be easy to consume, both by humans and statistical models, but we believe that the expressiveness of a five-point scale like ACR is limited. A successful evaluation of a system under test should therefore also include qualitative methods and user-focused testing procedures, if only to be able to know when votes are not necessarily reliable and—in a second instance—to know when they certainly are.

Bibliography

[1] Report bt.1082-1: Studies toward the unification of picture assessment methodology, 1990.

[2] Recommendation ITU-R BT.1788-0: Methodology for the subjective assessment of video quality in multimedia applications, 2007.

[3] Recommendation ITU-T E.800: Quality of telecommunication services: concepts, models, objectives and dependability planning – Terms and definitions related to the quality of telecommunication services, 2008.

[4] Recommendation ITU-T P.10/G.100: Vocabulary for performance and quality of service, Amendmend 2, 2008.

[5] Recommendation ITU-T P.910: Subjective video quality assessment methods for multimedia applications, 2008.

[6] Recommendation ITU-R BT.500-13: Methodology for the subjective assessment of the quality of television pictures, 2012.

[7] *Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003).* Patrick Le Callet, Sebastian Möller and Andrew Perkis, Lausanne, Switzerland, Version 1.1, June 3, 2012.

[8] Ralf Bender and Ulrich Grouven. Ordinal Logistic Regression in Medical Research. *Journal of the Royal College of Physicians of London*, 31(5):546–551, 1997.

[9] Norman M Bradburn, Seymour Sudman, and Brian Wansink. *Asking Questions: The Definitive Guide to Questionnaire Design–For Market Research, Political Polls, and Social and Health Questionnaires*. Wiley, 2004.

[10] S. Buchinger, S. Kriglstein, and H. Hlavacs. A Comprehensive View on User Studies: Survey and Open Issues for Mobile TV. *EuroITV – 7th European Conference on Interactive TV*, 2009.

[11] Shelley Buchinger, Ewald Hotop, Werner Robitza, Helmut Hlavacs, Francesca De Simone, Touradj Ebrahimi, Ulrich Engelke, Hans-Jürgen Zepernick, and Markus Fiedler. Towards a gesture-based video player interface. In *Third Euro-NF IA.7.5 Workshop on Socio-Economic Issues of Networks of the Future*, December 2010.

[12] Shelley Buchinger, Matej Nezveda, Werner Robitza, Patrik Hummelbrunner, and Helmut Hlavacs. Mobile TV coding. In *ConTEL 2009 – 10th International Conference on Telecommunications, 2009*, pages 465–465. IEEE, 2009.

[13] Shelley Buchinger, Werner Robitza, Patrik Hummelbrunner, Matej Nezveda, Martijn Sack, and Helmut Hlavacs. Slider or glove? Proposing an alternative quality rating methodology. In *Proceedings of VPQM'10*, Scottsdale, Arizona, 2010.

[14] Shelley Buchinger, Werner Robitza, Matej Nezveda, Patrik Hummelbrunner, and Helmut Hlavacs. Towards a comparable and reproducible subjective outdoor multimedia quality assessment. In *Third Euro-NF IA.7.5 Workshop on Socio-Economic Issues of Networks of the Future*, December 2010.

[15] Mathieu Desoubeaux, Gaetan Le Guelvouit, France Cesson-Sevigne, Yohann Pitrey, and William Puech. Subjective evaluation of Video DNA Watermarking under bitrate conservation constraints. In *QoEMCS workshop*, 2012.

[16] Ulrich Engelke, Anthony Maeder, and Hans-Jürgen Zepernick. Human observer confidence in image quality assessment. *Signal Processing: Image Communication*, 2012.

[17] Leena Eronen. User centered research for interactive television. In *Proceedings of the 2003 European Conference on Interactive Television: From Viewers to Actors (April 2-4, Brighton, UK)*, pages 5–12, 2003.

[18] Normand L Frigon. *Practical Guide to Experimental Design*. Wiley, 1997.

[19] David S Hands and SE Avons. Recency and Duration Neglect in Subjective Assessment of Television Picture Quality. *Applied Cognitive Psychology*, 15(6):639–657, 2001.

[20] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*. Wiley, 2013.

