Crowdsourcing Quality of Experience Experiments

Sebastian Egger-Lampl[1(✉)], Judith Redi[2], Tobias Hoßfeld[3], Matthias Hirth[4], Sebastian Möller[5], Babak Naderi[5], Christian Keimel[6], and Dietmar Saupe[7]

[1] Austrian Institute of Technology, Vienna, Austria
sebastian.egger-lampl@ait.ac.at
[2] Delft University of Technology, Delft, Netherlands
[3] University of Duisburg-Essen, Duisburg, Germany
[4] University of Würzburg, Würzburg, Germany
[5] TU Berlin, Berlin, Germany
[6] Technische Universität München, Munich, Germany
[7] University of Konstanz, Konstanz, Germany

1 Introduction

Understanding and measuring quality of multimedia and communication services and underlying communication networks from an end-user perspective (Quality of Experience, QoE) has attracted increased attention over the course of the last decade. For a better understanding of the QoE concept and its progression towards its actual conception and execution, it is helpful to make a brief review of the recent history of communications quality assessment.

In the early 1990s, the notion of Quality of Service (QoS) attracted considerable attention in telecommunications, nurtured by articles, for example, Parasuraman [76], in which the authors described their conceptual model of service quality and in which the ultimative instance for the service quality judgment was the respective customer. This user or customer centricity is also reflected in the ITU-T definition of QoS[1], which underlines the subjective roots of the service quality concept despite being oriented rather towards the view of a telecommunications provider or manufacturer:

> *Quality of Service is the totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service.* [46]

However, contrary to this original definition, most QoS-related work actually focused on the investigation of purely technical, objectively measurable network and service performance factors such as delay, jitter, bitrate, packet loss etc., thereby effectively reducing quality to a purely technology-centric perspective [7,85].

[1] ITU-T standards and work are frequently referred to in this introduction, as a number of initial and ongoing work in QoE is carried out within ITU-T study group 12.

D. Archambault et al. (Eds.): Evaluation in the Crowd, LNCS 10264, pp. 154–190, 2017.
DOI: 10.1007/978-3-319-66435-4_7

Due to this deviance from its subjective focus the concept of QoS got less attractive to domains such as audio and video research, where historically subjective quality assessment played a major role in comparing, for example, codec performance. A countermovement gained momentum which took up the notion of *Quality of Experience*, which was initially introduced in the context of broadcast technologies and television systems by Kubey and Csikszentmihalyi [62][2]. The notion of QoE was rapidly adopted not only in the context of mobile communications [99] but also in the domains of audio and video quality assessment [71,79,91,107]. However, each service type (voice, video, data services, etc.) tended to develop its own QoE community with its own research tradition and flavor. In addition, it has to be noted that some domains do not even use the notion of QoE but rather use the terms "subjective quality" or "user-perceived quality" although utilizing the conceptual model that goes back to QoE [3,5,27].

This has resulted in a number of parallel attempts to define QoE, as outlined by Reichl [85], accompanied by an equally large number of QoE frameworks and taxonomies (see Laghari et al. for a comprehensive overview [63]). However, today the definition by ITU-T Rec. P.10 (Amendment 2, 2008) is still the most widely used formulation of QoE, defining the concept as:

> *QoE is the overall acceptability of an application or service, as perceived subjectively by the end user.* [45]
> Note 1: includes the complete end-to-end system effects.
> Note 2: may be influenced by user expectations and context.

During discussions at the Dagstuhl Seminar 09192 in May 2009 it was pointed out that among others the notion of "acceptability" in the above definition is problematic as the concept of acceptability demands a certain (usage) context of the service [94] to yield reproducible results across different assessments of QoE or acceptability respectively. In addition, a new definition of acceptability was proposed as follows:

> *Acceptability is the outcome of a decision [yes/no] which is partially based on the Quality of Experience.* [70]

In an attempt to overcome this patchwork of definitions and additions, the COST Action IC 1003 has published a QoE definition whitepaper [7]. Version 1.2 of this whitepaper defines:

> *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.*

[2] It can not be figured out with 100% certainty who introduced the notion of QoE into the domain of multimedia quality assessment, however the work by Kubey and Csikszentmihalyi is one of the earliest ones that used the notion in the same understanding as it is still used nowadays [62].

Thus, it advances the ITU-T definition by going beyond merely binary acceptability and by emphasizing the importance of both pragmatic (utility) and hedonic (enjoyment) aspects of quality judgment formation[3].

In this respect, the above definition captures the essence of QoE by highlighting some of its main characteristics: subjectivity, user-centricity, and multidimensionality. Particularly concerning the latter aspect, most frameworks and definitions found in the literature highlight the fact that QoE is determined by a number of hard and soft *influence factors*: (a) user factors, (attributable either to the user him/herself), (b) system factors and (c) context factors (see Fig. 1 and [7]). This means that whether a user judges the quality of, for example, a mobile video service as good (or even excellent) not only depends on the user her- or himself (expectations, personal background, etc.), the performance of the technical system (including traditional network QoS as well as client and server performance),[4] but to a large extent also on the context (task, location, urgency, etc.) of the experience. The resulting level of complexity and broadness turns reliable and exact QoE assessment into a challenging problem.

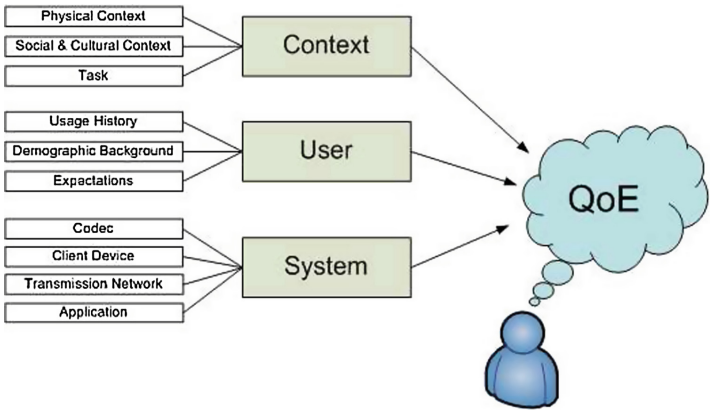

Fig. 1. QoE influence factors belonging to context, human user, and the technical system itself [95].

The very first and core step towards implementing this concept is the measurement of QoE.

In this respect, the QoE research and industrial community has typically favored a quantitative approach versus, for example, a more qualitative approach

[3] The definitions of the terms used as well as further details can be found in the QoE definition whitepaper [7] itself.

[4] Note that the technical system generally comprises of a chain of components (sender, transmission network elements, receiver) that connect the service provider with the end-user. All these elements can influence service quality (and thus QoE) on different layers, predominantly in terms of network- and application-level QoS.

taken towards User Experience (UX) in the Human Computer Interaction (HCI) domain. Psychometric techniques have been adapted to measure perceptions and preferences with respect to QoE, in what has been called QoE subjective testing. Subjective testing is, to date, the most common way to quantify users' QoE. Nevertheless, it is typically performed in highly controlled laboratory environments, to avoid bias and noise in the measurement due to undesired influence factors (see Fig. 1). This, of course, poses a limit to the quantity of test participants that can be involved, as well as on their diversity. For this reason, lately the community has started looking at crowdsourcing as an alternative approach to conduct large scale QoE experiments.

This chapter provides an overview of recent advances for QoE research in a crowdsourcing setting. To this end, it has firstly provided a general background to the QoE concept in the introduction above. The remainder of this chapter provides first an overview of QoE experiment types and commonly used scaling methodologies, followed by a discussion of specific QoE issues and experimental challenges for three different service categories: voice communication, audio-visual multimedia and web applications. Furthermore, specific challenges for transferring laboratory based experiments to the crowdsourcing context are reviewed for these three service categories. Finally, lessons learned are summarized in order to provide guidelines for setting up crowdsourced QoE tests to the interested reader. In the appendix to this chapter a novel approach towards using paired comparison in the crowdsourcing environment and related technologies for subsequent reconstruction of absolute category ratings is discussed.

2 Subjective QoE Experiments

The main goal of subjective QoE experiments is to sort stimuli (e.g., speech segments, audio tracks, images, videos,...) according to their perceived properties or attributes on a given scale, as defined by Engeldrum [19]. The scaling can be obtained by directly asking participants to (numerically) quantify QoE (in the so-called "direct" tests), or by deriving indices related to quality on the basis of other, intermediate measures ("indirect" tests). Such measures could for example be thresholds of perception (in classical psychophysics), physiological responses (such as skin conductance, EEG or EMG), or performance indicators (such as task success for an interaction task). All such tests can in principle be carried out both in a laboratory as well as in a crowd environment. However, different types of tests may set different requirements to the influence factors.

There are a number of criteria according to which experiments addressing the QoE of a system or service can be differentiated. A common classification is one used for standard psychophysical experiments, distinguishing amongst the following:

- Perceptual modality: Viewing tests, listening tests, viewing and listening tests, etc.
- Degree of activity: Passive (e.g. listening-only or viewing-only tests), active (e.g. speaking tests), interactive (e.g. conversation tests with different degree of interactivity)

- Presentation method: presentation of constant stimuli, with or without explicit reference (e.g. Absolute Category Rating tests, Paired Comparison tests, Comparison Category Ratings, Degradation Category Ratings) vs. adjustment of stimuli by the test participants
- Scaling method: Quantitative scaling of stimuli on a nominal, ordinal, interval or ratio scale

Whereas the first two items above do differ for different types of services and stimuli (discussed in Sects. 2.1, 2.2, and 2.3) the latter two items (and their variations) are rather common for all types of QoE experiments as they deal with the mapping of subjective experiences on certain (quantitative) descriptors. In the following a number of scaling methodologies that quantify subjective experiences are discussed.

The Paired Comparison (PC) method as described by David, Thurstone and Engeldrum is a classic psychometric technique that allows to precisely measure distances among stimuli in terms of Just Noticeable Differences (JNDs) [13,19, 100]. The experimental procedure consists of asking participants to compare each stimulus with all other stimuli in the set. As a result, even small differences between the stimuli can be detected, which makes the method particularly useful when stimuli close together in quality are to be sorted. On the other hand, the judgment effort grows as the square of the number of stimuli, hence this number must be limited.

Direct scaling techniques overcome this limitation by presenting the participant with a numerical (or categorical) scale on which each stimulus is evaluated (effort grows only linear with the number of stimuli). Participants have to quantify the QoE of the stimulus on such a scale; this judgment can depend on the comparison of the stimulus with a reference (Double Stimulus Methodology) or not (Single Stimulus Methodology). The Double Stimulus Impairment Scaling (DSIS) methodology as described in ITU-R BT-500 is often chosen for the assessment of audio or visual impairments [48]. DSIS judgments are expressed on an interval scale (typically, a five-point Absolute Category Rating - ACR - scale or a Degradation Category Rating scale [48]), as a (conscious) comparison of each impaired stimulus with its undistorted version. Being a double stimulus method, DSIS requires a moderate effort per judgment, but still allows the assessment of large datasets. A possible drawback of the method may be the categorical scale used for the assessment: the boundaries among categories (for example, “good” and “fair”) are blurred and depend on the participant; this may result in low inter-participant agreement as indicated by Engeldrum and Keelan [19,54]. Redi et al. have shown that to date the ACR scale is however the most widely used one in image and video subjective testing, also in a Single Stimulus settings (i.e., without an explicit reference to be presented to the participant) [82]. Both DSIS and Single Stimulus scaling can be performed also with numerical scales, both discrete or continuous as described in ITU-R BT-500 and Huynh-Thu et al. [42,48]. In all cases, the results of the tests are reported in terms of average score per stimulus (Mean Opinion Scores), expressed in the scale used for the experiment. These scores reflect human preference, though do not have a precise

psychophysical meaning. Indeed, the obtained scores may vary with the definition of the scale as shown by Engeldrum [19], as well as with the quality range spanned by the stimuli as shown by de Ridder [89]. This suggests that comparing results of different experiments may be problematic, possibly inducing inconsistencies when merging these data in a single, larger dataset.