[21] T Hoßfeld, Raimund Schatz, and Sebastian Egger. SOS: The MOS is not enough! In *QoMEx 2011 – Third International Workshop on Quality of Multimedia Experience*, pages 131–136. IEEE, 2011.

[22] Quan Huynh-Thu, M. N Garcia, F. Speranza, P. Corriveau, and A. Raake. Study of Rating Scales for Subjective Quality Assessment of High-Definition Video. *IEEE Transactions on Broadcasting*, 57(1):1–14, 2011.

[23] S. Ishihara. *Tests For Colour-Blindness*. H.K. Lewis, London, 1957.

[24] Susan Jamieson et al. Likert scales: how to (ab) use them. *Medical education*, 38(12):1217–1218, 2004.

[25] Satu Jumisko-Pyykkö. *User-Centered Quality of Experience and Its Evaluation Methods for Mobile Television*. PhD thesis, Tampere University of Technology, 2011.

[26] Satu Jumisko-Pyykkö and Miska M Hannuksela. Does Context Matter in Quality Evaluation of Mobile Television? In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*, pages 63–72. ACM, 2008.

[27] Evangelos Karapanos, Jean-Bernard Martens, and Marc Hassenzahl. Accounting for diversity in subjective judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 639–648. ACM, 2009.

[28] Franc Kozamernik, Paola Sunna, Emmanuel Wyckens, and Dag Inge Pettersen. Subjective Quality of Internet Video Codecs — Phase 2 Evaluations Using SAMVIQ. Technical report, EBU, 2005.

[29] Lenny Lipton. The stereoscopic cinema: From film to digital projection. *SMPTE journal*, 110(9):586–593, 2001.

[30] Matej Nezveda, Shelley Buchinger, Werner Robitza, Ewald Hotop, Patrik Hummelbrunner, and Helmut Hlavacs. Test persons for subjective video quality testing: Experts or non-experts? In *QoEMCS workshop*, Tampere, Finland, 2010.

[31] Abraham Naftali Oppenheim. *Questionnaire design, interviewing and attitude measurement*. Continuum International Publishing Group, 2000.

[32] Oxford Dictionaries. Oxford University Press. "questionnaire".

[33] Margaret Pinson, Lucjan Janowski, Romuald Pépion, Quan Huynh-Thu, Christian Schmidmer, Phillip Corriveau, Audrey Younkin, Patrick Le Callet, Marcus Barkowsky, and William Ingram. The influence of subjects and environment on audiovisual subjective tests: An international study. 2011.

[34] Margaret Pinson and Stephen Wolf. Video Quality Measurement Techniques. Technical report, National Telecommunications and Information Administration, 2002.

[35] Yohann Pitrey, Marcus Barkowsky, Romuald Pépion, Patrick Le Callet, Helmut Hlavacs, et al. Influence of the source content and encoding configuration on the perceived quality for scalable video coding. *SPIE Human Vision in Elect. Imaging*, 2012.

[36] Yohann Pitrey, Werner Robitza, and Helmut Hlavacs. Instance Selection Techniques for Subjective Quality of Experience Evaluation. In *EuroITV – 10th European Conference on Interactive TV*, 2012.

[37] D.H. Pritchard. US Color Television Fundamentals: A Review. *SMPTE Journal*, 86(11):819–828, 1977.

[38] Ulf-Dietrich Reips and Frederik Funke. Interval-level measurement with visual analogue scales in internet-based research: Vas generator. *Behavior Research Methods*, 40(3):699–704, 2008.

[39] Einar Risvik, Jean A McEwan, Janet S Colwill, Regina Rogers, and David H Lyon. Projective mapping: A tool for sensory analysis and consumer research. *Food quality and preference*, 5(4):263–269, 1994.

[40] Einar Risvik, Jean A McEwan, and Marit Rødbotten. Evaluation of sensory profiling and projective mapping data. *Food quality and preference*, 8(1):63–71, 1997.

[41] Werner Robitza, Shelley Buchinger, and Helmut Hlavacs. Impact of Reduced Quality Encoding on Object Identification in Stereoscopic Video. In *EuroITV - 9th European Conference on Interactive TV*, Lisbon, Portugal, June 2011.

[42] Werner Robitza, Shelley Buchinger, Patrik Hummelbrunner, and Helmut Hlavacs. Acceptance of mobile TV channel switching delays. In *Quality of Multimedia Experience (QoMEX), 2010 Second International Workshop on*, pages 236–241, 2010.