The methods briefly described above are commonly used across the media domains considered in this chapter: audio, image/video and web. On the other hand, for each of these domains, the dominant influencing factors may change; as such, specific methodological choices and recommendations to conduct subjective QoE experiments were developed. In the following subsection we cover this specificity, separately per domain.

2.1 Experiments Addressing Speech and Audio QoE

2.1.1 Experiments Addressing Speech QoE

Speech quality has been an object of investigation for more than a century, and the corresponding methodologies assessing speech QoE are rather well established. Common types of experiments include listening-only tests (Absolute Category Rating, Paired Comparison, Comparison Category Rating, Degradation Category Rating), third-party listening tests (listening to a conversation between two other persons), speaking-only tests, as well as conversation tests. More recently, diagnostic tests targeting individual listening-quality dimensions, conversational dimensions, as well as technical sources of quality degradations, have been a focus of research. The most common methods are described in the P.800 series of Recommendations issued by the International telecommunication Union, ITU-T, in particular ITU-T Rec. P.800 for listening-only and conversation tests, ITU-T Rec. P.805 for conversation tests with differing degree of interactivity, ITU-T Rec. P.806 for multi-dimensional assessment of listening-only quality, or ITU-T Rec. P.830 regarding quality assessment of coded speech. All of these methods can be considered as good practice for speech related QoE assessment and are frequently used for the different speech application fields.

These recommendations also specify a number of influence factors. User influence factors that have to be controlled are the participants' hearing ability, their language skill, and potentially their expertise with the domain of speech quality in the case that diagnostic listening for identifying technical sources of degradations is of interest. Whereas these characteristics can easily be controlled in a laboratory setting, they are more difficult to verify in a crowd setting, where participants may have the possibility to cheat in the case that self-reported abilities are used.

System influence factors are the ones most frequently under study. They include the source speech material (commonly collected from a variety of speakers, using different types of text material), the technical characteristics of the signal processing chain as well as the presentation device used by the listening participant. In the case of speaking or conversational tests, this deletes the source material from the list of influence factors which can be controlled for, however this puts additional requirements for the speaking and listening devices. Context

factors which can be expected to carry an impact on the results are the listening environment (especially the background noise and reverberation), as well as the test task given to the participants. The latter has shown to significantly impact quality judgments in the case of conversational test situations.

2.1.2 Experiments Addressing Audio QoE

Audio quality is in principle addressed similarly to speech quality. However, as the level of quality is commonly expected to be much higher, the test methodologies are commonly focusing on a more sensitive distinction between different processing chains of reproduction devices, and the requirements for the test equipment and listening situation are commonly higher. Test paradigms which are followed in audio quality assessment are, for example, double-blind triple-stimulus tests with hidden anchor, where test participants first have to distinguish between a degraded stimulus and a hidden reference, and then have to rate the perceived degradation on a category scale; or the multiple-stimulus test with hidden reference and anchor (MUSHRA), where the quality of multiple stimuli presented in parallel to the test participants has to be rated in relationship to each other, and is anchored by the use of a scale with absolute labels. With respect to factors influencing audio QoE, the same influence factors do apply as mentioned for speech above.

2.2 Experiments Addressing Image and Video QoE

Research on subjective image and video quality has, so far, mostly focused on determining user sensitivity to visual impairments and quantifying the annoyance generated by their visible presence. Multiple psychometric methodologies have been developed for this purpose, and adapted for the measurement of image and video quality in standardized conditions [44,47,48,50,55].

Methodologies such as DSIS, Paired Comparison and Single Stimulus evaluation with an ACR scale defined in ITU-R BT-500 and ITU-T P.910 are typically used to conduct both image and video subjective QoE assessments [47,48]. In addition, the Quality Ruler (QR) method deserves a mention, as an middle-ground alternative between the direct scaling methodologies (DSIS, Single Stimulus) and Paired Comparison. The QR method was first described by Keelan in [54], and subsequently adopted as an international ISO standard for psychometric experiments for image quality estimation [55] and video quality estimation in ITU-R BT-500 [48]. The core idea of the QR method is to provide the participant with a set of reference images, anchored along a calibrated quality scale, to compare a test image with. The task of the participant is to find the reference image closest in quality to the test image by visual matching. Reference images (1) depict a single scene and vary in only one perceptual attribute (i.e., blur, blockiness, color saturation); (2) are closely spaced in quality, but altogether span a wide range of quality. They are presented in a way that easily allows detection of the quality difference between them, and their close spacing in quality should allow the participant to score with higher confidence, decreasing the risk of inversions and range effects. In practice, participants perform several comparisons of reference-test stimuli to complete a single assessment, until

they find the reference stimulus that matches the quality of the test one. The advantage of this procedure is that, as long as the reference stimuli are kept the same, subjective scores obtained from a QR experiment always refer to the ruler scale, and not to the quality range spanned by the test stimuli. This minimizes range effects. Furthermore, Redi et al. have shown that the visual matching procedure reduces inter-participant variability [82]. This method has been successfully implemented for images, and recently Freitas et al. have proposed to use it for video quality assessment with promising results [22].

As mentioned earlier for audio and speech quality assessment, recommendations and standards enlist a series of influencing factors that impact on subjective quality assessment of images and videos. Among user influencing factors, we can distinguish between physiological (e.g. visual acuity, color blindness, stereo blindness) and psychological factors (preference for image material, personality and culture). To limit the influence of physiological factors on the test outcomes, ITU-R BT-500 advises to screen participants for (corrected to) normal visual acuity (e.g. by means of the Snellen or Landolt charts), and for normal color vision (e.g. via the Ishihara test) [48]. Limiting the influence of psychological factors is more complex; questionnaires investigating individual characteristics (e.g. personality) can be administered pre- or post-test and their outcomes used as co-variates in the rating analysis; a large number of observers and the careful selection of diverse image material can also help averaging out individual differences. Due to the visual nature of the stimuli, their physical representation towards the human participants is crucial. Hence, representation characteristics of the display device are the most important system influence factor. Examples of such characteristics are the achievable contrast ratio, the representable color space as well as the dynamic range of the display. Depending on the independent variable varied, one of these characteristics might be of utmost importance, for example, dynamic range for experiments addressing HDR representations of images or videos. With respect to context influencing factors, visibility conditions (monitor resolution and calibration, distance to screen, lighting) need to be controlled for and made uniform (most recommendations prescribe specific settings in this respect). The ambience (or context) in which the experiment is carried out also influences evaluations: Jumisko-Pyykkö and Hannukselausers found users to be more tolerant towards visual artifacts when evaluating them in realistic viewing conditions (laboratories with a living room appearance, bus, cafes) than in traditional laboratories [53].

2.3 Experiments Addressing Web QoE

Web-QoE, defined as "Quality of Experience of interactive services that are based on the HTTP protocol and accessed via a browser" by Hossfeld et al. [39], focuses on the optimization of web services by understanding the end-users' perception of overall system performance. The critical issue in this context are perceived waiting durations which occur after requesting a web-site until it has been fully loaded in the visible browser window.

Therefore, it is important to instrument waiting durations as the key metric for assessing Quality of Experience for web-based services. Furthermore, it is important to go beyond single page requests to a series of consecutive page views in order to accommodate for the interactive nature of web browsing activities. Especially interactivity and the related tasks which users want to accomplish are major QoE influencing factors beyond network-related performance parameters and have to be accounted for. The main characteristics of such subjective web browsing QoE tests as described in ITU-T P.1501 are [49] to simulate realistic web browsing where users are browsing and interacting with webpages in order to acquire certain information. The procedure they go through within this methodology has to ensure that users get into a browsing mode rather than a pure page loading mode. From a system factor perspective it must be ensured that participants are exposed to a certain QoS level over a period of time rather than for one event, in order to grasp several request-response cycles for the subjective evaluation. Additionally, it has to be ensured that the manipulated parameters (e.g., delay, packet loss, downlink bandwidth) can be set to the desired values and that these settings can be verified by *a posteriori* analysis (e.g., traffic traces). Accommodating for all these characteristics and at the same time ensuring that waiting times are properly instrumented is typically addressed by two approaches: (a) utilizing network emulators [16,94] that shape traffic such that the loading behavior of normal webpages is manipulated or (b) developing special webpages where waiting times are directly instrumented via, for example, Javascript [12,17,20,112].

With respect to the context of use, Strohmeier et al. showed that the task assigned (i.e. context of use) to the test participants has a considerable impact on QoE [12]. This is important to keep in mind when using certain tasks to stimulate the interaction between the webpage and the participant for each test condition. In addition to the assigned task, the webpage must be interactive and has to provide sufficient content such that the participant can browse through it over several conditions, without getting bored. As for the other services discussed above, human influence factors have to be considered for Web QoE as well. Varela et al. have shown that despite the ubiquitous usage of web sites across the globe, there are nevertheless differences with respect to archetypical web site arrangement and structuring as well as web site design and visual appeal [103]. Additionally, Sackl et al. showed that user expectations with respect to downlink performance and web page loading times have to be considered as well [92].

3 Transferring QoE Lab Experiments to the Crowd

The previous section has shown that QoE testing in laboratory environments is an established approach known for producing valid and reliable results. The major disadvantage of such laboratory-based experiments is the fact that they not only require expensive facilities and testing expertise but also incur significant expenses and relatively long campaign setup and turnaround times (typically in the order of weeks). Therefore, laboratory experiments are not suitable

for testing a large number of technical conditions in proof of concept tests or for comparing a large number of prototype implementations during the development phase.

Crowdsourcing, with its outreach to thousands of users concurrently, represents a very appealing option for subjective QoE experiments. Nevertheless, crowdsourcing of QoE experiments also faces certain challenges. In order to properly transfer QoE experiments from the laboratory to the crowd testing environment, dedicated solutions and great care has to be put into the test design. Within this section we discuss specific challenges that are connected with QoE experiments in crowdsourcing such as experiment duration and human, system and context influence factors.

3.1 Influence Factors Particularly Relevant for QoE Tests in Crowdsourcing

3.1.1 Test Duration and Design

Independent on the type of media/signal on which the QoE assessment is carried out, QoE testing is typically performed in a within-subjects fashion. In the simplest case, the experimenter wants to evaluate the impact of a set of F system factors s_f, with $f = 1, .., F$, on a specific type of media. To do so, (1) a set of K diverse, unimpaired media contents $O_k, k = 1, ..., K$ is selected and (2) a set of levels L_f is determined per each factor s_f to be applied, in isolation or combination with other levels of other factors, to the K selected contents. This results in a number N of impaired stimuli, which can be described as $O_k(s_1(i_1), s_2(i_2), ...s_F(i_F))$, where $f = 1, ..., F$ and $i_f = 1, ..., L_F$ in case of full factorial design. A pool of M users is then asked to evaluate the quality of *all* impaired stimuli (within-subjects design), within one or multiple sessions. This setup, also denoted as complete block design, allows to control for individual differences in quality perception (by modeling users as a random factor); nevertheless, it results in long experimental sessions, especially when the number of conditions N to be tested is large.

In crowdsourcing, long test sessions should be avoided. As pointed out by Hossfeld et al., short durations will favor engagement of the workers with the tasks, thereby favoring reliable executions and commitment [33,38]. Hour-long crowd-based tests would most likely result in poorly reliable executions and, therefore, poor results. For this reason, crowd-based evaluations see the transformation of complete block designs into incomplete ones. That is, the set of stimuli to be evaluated is divided in subsets, each evaluated by separate groups of workers in different campaigns. While considerably shortening the task duration, this practice has implications for the validity and reliability of the evaluations, namely:

- Redi et al. showed that it increases the risk of context effects, since the quality range spanned by each block of stimuli can hardly be kept constant across blocks [81]

- In the case of interaction between worker and influencing factors (i.e. different impairment perceptions depending on the worker, which quite often occurs), the non-systematic structure of the test will make the results difficult to analyze and interpret
- It further complicates the analysis, given that the incomplete block design becomes unbalanced, and that the same worker may participate to different campaigns (which is often the case, but can be controlled for on certain platforms).

3.1.2 Crowd Diversity and Expectations

With respect to user diversity, crowdsourcing platforms have a different reach-out to the population compared to typical laboratory tests. As crowdsourcing platforms are online platforms, only computer-literate persons will participate in the tasks, and due to the prevailing financial motivator the group will also show certain income characteristics, which certainly will differ from test participants recruited for laboratory tests (e.g. at academic institutions, through marketing companies, through newspapers, etc.).

Furthermore, visual and hearing characteristics, which are important for a number of QoE experiments, are usually rather widespread and can be only controlled to a certain extent in crowdsourcing settings. Due to the shorter crowdsourcing task length compared to the laboratory (see above), a higher number of different crowdworkers is required to collect the same number of ratings. Along with this increased user diversity, the diversity in user ratings increases as well. Another indirect factor of QoE perception on the user level can be the users' expectations: those used to lower quality (e.g. low video resolution) will rate differently than those typically consuming higher quality (e.g. high video resolution). Sackl et al. proved that the expectation level may be closely related to the usage experience of services and to the country of the crowdworkers [92]. In line with these findings, Hirth et al. showed that users from different regions may have different expectations about the provided content quality [32]. As a countermeasure to crowd diversity and expectations, training tasks or jobs can be integrated in crowdsourcing campaigns. In the training job, anchor stimuli (see Sect. 5.1) are presented to the worker and rated by them. Proper identification of such anchor degradations should be clearly visible in the respective worker's ratings. Anchor stimuli act as a standard reference, and their aim is to introduce the entire quality range to the workers with more consistent results in the end. Gardlo, Egger and Hossfeld have shown that proper training sessions help workers to use the entire range of the scale [24]. Another approach of temporarily expiring training certificates as a prerequisite qualification for crowdworkers has been proposed by Polzehl et al. [78]. The authors showed that training certificates valid for 40 min were able to clearly improve the correlation of crowdsourced and laboratory test results. As this subsection has only discussed QoE-related user factors the interested reader is pointed towards Chap. 3 for a more in depth discussion about demographic factors and challenges with respect to crowdsourcing.

3.1.3 Equipment

In contrast to laboratory experiments where presentation hardware can be closely monitored and controlled, workers in crowdsourcing tests typically use their own devices, in their current environments (e.g. wherever they are in case of mobile crowdsourcing). These devices may differ in terms of hardware (e.g. display, brightness sound output device, connected headphone, volume settings), software (e.g. OS, installed codecs) and connectivity (e.g. the bandwidth or delay of the Internet connection may vary). Furthermore workers may use their devices in different ways (e.g. monaural/binaural listening, concurrent use of other devices and/or applications) which can not be controlled and barely monitored. Therefore, it is important to either detect the device type and the device usage, or to ask users about their used hardware and settings. Another limitation (rather than an influence factor) with respect to the equipment are crowdsourced QoE tests of specific technologies, which require dedicated equipment. This might not be feasible, due the lack of diffusion of such equipment, and for the difficulty in emulating it. For example, immersive media technologies such as augmented reality and/or virtual reality, mulsemedia etc. The same holds true for contingent equipment that is often used to assist QoE experiments: eyetrackers and physiological sensors. In the case of eye-tracking, recent developments by Lebreton et al. have made it possible to track eye-movements of the worker while doing the task, although there is room for improvement [64]. Measurements through physiological sensors can not be achieved, for now.

3.1.4 Context

Because of the remoteness of the workers and the heterogeneity of the used soft- and hardware, it is necessary to monitor the users' environment in order to identify additional influence factors on the QoE assessment (see Sect. 1 for QoE influencing factors). Due to the unknown context in which the QoE assessment is performed by the workers in QoE crowdsourcing tests, these influence factors are not known beforehand, but are hidden, yet still influence the users' QoE ratings. In general, we have three options to cope with the unknown context and the resulting hidden influence factors. We can either monitor the appropriate context parameters, adapt the context or try to prevent the undesirable context itself in our test design. The environment in which the workers evaluate the stimuli in QoE crowdsourcing tests may impact the overall QoE and thus the application should be able to detect such factors. For visual stimuli, the general viewing conditions represented by the background illumination or the screen resolution can be influencing factors.

One option to adapt the conditions of the workers' environment is to provide them with simple test patterns that allow them to either calibrate their devices or enable the quantification of the deviation of a device's stimuli representation from the desired target. For visual stimuli, Gardlo et al. showed that basic test patterns similar to the test patterns used for calibration of monitor contrast and illumination in a professional environment can be utilized to quantify the users' viewing conditions, for example by asking how many gray steps on a grayscale

step-wedge are visible [25]. Moreover, such patterns can also be used to instruct workers how to calibrate their display.

Similarly, we can prevent an undesirable context from the technical perspective, for example for video QoE assessment, by pre-loading videos with included distortions in the remote browser, so that additional distortions introduced by the transmission do not affect the playback [15,40]. Hence, influence of the users' context with respect to bandwidth is no longer an issue. But even by doing so, the resulting initial delays may also be too long and influence the user rating. In both cases, it is evident that monitoring on system or application level is required. As a possible solution, download speed and latency may also be measured before the actual test, and then only users with suitable connection speed and latency are selected.

3.2 Speech and Audio QoE

As speech and audio QoE tests span a wide range from pure listen-only tests to interactive conversational tests (see Sect. 2.1), the challenges for conducting such tests through crowdsourcing are manifold as well. Therefore, we exemplify only challenges and solutions that are applicable across all these test types. As sound reproduction is key for speech and audio tests, respective human and system characteristics have to be carefully considered. With respect to hearing abilities of the crowdworkers, and when the workers hearing level is not an independent factor under the study design, hearing levels of all participating workers should be examined in screening tasks. Candidates with normal hearing levels should then be qualified for participating in the main campaign. Alternatively, workers with different hearing levels should equally be distributed throughout different campaigns and test conditions. Besides the human hearing characteristics, Cooke et al. showed that system characteristics can be assessed during such screening tasks (e.g. type of hardware, OS, mon- or bi-aural output devices) [10]. With respect to context factors, either question-based or measurement-based context assessment is feasible. Measurement-based approaches as introduced by Naderi et al. on mobile devices allow for identification of worker mobility (or the worker being stationary through motion sensors, location data) or if he is in a silent or noisy environment (through the device microphone) [73].

Initial suggestions on how to design crowdsourcing tasks for subjective speech and audio QoE have been provided by Naderi and Pozehl [74,78], as well as guidelines of resulting data analysis by Ribeiro [88]. Furthermore, within ITU-T study group 12 a work item has been started towards a recommendation on subjective methods for assessing audio quality in crowdsourcing environments.

In the domain of audio and speech quality, crowdsourcing is to date used for subjective speech quality ratings (e.g. listening-only-tests [74,78]), naturalness [88], intelligibility test [68,108] and preference tests of speech synthesis systems, followed by data collection studies (e.g. to explore factors from wireless networks that impact mobile voice quality [75], evaluating voice-over-IP services [93] and Skype call quality assessments [111]).

3.3 Image and Video QoE

Image and video quality assessment is done for a range of different application areas: from the visual quality evaluation of picture and video coding technologies and processing algorithms to the influence of network delays and packet loss in case of video quality. The QoE of image and video is usually determined in a well-defined testing environment with subjective methodologies, as described in standards [47,48,55].

The first challenges we face result from the differences of crowdsourcing compared to the structure and procedures of subjective video quality assessment in an laboratory environment. Crowdsourcing tasks are typically small tasks that can be done by the workers both fast and easily and while image and video quality assessment is usually a comparably easy task, laboratory-based assessment sessions can last up to 30 min as in ITU-R BT-500 and ISO 20462 [48,55]. Hence, it is not possible to just run a test designed for a laboratory environment without modifications; it rather needs to be partitioned into several crowdsourcing campaigns, for example, its basic test cells (BTCs) or only a small subset of BTCs compared to a laboratory-based assessment will be included in each crowdsourcing campaign and its underlying tasks. The necessary breaking up of the structure of the laboratory-based assessment makes the adherence to design rules aiming at avoiding contextual effects, therefore more challenging. Moreover, compared to the approach taken in laboratory-based assessment, Keimel et al. and Redi et al. showed that workers will usually only assess a subset of all image or video sequences under test [56,84].

In contrast to images, video is more challenging from a resource perspective (e.g. bandwidth requirements or download volume for long video sequences) in crowdsourced quality assessments. Obviously, we are neither able to control the setup of the testing environment itself (e.g. room illumination), nor the used equipment (e.g. displays). This, however, also implies that evaluations requiring explicitly a controlled environment, for example, for determining the thresholds of just noticeable differences of stimuli, are not suitable for crowd-based evaluation. Also research questions utilizing new technologies for the visual stimuli are not yet widely deployed in consumer equipment. For example, Hanhart et al. have claimed that questions related to high dynamic range (HDR) displays, can not easily be answered using crowdsourcing as respective displays are not widely available to crowdworkers [28,29]. Even though image downloads and video streaming is nowadays a generally used service, crowdsourced image video quality assessment faces some additionally challenges compared to the laboratory environment. Firstly, we need to consider that in general the worker's web-browser and plug-ins cannot be assumed to support the original encoding format of images and videos, especially lossless compression. On the one hand, this limits the possibility to asses new coding technologies or other processing algorithms which are neither supported nor can be emulated using generic web technologies. On the other hand, double stimulus methodologies requiring an undistorted version of the stimuli under test (e.g. DSIS, as defined in ITU-R BT.500 [48]) can also not be used. Even though this last point can be circumvented by re-encoded

images and videos for the delivery with common lossy coding techniques supported by common web-browsers, the test case then differs even stronger from the laboratory setup, as also the artefacts introduced by this additional compression will be implicitly assessed. Secondly, in case of video QoE the bitrate needed for smooth video playback can be substantial and this can limit the pool of potential workers. Buffering the video can help in lifting this limitation, but buffering will extend the time needed per test case, limiting in turn the number of test cases that can be assessed per crowdsourcing task and thus further deviating from the laboratory-based setup.