[43] Werner Robitza, Yohann Pitrey, and Helmut Hlavacs. The NappingPlayer — Projective Mapping Experiments on Android Tablets. In *PQS – Fourth International Workshop on Perceptual Quality of Systems*, Vienna, Austria, 2013.

[44] Werner Robitza, Yohann Pitrey, Matej Nezveda, Shelley Buchinger, and Helmut Hlavacs. Made for Mobile: A Video Database Designed for Mobile Television. In *VPQM - Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2012.

[45] Morris Rosenberg. The Measurement of Self-Esteem. *Society and the Adolescent Self Image*, 297, 1965.

[46] Jens Schumacher and Karin Feurstein. Living labs–the user as co-creator. In *ICE 2007 Proceedings: 13th International Conference on Concurrent Enterprising*. The Free Press, 2007.

[47] Heiko Schwarz, Detlev Marpe, and Thomas Wiegand. Overview of the scalable video coding extension of the H. 264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(9):1103–1120, 2007.

[48] Dominik Strohmeier. *Open Profiling of Quality: A Mixed Methods Research Approach for Audiovisual Quality Evaluations*. PhD thesis, Technische Universität Ilmenau, 2011.

[49] Felix B Tan and M Gordon Hunter. The Repertory Grid Technique: A Method for the Study of Cognition in Information Systems. *MIS Quarterly*, pages 39–57, 2002.

[50] Evelyne F Vallieres and Robert J Vallerand. Traduction et validation canadienne-française de l'échelle de l'estime de soi de Rosenberg. *International Journal of Psychology*, 25(2):305–316, 1990.

[51] Video Quality Experts Group. Final Report of VQEG's Multimedia Phase I Validation Test. Technical report, 2008.

[52] Holly Willis. *New Digital Cinema: Reinventing the Moving Image*, volume 25. Wallflower Press, 2005.

[53] Stefan Winkler. On the properties of subjective ratings in video quality experiments. In *QoMEx 2009 – International Workshop on Quality of Multimedia Experience*, pages 139–144. IEEE, 2009.

[54] Robert K Yin. *Case study research: Design and methods*, volume 5. Sage, 2009.

List of Abbreviations

ACR	Absolute Category Rating
AVC	Advanced Video Coding (→ H.264 / MPEG-4 Part 10)
DCR	Degradation Category Rating
DMOS	Difference → MOS
EBU	European Broadcasting Union
GOP	Group of Pictures
HRC	Hypothetical Reference Circuit
IRCCyN	Institut de Recherche en Communications et Cybernétique de Nantes
IS	Impairment Scale
ITU	International Telecommunication Union
MCS	Mean Confidence Score
MMOS	Middle Mean Opinion Score
MOS	Mean Opinion Score
PVS	Processed Video Sequence
QoE	Quality of Experience
QoS	Quality of Service
RSES	Rosenberg Self-Esteem Scale
SRC	Source
SVC	Scalable Video Coding (→ H.264 / MPEG-4 Part 10)
VAS	Visual Analog Scale
VGA	Visual Graphics Array(640 × 480 pixels)

Abstract

How much can we trust our data? How much confidence can we put into the ratings of our observers? Even more so: How confident were our observers when they were rating?

No computer-generated estimate can substitute subjective tests with human observers when it comes to evaluating the Quality of Experience (QoE) of today's multimedia services. Automated Quality of Service (QoS) measurements that take into account factors such as the bitrate, packet loss, or signal to noise ratio may give an estimation of the resulting quality for the end user, but QoS-based methods have been proven inefficient at predicting the experienced quality, only offering a rough estimate. In turn, QoE experiments are conducted in order to give ground truth data for creating models that predict QoE on the base of QoS data. To generate accurate models, one needs to know whether the acquired data itself is accurate.

There exist various documents by the ITU, such as ITU-T BT.500-13 or ITU-T Rec. P.910 which describe the way subjective multimedia quality experiments have to be conducted. They also include procedures on data analysis, which specify how experiment data has to be reported, and test persons have to be removed from the pool when their behavior is deviating from the others.