Despite these differences between crowdsourced and laboratory-based image and video quality assessment, crowdsourced image and video quality assessment has been used so far successfully as a replacement for laboratory-based QoE assessments for a number of different research questions: Image recognizability and aesthetic appeal [81,83,84], selfie portrait images perception in a recruitment context [69], privacy in HDR images and video [59,60,86], QoE of video coding in general [57,58], audio-visual QoE of Internet-based applications in [8,9,109], and influence of stalling events and initial delays [34,36] on the QoE of video streaming applications. In addition, a general discussion using crowdsourcing for image and video QoE is provided by Hossfeld et al. [33,84].

3.4 Web QoE

In the context of interactive services accessed via the browser, waiting times are the key influence factor for the user's perception of performance. Thus, proper manipulation of these waiting times is of utmost importance in evaluation studies. For crowdsourced tests this is a particular challenge. Due to the limited control of the network connection (traffic shaping, as shown by Schatz and Egger [94] or delay of certain page elements in the downlink path as shown by Shaikh et al. [97]), such a manipulation can only be achieved through the development of special web sites that are able to instrument certain page loading behavior and respective waiting times until the content is displayed. A further complicating factor for this aspect is the realistic appeal of the resulting web sites as deemed important in ITU-T P.1501, which necessitates a certain content depth of the created web sites. This results in a large set of content to be acquired for, for example, a news look-a-like web site [49]. Furthermore, comparable to other services such as video and speech, test duration, testing equipment of the worker and crowd diversity do pose certain challenges for conducting Web QoE studies in a crowdsourcing environment. Due to the limited time for the test duration per user and incomplete block designs a large number of workers have to be chosen. Differences in testing equipment can not be *a priori* defined by the nature of crowdsourcing, however logging of numerous equipment factors important for Web QoE (e.g. screen size and resolution, terminal category etc.) is possible. This enables the researcher to consider equipment factors in the data analysis as an additional dimension. Contrary to video and speech services where reproduction fidelity of the end user device is of high importance for resulting media fidelity, reproduction fidelity of web sites is not bound to media fidelity as long

as correct rendering can be ensured. Hence, visual characteristics of the display such as color accuracy or brightness of the display are not as important. Diversity of crowd and workers is of course a complicating factor but can be controlled either *a priori* by proper crowd selection or *a posteriori* by respective reliability analysis approaches (see Sect. 5.2). Also context factors do exert certain influences on Web QoE. Strohmeier et al. have shown that the task context while web browsing does impact users' QoE ratings [12]. On the other hand, results from Guse et al. have shown that physical context (laboratory vs. metropolitan transport) did not lead to significant differences in the QoE ratings [26].

Despite these challenges certain successful work on Web QoE in a crowdsourcing context has been presented. In order to overcome the web site content challenge, the work from Egger and Schatz [17], ETSI[5] [20], and Zinner et al. [112] present open source solutions that make it easy to create web sites with instrumentable loading times and realistic appeal [17,20,112]. To date, no crowdsourced results with these solutions have been published but will appear shortly. With respect to crowd and worker diversity the work in Varela et al. has studied the impact of design and visual appeal on web QoE for geographically differing societies and showed that there are different degrees of influence and different preferences of design as well [103]. A further study from Varela et al. showed that changes in visual appeal do impact perceived performance of web sites despite technically identical loading times [104].

4 Crowdsourcing Frameworks for QoE Testing

Crowdsourcing has been widely used by researchers in domains other than QoE so far and consequently numerous different tools (e.g., Turkit [67]), have already been developed to ease the application of Crowdsourcing for their purposes. While some tools, like Turkit, focus on general problems, for example, providing control flow for consecutive crowdsourcing tasks, other software tools or frameworks are designed for a specific use case.

Using crowdsourcing to conduct QoE assessments seems to be a promising way to quickly collect a large number of test results in real world usage settings. However, it imposes new and different challenges compared to similar tests in laboratory environments. The first challenge is to find an appropriate pool of workers for the test and a crowd provider providing a flexible enough interface to run the experimental tasks. The second major challenge is the delivery of the test to the workers. It is often necessary to redesign the test to a web-based version which allows the access for the globally distributed workers and – in the best case – does not require the workers to install any software on their device. During this process a significant software development effort is needed that can be reduced significantly by reusing existing frameworks.

Web-based crowdsourcing frameworks for multimedia quality assessment represent a conceptual approach with programming tools to develop subjective tests that can be executed in a web browser. In particular, such frameworks allow

[5] European Telecommunications Standards Institute.

multimedia content to be displayed in a browser for workers to evaluate the quality using web forms. The test logic may be implemented at the client-side (e.g. Javascript) or at the server-side (e.g., PHP). Such frameworks enable the execution of the tests utilizing typical crowd-provider platforms. The basic functionality of a framework includes (a) the creation of the test (by supporting common testing methodologies like ACR, DCR, PC), (b) the execution of the test (by supporting training, task design, task order, screening), and (c) the storage and access to the result data. In the following we give an overview of existing crowdsourcing frameworks that have been specifically developed for QoE tests[6] and their available features.

This overview is structured along specific criteria such as the test design, the applied test methodology, the type of media to evaluate, and the hardware and software environment. In the remainder of this section, we focus on frameworks especially for crowdsourced QoE studies. Hoßfeld et al.provided a survey of widely used frameworks for this purpose in [37]. We summarize the considered frameworks therein and additionally consider Crowdee[7], which has a major focus on quality testing.

Quadrant of Euphoria

Initially proposed by Chen et al. in [9] and extended by Wu et al. in [109], Quadrant of Euphoria mainly focuses on the QoE evaluation of audio, visual, and audio-visual stimuli. It allows for a pairwise comparison of two different stimuli in an interactive web-interface, where the worker can judge which of the two stimuli has a higher QoE. Reliability assessments are based on the actual user ratings under the assumption that the preferences of users are a transitive relation, expressed by the Transitivity Satisfaction Rate.

crowdMOS

The crowdMOS framework for subjective user studies was proposed by Ribeiro et al. [88] and is an open-source project that initially focused on subjective audio testing using the ACR and MUSHRA audio quality assessment methodologies. Ribeiro later extended the crowdMOS framework to image quality assessments [87] with ACR for video from ITU-T P.910 [47]. For assessing the reliability of users, the sample correlation coefficient between the average user rating of a worker and the global average rating is used.

QualityCrowd

QualityCrowd is an open-source project by Keimel et al. that provides a multitude of different options for the test design [57]. In this framework, a test can consist of any number of questions and can contain videos, sounds or images or any combination. Moreover, it allows the use of different testing methodologies

[6] As each crowdsourcing test is somewhat unique, it is very difficult to find a framework that can be used directly without any modification. However, using an existing framework as a starting base and modifying it to fit the requirements of the test design needed is a highly valuable alternative.

[7] http://crowdee.de last accessed 14 Jun 2017.

(e.g. single stimulus or double stimulus), and different scales (e.g., discrete or continuous quality or impairment scales). In its latest iteration (QualityCrowd2[8]) a simple scripting language has been introduced that allows for the creation of test campaigns with high flexibility. This is not only achieved by enabling the combination of different stimuli and testing methodologies, but also by the possibility to specify training sessions and/or the introduction of control questions for the identification of reliable user ratings in order to ensure high data quality.

WESP

Rainer et al.describe an open source[9] Web-based subjective evaluation platform (WESP), which was initially developed for subjective quality assessments of sensory experience but can also be used for general-purpose QoE assessments [80]. WESP provides a management and presentation layer for configuring the task design and for the presentation of the actual user study, respectively. The management layer allows the configuration of each component (e.g. pre-questionnaire, voting mechanism, rating scale, and control questions), independently and thus provides enough flexibility for a wide range of different methodologies (e.g., single stimulus, double stimulus, pair comparison or continuous quality evaluation). Additionally, any new methodology can be implemented through the management layer. The presentation layer presents the content (e.g. video using HTML5 or Flash), to the workers and is based on standard HTML elements. In particular, it allows the collection of explicit and implicit user input: the former is entered by the user via explicit user input elements (e.g. voting using a slider for a given rating scale), compared to the latter describing implicit input represented by data from the browser window (e.g. window focus or duration of the test).

BeaqleJS

The BeaqleJS framework is developed by Kraft and Zölzer and focuses on subjective audio studies [61]. It is written in Javascript and PHP, and HTML5 is used to playback the audio clips[10]. Several audio formats are supported, including an uncompressed WAV PCM format. The framework allows the implementation of different testing methodologies via code extensions, with two evaluation methodologies already implemented: the ABX methodology and MUSHRA. Currently, there is no support for reliability detection and evaluation results are emailed to the organizer of an evaluation in a text file.

in-momento crowdsourcing

Gardlo et al. [25] introduced the *in-momento* crowdsourcing framework, combining careful user-interface design together with the best known practices for QoE crowdsourcing tests from Hossfeld et al. [33]. Instead of *a posteriori* data analysis and subsequent removal of unreliable data, this framework aims at live

[8] https://github.com/ldvpublic/QualityCrowd2 last accessed 14 Jun 2017.
[9] http://selab.itec.aau.at/ last accessed 14 Jun 2017.
[10] https://github.com/HSU-ANT/beaqlejs last accessed 14 Jun 2017.

or *in-momento* evaluation of the user's behavior: as the user proceeds with the assessment, the reliability of the user is continuously updated and a reliability profile is built which is used for screening. Users are able to quit the assessment task at any point unlike in other frameworks. The aim is to avoid forcing a user to continue with the test even though they are bored or have lost interest, as these two issues are closely related to unreliable behavior. Since the reliability profile is known at each stage of the assessment, it is possible to offer reliable users additional tasks for an increased reward.

Crowdee

Crowdee is a mobile crowdsourcing micro-task platform which is developed and actively supported by a research group of the Quality and Usability Laboratory, Technische Universität Berlin. Besides the fundamental functionalities provided by crowdsourcing platforms, Crowdee brings worker mobility to crowdsourcing user studies. Workers use a mobile application to find and perform micro-tasks available in the platform. As a result they are able to perform QoE tasks wherever and whenever they want, which facilitates conducting QoE studies in field settings [73]. With respect to modalities, crowdee enables image, audio or video content for testing. As a further option for media playout the researcher can force multimedia content to be preloaded before the start of a task to avoid influence of network distortions. Scales and questions can be selected among free text, single or multiple choice, sliders, taking a picture, or recording audio and video.