The ratings acquired from viewers during experiment sessions are often simply put in a bowl. This is what we call the "Mean Opinion Score" (MOS)—the average score all observers assigned to a stimulus. The MOS does not take into account inter-personal differences or the fact that observers might not have been too sure on what they were even rating. Often, MOS are presented along with their 95% confidence intervals (CI). The CI is a good sign of agreement between observers, but only in the sense of how certain one can be that the found MOS conforms to the "true" MOS. To dive deeper into understanding the causes for (dis)agreement between observers, a new rating methodology that focuses on their confidence is evaluated over the course of seven different multimedia quality experiment sessions, conducted at the University of Vienna and the Institut de Recherche en Communications et Cybernétique de Nantes in France.

Focusing on the confidence of observers, it becomes obvious that the estimated quality may not only depend on the actual stimulus, but even outside factors such as the test situation or the personality. Even the scale used for assigning quality values could have an influence on how confident observers might feel during a session. Also, with new emerging multimedia services such as 3D vision, one cannot assume previous experience of the observers with the technology, which might lower the confidence they put in their votes.

In this thesis, we address multiple hypotheses, such as whether confidence can be measured effectively during experiments, what personal factors influence the voting behavior, and how the confidence of observers influences their quality votes. In our experiments, we also take into account personality traits and hidden measurements such as the reaction time of observers. We show that rating behavior differs from person to person. We propose new reporting and data analysis methods and formulate recommendations for the conduction of QoE experiments that will allow much deeper insight into the acquired data.

Abstract (Deutsch)

Können wir unseren Daten vertrauen? Wie viel Vertrauen kann man in die Bewertungen von VersuchsteilnehmerInnen setzen? Anders gefragt: Wie sicher waren sich die TeilnehmerInnen bei der Bewertung selbst?

Wenn es um die Evaluierung von Multimedia-Qualität geht, können computer-generierte Schätzungen kaum subjektive Tests mit menschlichen TeilnehmerInnen ersetzen. Automatisierte Quality of Service (QoS) Messungen können zwar Faktoren wie Bitrate, Paketverlust, oder Signal to Noise Ratio mit einbeziehen und eine Schätzung über die resultierende Qualität für den User liefern, jedoch werden diese Methoden als ineffizient angesehen, da sie ineffizient und ungenau die Quality of Experience (QoE) voraussagen. Daher werden QoE-Experimente durchgeführt, um Ground-Truth-Daten für Modelle zu liefern, welche wiederum QoE auf Basis von QoS-Daten berechnen können. Um jedoch genaue Modelle zu generieren, müssen wir wissen, wie genau die Daten sind, die von ExperimentteilnehmerInnen geliefert wurden.

Dokumente der ITU – wie etwa ITU-T BT.500-13 oder ITU-T Rec. P.910 – beschreiben, wie subjektive Experimente zur Messung von Multimedia-Qualität durchgeführt werden sollen. Sie inkludieren Prozeduren für die Datenanalyse und -auswertung. Außerdem wird beschrieben, welche Datensätze entfernt werden müssen, sollten Testpersonen in ihren Ergebnissen zu stark von den anderen TeilnehmerInnen abweichen.

Die Bewertungen, die TeilnehmerInnen in Experimenten abgeben, werden typischerweise gemittelt – dies ist der "Mean Opinion Score", also due Durchschnittsbewertung für einen Stimulus, über alle Testpersonen gesehen. Dieser MOS berücksichtigt jedoch nicht die Unterschiede zwischen den TeilnehmerInnen, oder etwa die Tatsache, dass sich ein(e) TeilnehmerIn bei der Bewertung nicht sicher gewesen ist und möglicherweise ungültige Daten abgegeben hat. MOS werden häufig mit ihrem 95% Konfidenzintervall präsentiert. Das Konfidenzintervall ist zwar ein gutes Zeichen für die Streuung der Bewertungen, aber zeigt nur, wie sehr der gefundene MOS sich dem tatsächlichen MOS nähert. Um die Gründe für Übereinstimmung zwischen Bewertungen verschiedener TeilnehmerInnen genauer zu erforschen, benötigen wir jedoch neue Bewertungsmethoden. In sieben Experimentreihen, durchgeführt an der Universität Wien sowie am Institut de Recherche en Communications et Cybernétique de Nantes in Frankreich, erforschen wir eine solche Methode, die die Selbstsicherheit der TeilnehmerInnen zum Hauptaugenmerk hat.