In addition, the platform supports dynamic worker profiles. Profile values can automatically or manually be assigned on response submission or approval time. Profile keys can be used to specify necessary qualifications and profile values for granting permission to perform a job. Polzehl et al. used these temporal profile entries in order to specify training qualification validity periods and were able to significantly improve the quality of responses in a crowdsourcing speech quality assessment task [78].

Discussion of Frameworks: Pros and Cons

Table 1 compares the different crowdsourcing frameworks for QoE assessment. The frameworks differ mainly in the testing methodology they natively support and which kind of multimedia content can be used.

There are some platforms (CrowdMOS and BeaqleJS), which focus on audio quality assessment and implement specific methodologies for subjective evaluation of audio quality like MUSHRA. Other platforms like WESP or Crowdee allow full flexibility by providing programming interfaces or making the source code publicly available. Concerning the task design and the possibility to add additional reliability questions beyond the rating task (e.g. content questions to check reliability of users), this feature is only provided by CrowdMOS or QualityCrowd2. Others implement basic screening mechanisms instead. Quadrant of Euphoria from Chen et al. uses the transitivity index [9]; CrowdMOS uses 95% confidence intervals which does however not allow to check reliability of workers properly as described by Hossfeld et al. [33]; the *in-momento* approach from

Table 1. Comparison of crowdsourcing frameworks for QoE assessment.

Frame-work Ref.	Euphoria [9]	Crowd MOS [88]	Quality Crowd2 [57]	WESP [80]	BeaqleJS [61]	*in-momento* [25]	Crowdee [73]
Multimedia types							
Image	x	x	x	x		x	x
Video	x		x	x		x	x
Audio	x	x	x	x	x		x
Testing methodology and scale							
Single		x	x	x		x	x
Double	x		x	x	x		x
Mushra		x			x		x
Cont. Scale			x	x		x	x
Questionnaire and task design							
Add. questions		x	x	x			x
Custom template		x	x				x
Random order	x	x		x		x	x
Screening	x	x				x	x

Gardlo et al. computes a reliability score of a worker during the test used for reliability screening [25]. Crowdee differs as it is a crowdsourcing platform provider and can therefore provide a (historic) reliability profile of its workers.

The frameworks considered here are designed for different purposes: either to support concrete methodologies like MUSHRA or paired comparison, or to evaluate quality of certain multimedia types. However, as each crowdsourcing test is somewhat unique, Hossfeld et al. have shown that it is very difficult to find a framework that can be used directly without any modification [37]. Still, the provided overview may help the researcher to select an existing framework as starting base which may then be modified to the purposes of the own test.

5 Lessons Learned

5.1 Scale and Anchoring

Whereas multiple criteria should be adopted to select the methodology most appropriate to investigate a specific problem, direct scaling via, for example, Absolute Category Rating (ACR), has been extensively used in laboratory-based QoE testing [21] due to its ease in implementation and straightforwardness in the interpretation of results. As mentioned earlier, ACR entails users to visualize the stimulus once (Single Stimulus setup), and quantify its quality/level of impairment on a discrete scale, along which qualitative labels (adjectives) are reported (bad-poor-fair-good-excellent for the quality scale). Workers are required to indicate which of these five adjectives better expresses the quality level of the stimulus.

Although direct scaling fits perfectly many of the requirements of crowdtesting (ease of implementation, task simplicity and fast completion), it is important that the task designer takes into account one of its major drawbacks: the risk of returning scores suffering from context effects as shown by Corriveau et al. and de Ridder [11,89]. Context effects derive from the cognitive bias that leads subjects to use the entirety of a scoring scale (in case of ACR, until 'bad'), to express the quality range that is visualized in the stimulus set. So, having a stimulus set having true quality values covering a range $[0, A]$, and a second set of stimuli covering the range $[A/2, A]$, it is quite likely that the worst stimulus of the second range will still obtain a Mean Opinion Score (MOS) close to 'bad' (although in reality is not as bad as other stimuli in the first set, with a true quality value $<A/2$). Pitrey et al. showed the solution to this issue is re-alignment [77].

In order to overcome these issues with direct scaling the work by Wu and Chen proposes to use comparison rating procedures instead [9,109]. An elaborate discussion of this approach can be find in the appendix of this chapter.

A possible solution to context effects derived from the fragmentation of QoE evaluations in crowdsourcing was proposed by Hossfeld et al. and Redi et al. [33,84]. The authors suggest to introduce a small number of stimuli in each evaluation campaign, kept equal for all sub-tasks and spanning a wide range of quality. These stimuli, named "anchors", have the purpose of limiting context effects by fixing the extreme values of aesthetic appeal to be seen in each subtask. For this reason, at least one of the anchors should present extremely bad quality, possibly lower than that of the entire stimulus set, and at least one should have excellent quality (as known, for example, from a small pilot study prior to the main campaign). Redi et al. showed that the use of anchors was effectively limiting context effects [84]. The authors had a set of 200 images to be rated with respect to aesthetic quality in a crowdsourcing environment. They divided the set in 13 subsets, to be evaluated in as many campaigns. Then, they added to each campaign five images whose quality values corresponded to the minimum, maximum and 25th, 75th and 50th percentile of the distribution of the quality values of the entire image set (as known from a previous laboratory experiment, see Fig. 2). In analyzing the data, the authors performed a re-alignment of the image MOS across campaigns, only to conclude it was unnecessary and their ordering would not change significantly after realignment, thereby proving the effectiveness of the anchors.

In terms of language and scale design crowdsourcing workers are quite heterogeneous regarding their native language and their cultural background. Therefore, they often receive instructions and scale descriptors different from their native language. As the language cannot be relied on in terms of scale description, different scale designs can influence the scale usage and the resulting mean opinion scores. Therefore, the unambiguous design of rating scales is essential for acquiring proper results from crowdsourcing campaigns.

Based on these assumptions a comparison of different scale types and designs Gardlo, Egger and Hossfeld have revealed that an ACR 5 scale with non-clickable anchor points and traffic-light semaphore design as depicted in Fig. 3 yields reliable results and is most efficient in terms of the relative number of outliers [24].

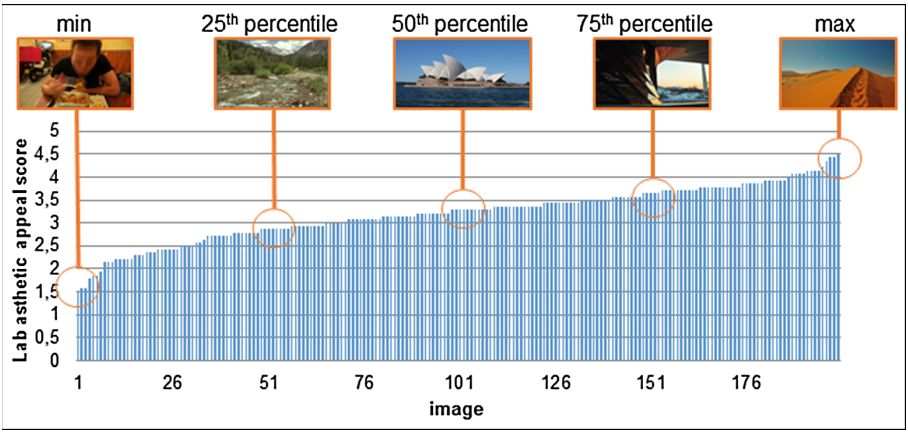

Fig. 2. Anchors used in the crowdsourcing-based image aesthetic quality assessments reported in [84]

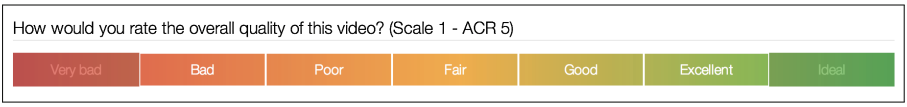

Fig. 3. ACR-5 scale with non clickable anchor points and a traffic-light semaphore design. The scale designs is available under Creative Commons Attribution 3.0 Austria License at https://github.com/St1c/ratings last accessed 14 Jun 2017.

5.2 Reliability Checks

QoE evaluations by their very nature are highly subjective and may differ significantly among the workers. Consequently, it is impossible to categorize subjective ratings as either 'correct' or 'incorrect'. To overcome this issue, reliability checks have to be added to a task in order to estimate the trustworthiness or reliability of a user. In particular, Hossfeld et al. propose to add one of the following elements in the test design to check the reliability of the users.

- Verification tests as reported by Alonso et al. and Downs et al. help in identifying automatization in the form of scripts, but can also be an indicator for sloppy workers and random clickers [1,14]. They include captchas or computation of simple text equations: "two plus 3=?", "Which of these countries contains a major city called Cairo? (Brazil, Canada, Egypt, Japan)".
- Consistency tests estimate the validity of a user's answer by asking, for example, at the beginning of the test, "In which country do you live?", followed later in the test by the question "In which continent do you live?"
- Content questions about the test allow to assess the attention of the user, for example, one can ask after showing a video clip "Which animal did you see in the video? (Lion, Bird, Rabbit, Fish)".

- If the correct result for certain test cases is known in advance, Hsueh et al. showed that so-called gold standard data can by utilized [41]: when a video clip under test, for example, does not contain any stalling, the following question could be asked: "Did you notice any stops to the video you just watched? (Yes, No)". Note, however that such questions can only be used to check for obvious impairments and not for the resulting ratings themselves.
- The repetition of test conditions can be used to check consistent user rating behavior. This can be seen as a special kind of consistency check but based on user ratings instead of additional information.
- Independent of the ratings or additional consistency questions, the general interactions of the user with the task interface can be monitored to unveil deviant behavior. Typically, the focus time of a video clip or the time it take the users to answer questions is monitored. Based on preliminary tests about how trustworthy users behave (used to identify 'normal behavior' or focus and answering times), an additional reliability score based thereon is computed.

Combining these elements also leads to an improved reliability of the results. These reliability tests may either be employed *a posteriori* after the test or alternatively already during the test. The *in momento* reliability checking proposed by Gardlo et al. also allows to identify reliable workers during the test, which allows to engage reliable users with more tasks directly in the current test [25].

After the conclusion of the test, commonly used outlier detection methodologies for the subjective ratings can also be used to detect users whose ratings significantly deviate from the average evaluations as usually represented by the Mean Opinion Scores (MOS), and in a non-systematic way, i.e. their ratings are not systematically above or below the average. For ACR or interval scales, the procedure proposed in ITU-R BT.500 [48] is most suitable. For paired-comparison based tests rating inversions as introduced by Xu and Chen, can be utilized [9,110]. Outlier detection should also include assessing the task execution time since it is a good indicator for the reliable task completion as proposed by Hossfeld, Redi and Korshunov, as workers may skip stimuli too fast without taking the time to properly evaluate them, intertwine the rating task with another task (e.g. web surfing), or get distracted during at least one test case by their environment (e.g. a phone call) [38,60,81]. The first two cases can be identified using the outlier detection from ITU-R BT.500 as workers identified to repeatedly score in an amount of time which is significantly lower or higher than average can be deemed unreliable. The last case can be identified by detecting unusually high evaluation times for a single stimulus. Redi et al. showed that this can be done by observing the standard deviation of the time taken by each worker to evaluate each stimulus as suggested in [81,83].