Es stellt sich heraus, dass die bewertete Qualität nicht nur von dem Stimulus an sich, sondern auch von äußeren Faktoren, wie etwa der Testsituation oder der Persönlichkeit der Testperson abhängt. Auch die Bewertungsskala kann einen Einfluss auf die Selbstsicherheit haben. Gerade bei neuen Technologien wie etwa 3D-Fernsehen und -kino können WissenschafterInnen nicht zwangsläufig vorherige Ergebnisse heranziehen, um Qualitätsbewertungen vorzunehmen. Hier ist es wichtig, auch abzuschätzen, inwieweit neue Technologien TeilnehmerInnen verunsichern und damit ihre Bewertungen verfälschen.

In dieser Arbeit soll mehreren Fragen nachgegangen werden, unter anderem, ob die Selbstsicherheit von ExperimentteilnehmerInnen effektiv gemessen werden kann, welche persönlichen Faktoren das Bewertungsverhalten beeinflussen, und welche Auswirkungen die Sicherheit wiederum auf die Qualitätsbewertungen hat. In unseren Experimenten berücksichtigen wir auch Persönlichkeitsmerkmale und versteckte Messungen, wie etwa die Bewertungszeit der TeilnehmerInnen. Wir zeigen auf, wie stark sich das individuelle Bewertungsverhalten zwischen Personen unterscheiden kann, und schlagen neue Analysemethoden für QoE-Experimente vor. Diese erlauben bessere Einblicke in Experimentdaten und sollen WissenschafterInnen helfen, QoE besser vorauszusagen.

Acknowledgements

First and foremost, I would like to thank my supervisors Dr. Shelley Buchinger and Univ.-Prof. Dipl.-Ing. Dr. Helmut Hlavacs, who offered me the chance to start working as a researcher at the Faculty of Computer Science in 2009. Without their support and feedback I would never have been able to follow my goals.

Thank you also to Dr. Yohann Pitrey, who was a great mentor to talk to in the last year before writing the thesis. I want to thank Ulrich Engelke for his ideas and interesting phone calls on the subject of this work. Credits go to all my colleagues working in the projects CACMTV and QUASSUMM: Ronny Aderkast, Christoph Czepa, Patrik Hummelbrunner, Benjamin Kitzinger, Matej Nezveda (who also wrote the foreground-background separation software and the MATLAB interface for conducting most of the tests), Martijn Sack (who built our own slider and made use of my electronic gloves) and David Selig. Thank you also to Romuald Pépion for conducting the tests at the IRCCyN in France for us.

Finally, my wife Lisa for always supporting me whenever I needed motivation—thank you.

Curriculum Vitae

Name:	Werner Robitza
E-Mail:	werner.robitza@gmail.com
Nationality:	Austria

Education

2011–2014	Master's program for Media Informatics, University of Vienna
2007–2011	Bachelor's program for Informatics, University of Vienna
2006	Matriculation at BG/BRG Bruck an der Leitha
1998–2006	Secondary school at BG/BRG Bruck an der Leitha

Academics and Research Experience

Oct. 2012 – Jan. 2013	Tutor for "Multimedia Retrieval", Multimedia Information Systems Group, University of Vienna
Oct. 2012 – Dec. 2012	Researcher, Entertainment Computing Research Group, project "QUASSUMM", University of Vienna
Mar. 2009 – Aug. 2012	Researcher, Entertainment Computing Research Group, project "Content Aware Coding for Mobile TV", University of Vienna
Oct. 2010 – Oct. 2011	Staff tutor, Multimedia Information Systems Group, University of Vienna
Mar. 2010 – Jun. 2010	Tutor for "Grundlagen wissenschaftlichen Arbeitens", University of Vienna
Oct. 2009 – Mar. 2010	Tutor for "Software Architectures and Web Engineering", University of Vienna

Employment

2013–2014	Software engineer and project manager at salesXp GmbH, Vienna
2008–2012	IT and network support for 3s Unternehmensberatung GmbH, Vienna
2006–2007	Civil service at the Austrian Red Cross, Hainburg an der Donau

A full list of publications can be found at:

`http://homepage.univie.ac.at/werner.robitza/publications`