In their Best Practices for Crowdsourcing paper, Hossfeld et al. note that it is important to filter out all ratings from suspicious workers rather than individual ratings, as there may be hidden, not monitored influence factors for that worker (e.g. bad light conditions) or workers not conducting the task properly (e.g. wrongly understood instructions) [33].

5.3 Duration

In QoE crowdsourcing, Hossfeld et al. recommend campaigns to be fairly short (up to 10 min) to avoid boredom and unreliable behavior [33,38]. Traditional QoE tests typically involve tens or hundreds of stimuli, requiring participants to score for much longer timespans (typically between 30 min and one hour). Thus, to collect QoE scores for a large set of stimuli, researchers usually have to decompose the scoring task in a set of smaller tasks (i.e., campaigns), each one including a sub-set of the stimuli. Redi et al. replicated a laboratory-based experiment in crowdsourcing [81]. In the laboratory experiments, all participants evaluated a total of 200 images in a single session (with three small breaks in between), taking in total 40 min, approximately. Such long task duration could not be replicated in crowdsourcing; hence, the authors split-up the evaluation task into a number of sub-tasks (campaigns) including 20 images each. However, this approach increased tenfold the risk of context effects.

5.4 Payment

There are different motivations for users to participate in crowdsourcing as pointed out in Chap. 3, which aims at understanding the crowd and especially who they are and what their motivations are. As a key result of that chapter, payment is the major motivation for the crowd in commercial crowdsourcing platforms, and all other motivations are secondary. Still, it was observed that higher payments do not guarantee more success or better quality work. Also faster batch completion times cannot be achieved with a higher payment in general, even if some studies indicate that crowdsourcing users tend to choose mainly tasks with high rewards [2,96].

Varela et al. established two identical crowdsourcing campaigns on Web QoE assessment, which only differed in the reward to the workers [102]. In the second campaign, the users earned three times more money than the workers in the first campaign. The higher paid campaign led to significantly shorter completion time (3 h vs. 173 h), but the ratio of reliable users was lower (66% vs. 72%). As a result of the shorter completion times, the demographics of users was narrowed which may be caused by higher motivation of users to participate, time-zones of users, and the start time of the campaign. This effect may be considered when starting a crowdsourced QoE campaign, for example, by possibly throttling the execution, or by selecting users with a certain demographic background in order to obtain more representative population samples.

However, the major observation was that the mean user rating across all test conditions was slightly higher for the higher paid group (3.80 vs 3.60) [102]. A detailed analysis showed that the difference was statistically significant. A possible explanation may be that the users wanted to ensure to earn the reward by ‘pleasing’ the employer which was leading to higher ratings. In the tests, the normalization of the user ratings based on z-scores lead however to the same main effects and interactions. Thus, the normalized user ratings allowed to properly derive a QoE model. Redi et al. compared paid workers and volunteers when

rating the beauty of images and observed that paid users are more likely to commit to the execution of a crowdsourcing task [83]. However, again a bias of paid users to rate quality towards the higher end of the quality scale in contrast to voluntary users was observed.

There are however no general conclusions on payments and incentive design for crowdsourcing studies. For other applications of crowdsourcing beyond QoE testing, different results were observed. Harris et al. used crowdsourcing for screening a number of candidates applying for a job at a company and to conduct resume reviews [31]. Better incentive schemes increased the quality of work.

From these examples we summarize that the influence of payments needs to be considered, (a) in the analysis of the results, for example, using z-scores [57, 102], or removing worker bias [51], and (b) in the test design to ensure that the workers do not want to please the employer and use the entire scale, for example, by proper instructions and training [35,38]. Further, (c) reporting of payments is crucial in publications of crowdsourced QoE studies.

6 Conclusions

Crowdsourcing for QoE testing has seen a steep take up in numerous crowdsourced tests due to its promise to reach out to a large, diverse and global crowd in real life environments with short turn-around times. However, research practice has shown that crowdsourcing also hides many pitfalls due to the lack of direct visual and crowdworkers' feedback. Related to these pitfalls, this chapter discusses a number of these issues and their solutions by people that adopted crowdsourcing for QoE testing. Furthermore, a number of existing crowdsourcing platforms are reviewed and discussed with respect to their abilities for different types of QoE tests. The final overview of lessons learned can serve as guidelines for best practices in the experimental setup, data analytics and monetary incentives to be used for QoE testing in the crowd.

Acknowledgement. The authors want to thank Schloss Dagstuhl Leibniz-Zentrum für Informatik, the participants of Dagstuhl Seminar 15481 *Evaluation in the Crowd: Crowdsourcing and Human-Centred Experiments* as well as the Qualinet members that participated in the creation of *Best Practices and Recommendations for Crowdsourced QoE - Lessons learned from the Qualinet Task Force* [38]. Furthermore, this work was supported by the Deutsche Forschungsgemeinschaft (DFG) under Grants HO4770/2-1 and TR257/38-1.

A Appendix

Absolute vs. Comparative Tasks

In the chapter above we pointed out several weaknesses that are inherent to crowdsourced tests: for example, the lack of control, the diversity of the typically international crowdworker pool and problems with ACR scaling methodologies.

However, the feasibility of large-scale crowdsourced tests motivates to consider other options for judging stimuli besides the single stimulus absolute category rating scheme that is commonly used in QoE crowd testing. In this appendix we discuss the limitations of this traditional approach, and discuss the method of paired comparison as an alternative. We then outline psychometric scaling methods that can reconstruct qualities of stimuli from paired comparisons and conclude by stating the limitations of this approach.

Limitations of Absolute Category Rating

A major category of subjective testing in laboratory environments is aimed at assessing *quality of experience* (QoE) that commonly is defined as an expression of human expectations, feelings, perceptions, cognition and satisfaction with respect to a particular product, service or application. For such tasks where subjects are directly asked to express their subjective perception of a sensory event (visual or auditory event, encounter with a certain system etc.), it is necessary to assign (numerical) values to the related event. Typically, such assignments are achieved by using certain scales. As events can differ strongly, a number of different scales can be used. Among these, absolute category rating (ACR) scales [43] and Comparison Category Rating (CCR) scales [43] have emerged as well established examples for absolute rating or comparative rating tasks in laboratory settings. However, in recent years industry and research has rather shifted towards absolute category ratings (ACR) as they compare well to several other customer satisfaction measures that are typically used to assess product offerings, as well as questions about various aspects of customer interaction with services, products or companies [90]. Such ACR scales have several drawbacks:

- Their usage often varies between different users as they have different understandings of how to map their personal perception on the ACR scale.
- Users tend to avoid both ends of the scale, thus the votes tend to saturate before reaching the end points as shown by Keimel et al. and Gardlo et al. [23,56].
- Language and cultural differences regarding the 'distance' between scale labels for a given International Telecommunication Union (ITU) scale as reported by Jones et al. and Virtanen et al. make it difficult to compare results across cultural or international boundaries [52,105]. Rossi et al. termed these different usage patterns the *scale usage heterogeneity problem* [90].
- There is no well established method for detection of unreliable ratings (in the QoE domain; in other domains such as image labeling, there are established methods to build and use ground truths). Crowdworkers of a study may lack the necessary care and attention to give proper ratings.

One solution to address such issues is the usage of appropriate scale design as reported by Gardlo et al. that have shown to overcome certain limitations of ACR scales [24]. Possible other solutions can be the usage of training sessions within a task as described by Hossfeld et al. that help to align rating diversity and

scales usage across different subjects [38]. With such measures and controlled laboratory setups, MOS (mean opinion score) test results can be reproduced quite well in different laboratories.

Crowdsourcing has become an attractive alternative to laboratory studies for QoE assessments because of its efficiency in time and cost, the easy accessibility of crowdworkers form different parts of the world, and the availability of commercial platforms for crowdsourcing. However, with the crowdsourcing approach the limitations of ACR scales are even more severe. The workers can be expected to have a wider range of behavioral patterns with respect to the rating tasks, cultural differences may strongly vary, and their reliability can be poor. Moreover, in crowdsourcing environments it is important to work with a low number of training sessions in order not to lose crowdworkers' attention as shown by Hossfeld et al. [38].

Advantages of Quality Estimation by Paired Comparison

A promising replacement for ACR tests in the crowdsourcing environment is provided by paired testing via CCR procedures. It eliminates offsets between different crowdsourcing campaigns (and laboratory tests, too) as proposed by Chen and Wu [9,109]. In the following we discuss this approach, its properties and advantages.

In paired comparison studies, participants simply express their preference for one or the other of two presented stimuli. If desired, the option for a tie or the degree of preference ('slightly better', 'better', 'much better') may be offered as well. Therefore, training procedures to properly align user ratings with an ACR scale (Bad, Poor, Fair, Good, and Excellent) in the context of a specific application like quality of speech synthesis are not necessary for paired comparison studies. Moreover, the response time yielding a preference for a given pair of stimuli can be expected to be significantly smaller than for a single absolute category rating as participants need not remember and recall the appropriate quality levels from the training sessions for each pair over and over again.

Another important advantage of paired comparisons is that checking for consistency in the answers of an individual as well as for a group of participants is straightforward, by use of the transitivity property. If stimulus X is regarded superior to stimulus Y, and Y superior to Z, then the judgment for the pair (X, Z) should be in favor of X, of course. Therefore, consistency can be expressed as the fraction of judgments that adhere to the expectation due to the transitivity rule, with a fraction of 1.0 giving perfect consistency. A consistency fraction below some threshold may be an indication that the results of the corresponding worker in the crowdsourcing study is not reliable. The notion of consistency can be generalized to the case where paired comparisons are repeated many times, as in a study with several participants, by the well-known concepts of weak, moderate, and strong stochastic transitivity as reported in Bossuty et al. [4].

Reconstruction of Absolute Ratings from Paired Comparisons

Lastly, and perhaps most importantly, there are ways to reconstruct an absolute rating from the relative ones provided by paired comparisons. The simplest such procedures are common scoring schemes as in round robin sport tournaments. Each player (or team) carries out a match with each other player (or team). In each match points are given for a win or a draw. Then after all possible matches have been carried out the players' respective teams can be ranked according to the accumulated scores, which can also serve as absolute measures of performance.

However, the most commonly applied models applied to the problem of assigning a scalar value to some object quality, based on paired comparison, are probabilistic or statistical in nature. Assume that the object qualities are (uncorrelated) continuous random variables ordered along a line. Then distances of their respective means reflect their relative qualities which can be assessed empirically by pairwise comparisons. Each of these random variables has a corresponding probability density function (PDF) and with such a linear model it is the task on hand to estimate their unknown means. A judge, asked to compare any two of the objects, respectively their qualities, say A and B, can then be modeled as follows. A sample is drawn from each of the two distributions and the larger sample drawn determines the winner of the comparison. Repeating this procedure yields $N_{A,B}$ preferences of A over B and $N_{B,A}$ preferences of B over A. The fraction $N_{A,B}/(N_{A,B} + N_{B,A})$ can be regarded as an estimate of $P(A > B)$, the probability that A is better (larger) than B. When we assume certain PDFs for the distributions of the random variables with unknown means we can in principle calculate this probability $P(A > B)$ as a function of the difference of the means. On the other hand, replacing the probability $P(A > B)$ by its empirical estimate $P_{A,B} = N_{A,B}/(N_{A,B} + N_{B,A})$ and applying the inverse of this function will yield an estimate of the distance of the means.

In the classical model of Thurstone-Mosteller [72,100] the probability density functions are Gaussian, in the simplest case with equal variance. Here the estimate of the distance of the means simply is the inverse of the cumulative density function Φ of the standard normal distribution, applied to the empirical estimate $P_{A,B}$ of $P(A > B)$, up to a scale parameter that depends on the variance of the underlying distributions. Another popular linear model is the one of Bradley-Terry [6]. Here the logistic cumulative density function replaces the normal one, giving very similar results.

After deriving estimates for the distances $d_{i,j}$ between all the stimuli qualities, say A_i and A_j, we still need to reconstruct the linear ordering of all corresponding quality values A_i. This is a problem since a perfect one-dimensional embedding generally does not exist, as we cannot ensure that $d_{i,j} + d_{j,k} = d_{i,k}$ for all i, j, k. There are several approaches to define an appropriate ordering. The simplest one is given by the least-squares estimate, minimizing the sum of squared differences between the empirically estimated differences and the differences in the linear

ordering, i.e., $\sum_{i,j}(d_{i,j}-(A_i-A_j))^2$. With the adjustment that the mean quality is zero, $\sum_i A_i = 0$, one obtains the solution $A_j = \sum_i d_{i,j}$ for all j.

Another approach is given by the maximum likelihood method that has been shown to have significant advantages over the traditionally applied least-squares method. For the linear model of Thurstone-Mosteller the likelihood that a sample of the random variable with mean A_i is larger than a sample of the j-th random variable is proportional to $\Phi(A_i - A_j)$. Therefore, if $N_{i,j}$ denotes the number times the i-th stimulus was judged to be larger than the j-th in a pairwise comparison, the log-likelihood for these observations is proportional to $\sum_{i\neq j} N_{i,j} \log(\Phi(A_i - A_j))$. The minimization of this quantity is a convex optimization problem and, thus, readily solvable by numerical methods. Note that it is necessary to add a constraint such as $\sum_i A_i = 0$ in order to ensure an isolated, unique solution. It may be of advantage to generalize this approach to a maximum *a posteriori* estimate, for example, by including a Gaussian prior, amounting to subtracting $\frac{1}{2}\sum_i A_i^2$ from the above log-likelihood.

We conclude this short exposition about paired comparison by giving pointers to some selected literature that describes further details of the methods, their theory, and some examples, and by discussing the limitations of the method of paired comparisons and how they can be dealt with.

Selected References

The most comprehensive treatment of the overall subject matter of pairwise comparison, including a large chapter on linear models, can be found in the monograph *The method of paired comparisons* [13] by H.A. David (1988). In the technical report [101] the authors Tsukida and Gupta provided a modern and short account of the theory and practice of the linear models of Thurstone-Mosteller and Bradley-Terry, including some of the proofs. Moreover, the report studies the different models and computational approaches for them by simulation and lastly lists MATLAB code for the routines for the method of Thurstone-Mosteller. Wickelmaier and Schmid [106] presented details for improvements of the Bradley-Terry model including corresponding MATLAB functions. Wu et al. presented a comprehensive study comparing crowdsourcing using paired comparison with Mean Opinion Score for QoE of multimedia content [109]. They also introduced a general, systematic input validation framework for crowdsourced QoE assessments. Lee et al. proposed an extension of the Bradley-Terry linear model to generate intuitive measures of confidence besides the absolute quality scores [65].

One of the main applications of subjective quality assessment is the comparison of the performance of different, competing (objective) quality assessment algorithms. For example, the correlation between the subjective mean opinion scores and objective scores can be used to judge different algorithms. Hanhart et al. propose that one may also use the results of (subjective) paired comparisons directly without reconstructing absolute scalar quality ratings beforehand [30]. This can

be achieved by grouping responses for item pairs (A, B) into classes (e.g. $A > B$ and $A < B$) and then using a threshold t for any given objective quality measure μ to predict the corresponding classes for the same item pairs (A, B); i.e. (A, B) belongs to class $A > B$, if $\mu(A) > \mu(B)$. The performance of a quality assessment algorithm can finally be judged by classification error rates or the area under the corresponding receiver operator characteristic (ROC) curves (Area Under Curve).

Limitations of the Method of Paired Comparison

There are several problems with the method of pairwise comparisons that do not apply to direct absolute category rating.

The 0/1-Problem. When two stimuli presented in a paired comparison differ so strongly that all of the comparisons are in favor of one of them the so-called 0/1-problem occurs. Due to the infinite tail of the cumulative normal density function, the inverse of 0 or 1 will be infinite, yielding an infinitely large estimated distance between the stimuli qualities. One may simply ignore all such comparisons and base the calculations on such incomplete data. A better solution is to add a small amount (e.g. 1 vote) to the counts of the corresponding comparison outcomes. Still better yet is to apply the maximum likelihood method for the optimization as it does not apply the inverse of the cumulative density function and therefore does not require an artificial modification of the empirical data.

Scale and offset. The resulting values for the qualities A_i depend on an arbitrarily chosen scale (determined by the assumed variance of the corresponding random variables) and on an arbitrary offset (determined by the constraint $\sum_i A_i = 0$ or a similar one). Thus, for a comparison with some otherwise obtained ACR values an appropriate rescaling and shift must be carried out. For example, one may scale by equating the variance of the mean values A_i and shift to align the means of the means.

Complexity. Given N stimuli to be compared with each other there are $\frac{1}{2}N(N-1) = O(N^2)$ pairs of stimuli to be compared. A worker has to make a binary decision for each of these $O(N^2)$ pairs in order to generate a complete data set for the analysis. In comparison, for a study based on ACR each worker makes only N decisions, however, these are multiple choice instead of simply binary. In the practice of studies based on crowdsourcing and even more so in a laboratory setting the quadratic complexity may drive the cost and time for the experiment above the given limits for the study. The obvious way to deal with this issue is to carefully select the most relevant comparisons that should be made avoiding those that are more or less redundant. The methods for the analysis of the resulting comparisons have to be properly adapted to the fact that the data is incomplete. Several methods for such complexity reductions exist [18,66,98,110].

References

1. Alonso, O., Rose, D.E., Stewart, B.: Crowdsourcing for relevance evaluation. In: ACM SigIR Forum, vol. 42, pp. 9–15. ACM (2008)
2. Becker, M., Borchert, K., Hirth, M., Mewes, H., Hotho, A., Tran-Gia, P.: Microtrails: comparing hypotheses about task selection on a crowd sourcing platform. In: International Conference on Knowledge Technologies and Data-driven Business (I-KNOW), Graz, Austria, October 2015
3. Bhatti, N., Bouch, A., Kuchinsky, A.: Integrating User-Perceived quality into web server design. In: 9th International World Wide Web Conference, pp. 1–16 (2000)
4. Bossuyt, P.: A Comparison of Probabilistic Unfolding Theories for Paired Comparisons Data. Springer, Heidelberg (1990). doi:10.1007/978-3-642-84172-9
5. Bouch, A., Kuchinsky, A., Bhatti, N.: Quality is in the eye of the beholder: meeting users' requirements for internet quality of service. In: CHI 2000: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 297–304. ACM, New York (2000)
6. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
7. Callet, P.L., Möller, S., Perkis, A. (eds.): Qualinet white paper on definitions of Quality of Experience (2012)
8. Chen, K.T., Chang, C.J., Wu, C.C., Chang, Y.C., Lei, C.L.: Quadrant of euphoria: a crowdsourcing platform for QoE assessment. Network **24**(2) (2010)
9. Chen, K.T., Wu, C.C., Chang, Y.C., Lei, C.L.: A crowdsourceable QoE evaluation framework for multimedia content. In: Proceedings of the 17th ACM international conference on Multimedia, MM 2009, pp. 491–500. ACM (2009)
10. Cooke, M., Barker, J., Lecumberri, G., Wasilewski, K.: Crowdsourcing in Speech Perception. Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment, pp. 137–172 (2013)
11. Corriveau, P., Gojmerac, C., Hughes, B., Stelmach, L.: All subjective scales are not created equal: the effects of context on different scales. Sig. Process. **77**(1), 1–9 (1999)
12. Strohmeier, D., Jumisko-Pyykkö, S., Raake, A.: Toward task-dependent evaluation of Web-QoE: free exploration vs. who ate what? In: Globecom Workshops, pp. 1309–1313. IEEE (2012)
13. David, H.A.: The Method of Paired Comparisons. Griffin's statistical monographs, vol. 41, 2nd edn. Charles Griffin & Company Limited, London (1988)
14. Downs, J.S., Holbrook, M.B., Sheng, S., Cranor, L.F.: Are your participants gaming the system?: screening mechanical turk workers. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2399–2402. ACM (2010)
15. Egger, S., Gardlo, B., Seufert, M., Schatz, R.: The impact of adaptation strategies on perceived quality of HTTP adaptive streaming. In: Proceedings of the 2014 Workshop on Design, Quality and Deployment of Adaptive Video Streaming, pp. 31–36. ACM (2014)
16. Egger, S., Reichl, P., Hosfeld, T., Schatz, R.: time is bandwidth? narrowing the gap between subjective time perception and quality of experience. In: 2012 IEEE International Conference on Communications (ICC), pp. 1325–1330. IEEE (2012)
17. Egger, S., Schatz, R.: Interactive content for subjective studies on web browsing QoE: A Kepler derivative. In: ETSI STQ Workshop on Selected Items on Telecommunication Quality Matters, pp. 27–28 (2012)

18. Eichhorn, A., Ni, P., Eg, R.: Randomised pair comparison: an economic and robust method for audiovisual quality assessment. In: Proceedings of the 20th International Workshop on Network and Operating Systems Support for Digital Audio and Video, pp. 63–68. ACM (2010)
19. Engeldrum, P.G.: Psychometric Scaling: A Toolkit for Imaging Systems Development. Imcotek Press, Winchester (2000)
20. ETSI: Speech Processing, Transmission and Quality Aspects (STQ); Reference webpage for subjective testing. ETSI Standard TR 103 256, October 2014
21. Fliegel, K.: Qualinet multimedia databases v5. 5 (2014)
22. Freitas, P.G., Redi, J.A., Farias, M.C., Silva, A.F.: Video quality ruler: a new experimental methodology for assessing video quality. In: 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), pp. 1–6. IEEE (2015)
23. Gardlo, B.: Quality of experience evaluation methodology via crowdsourcing. Ph.D. thesis, University of Zilina (2012)
24. Gardlo, B., Egger, S., Hossfeld, T.: Do scale-design and training matter for video QoE assessments through crowdsourcing? In: Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia, pp. 15–20. ACM (2015)
25. Gardlo, B., Egger, S., Seufert, M., Schatz, R.: Crowdsourcing 2.0: enhancing execution speed and reliability of web-based QoE testing. In: International Conference on Communications, Sydney, AU, June 2014
26. Guse, D., Egger, S., Raake, A., Möller, S.: Web-QoE under real-world distractions: two test cases. In: 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 220–225. IEEE (2014)
27. Hands, D., Wilkins, M.: A study of the impact of network loss and burst size on video streaming quality and acceptability. In: Diaz, M., Owezarski, P., Sénac, P. (eds.) IDMS 1999. LNCS, vol. 1718, pp. 45–57. Springer, Heidelberg (1999). doi:10.1007/3-540-48109-5_5
28. Hanhart, P., Korshunov, P., Ebrahimi, T.: Crowd-based quality assessment of multiview video plus depth coding. In: IEEE International Conference on Image Processing, ICIP 2014. Paris France, April 2014
29. Hanhart, P., Korshunov, P., Ebrahimi, T.: Crowdsourcing evaluation of high dynamic range image compression. In: SPIE Optical Engineering + Applications. International Society for Optics and Photonics, San Diego, CA, USA, August 2014
30. Hanhart, P., Krasula, L., Le Callet, P., Ebrahimi, T.: How to benchmark objective quality metrics from paired comparison data? In: Quality of Multimedia Experience (QoMEX), pp. 1–6. IEEE (2016)
31. Harris, C.: You're hired! an examination of crowdsourcing incentive models in human resource tasks. In: WSDM Workshop on Crowdsourcing for Search and Data Mining (CSDM), pp. 15–18 (2011)
32. Hirth, M., Hoßfeld, T., Tran-Gia, P.: Anatomy of a crowdsourcing platform - using the example of Microworkers.com. In: Workshop on Future Internet and Next Generation Networks (FINGNet), Seoul, Korea, June 2011
33. Hossfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia, P.: Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. Trans. Multimedia **16**(2), 541–558 (2014)
34. Hoßfeld, T., Seufert, M., Hirth, M., Zinner, T., Tran-Gia, P., Schatz, R.: Quantification of YouTube QoE via crowdsourcing. In: Symposium on Multimedia, Dana Point, USA, December 2011

35. Hossfeld, T.: On training the crowd for subjective quality studies. VQEG eLetter **1**(1), 8 (2014)
36. Hoßfeld, T., Egger, S., Schatz, R., Fiedler, M., Masuch, K., Lorentzen, C.: Initial delay vs. interruptions: between the devil and the deep blue sea. In: QoMEX 2012, Yarra Valley, Australia, July 2012
37. Hoßfeld, T., Hirth, M., Korshunov, P., Hanhart, P., Gardlo, B., Keimel, C., Timmerer, C.: Survey of web-based crowdsourcing frameworks for subjective quality assessment. In: 2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–6. IEEE (2014)
38. Hoßfeld, T., Hirth, M., Redi, J., Mazza, F., Korshunov, P., Naderi, B., Seufert, M., Gardlo, B., Egger, S., Keimel, C.: Best practices and recommendations for crowdsourced QoE - lessons learned from the qualinet task force "Crowdsourcing" October 2014. https://hal.archives-ouvertes.fr/hal-01078761, lessons learned from the Qualinet Task Force "Crowdsourcing" COST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services (QUALINET)
39. Hoßfeld, T., Schatz, R., Biedermann, S., Platzer, A., Egger, S., Fiedler, M.: The memory effect and its implications on web QoE modeling. In: 23rd International Teletraffic Congress (ITC 2011), San Francisco, CA, USA (2011)
40. Hoßfeld, T., Schatz, R., Seufert, M., Hirth, M., Zinner, T., Tran-Gia, P.: Quantification of YouTube QoE via Crowdsourcing. In: IEEE International Workshop on Multimedia Quality of Experience - Modeling, Evaluation, and Directions (MQoE 2011), Dana Point, CA, USA, December 2011
41. Hsueh, P.Y., Melville, P., Sindhwani, V.: Data quality from crowdsourcing: a study of annotation selection criteria. In: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing, pp. 27–35. Association for Computational Linguistics (2009)
42. Huynh-Thu, Q., Garcia, M.N., Speranza, F., Corriveau, P., Raake, A.: Study of rating scales for subjective quality assessment of high-definition video. IEEE Trans. Broadcast. **57**(1), 1–14 (2011)
43. International Telecommunication Union: Methods for Subjective Determination of Transmission Quality. ITU-T Recommendation P.800, August 1996
44. International Telecommunication Union: Interactive test methods for audiovisual communications. ITU-T Recommendation P.920, May 2000
45. International Telecommunication Union: Vocabulary and effects of transmission parameters on customer opinion of transmission quality, amendment 2. ITU-T Recommendation P.10/G.100 (2006)
46. International Telecommunication Union: ITU-T recommendation e.800. Quality of Telecommunication Services: Concepts, models, objectives and dependability planning. terms and definitions related to the quality of telecommunication services. ITU-T Recommendation E.800, September 2008
47. International Telecommunication Union: Subjective video quality assessment methods for multimedia applications. ITU-T Recommendation P.910, April 2008
48. International Telecommunication Union: Methodology for the Subjective Assessment of the Quality of Television Pictures. ITU-R Recommendation BT.500-12, March 2009
49. International Telecommunication Union: Subjective Testing Methodology for web browsing. ITU-T Recommendation P.1501 (2013)
50. International Telecommunication Union: Subjective Methods for the Assessment of stereoscopic 3DTV Systems. ITU-R Recommendation BT.2021, July 2015

51. Janowski, L., Pinson, M.: Subject bias: introducing a theoretical user model. In: 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 251–256. IEEE (2014)
52. Jones, B.L., McManus, P.R.: Graphic scaling of qualitative terms. SMPTE J. **95**(11), 1166–1171 (1986)
53. Jumisko-Pyykkö, S., Hannuksela, M.M.: 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, pp. 63–72. ACM (2008)
54. Keelan, B.: Handbook of Image Quality: Characterization and Prediction. CRC Press, Boca Raton (2002)
55. Keelan, B.W., Urabe, H.: ISO 20462: a psychophysical image quality measurement standard. In: Electronic Imaging 2004, pp. 181–189. International Society for Optics and Photonics (2003)
56. Keimel, C., Habigt, J., Diepold, K.: Challenges in crowd-based video quality assessment. In: Forth International Workshop on Quality of Multimedia Experience (QoMEX 2012), Yarra Valey, Australia, July 2012
57. Keimel, C., Habigt, J., Horch, C., Diepold, K.: QualityCrowd - a framework for crowd-based quality evaluation. In: Picture Coding Symposium, Krakow, PL, May 2012
58. Keimel, C., Habigt, J., Horch, C., Diepold, K.: Video quality evaluation in the cloud. In: Packet Video Workshop, Munich, DE, May 2012
59. Korshunov, P., Cai, S., Ebrahimi, T.: Crowdsourcing approach for evaluation of privacy filters in video surveillance. In: 1st International ACM workshop on Crowdsourcing for Multimedia (CrowdMM 2012). ACM, Nara, October 2012
60. Korshunov, P., Nemoto, H., Skodras, A., Ebrahimi, T.: The effect of HDR images on privacy: crowdsourcing evaluation. In: SPIE Photonics Europe 2014, Optics, Photonics and Digital Technologies for Multimedia Applications, Brussels, Belgium, April 2014
61. Kraft, S., Zölzer, U.: BeaqleJS: HTML5 and JavaScript based framework for the subjective evaluation of audio quality. In: Linux Audio Conference, Karlsruhe, DE, May 2014
62. Kubey, R., Csikszentmihalyi, M.: Television and the Quality of Life: How Viewing Shapes Everyday Experience. A Volume in the Communication Series. L. Erlbaum Associates (1990). http://books.google.at/books?id=zk_Zg5fJSVwC
63. Laghari, K., Crespi, N., Connelly, K.: Toward total quality of experience: a QoE model in a communication ecosystem. IEEE Commun. Mag. **50**(4), 58–65 (2012)
64. Lebreton, P.R., Mäki, T., Skodras, E., Hupont, I., Hirth, M.: Bridging the gap between eye tracking and crowdsourcing. In: Human Vision and Electronic Imaging XX, San Francisco, California, USA, 9–12 February 2015, p. 93940W (2015)
65. Lee, J.S., De Simone, F., Ebrahimi, T.: Subjective quality evaluation via paired comparison: application to scalable video coding. IEEE Trans. Multimedia **13**(5), 882–893 (2011)
66. Li, J., Barkowsky, M., Le Callet, P.: Boosting paired comparison methodology in measuring visual discomfort of 3DTV: performances of three different designs. In: Proceeding SPIE Electronic Imaging-Stereoscopic Displays and Applications XXIV (2013)
67. Little, G., Chilton, L., Goldman, M., Miller, R.: TurKit: tools for iterative tasks on mechanical turk. In: Proceedings of the ACM SIGKDD Workshop on Human Computation, pp. 29–30. ACM (2009)

68. Mayo, C., Aubanel, V., Cooke, M.: Effect of prosodic changes on speech intelligibility. In: Interspeech. Citeseer (2012)
69. Mazza, F., Da Silva, M.P., Le Callet, P.: Would you hire me? Selfie portrait images perception in a recruitment context. In: IS&T/SPIE Electronic Imaging, p. 90140X. International Society for Optics and Photonics (2014)
70. Möller, S.: Quality Engineering - Qualität kommunikationstechnischer Systeme. Springer, Heidelberg (2010). doi:10.1007/978-3-642-11548-6
71. Möller, S., Raake, A.: Telephone speech quality prediction: towards network planning and monitoring models for modern network scenarios. Speech Commun. **38**, 47–75 (2002). http://portal.acm.org/citation.cfm?id=638078.638082, ACM ID: 638082
72. Mosteller, F.: Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. Psychometrika **16**(1), 3–9 (1951)
73. Naderi, B., Polzehl, T., Beyer, A., Pilz, T., Möller, S.: Crowdee: mobile crowdsourcing micro-task platform - for celebrating the diversity of languages. In: Proceeding of 15th Annual Conference of the International Speech Communication Assocation (Interspeech 2014) (2014)
74. Naderi, B., Polzehl, T., Wechsung, I., Köster, F., Möller, S.: Effect of trapping questions on the reliability of speech quality judgments in a crowdsourcing paradigm. In: 16th Annual Conference of the International Speech Communication Assocation (Interspeech 2015), ISCA, pp. 2799–2803 (2015)
75. Ouyang, Y., Yan, T., Wang, G.: CrowdMi: scalable and diagnosable mobile voice quality assessment through wireless analytics. IEEE Internet Things J. **2**(4), 287–294 (2015)
76. Parasuraman, A., Zeithaml, V.A., Berry, L.L.: A conceptual model of service quality and its implications for future research. J. Market. **49**, 41–50 (1985)
77. Pitrey, Y., Engelke, U., Barkowsky, M., Pépion, R., Le Callet, P.: Aligning subjective tests using a low cost common set. In: Euro ITV, IRCCyN-Contribution (2011)
78. Polzehl, T., Naderi, B., Köster, F., Möller, S.: Robustness in speech quality assessment and temporal training expiry in mobile crowdsourcing environments. In: 16th Annual Conference of the International Speech Communication Assocation (Interspeech 2015), ISCA, pp. 2794–2798 (2015)
79. Raake, A.: Speech Quality of VoIP: Assessment and Prediction. Wiley, New York (2006)
80. Rainer, B., Waltl, M., Timmerer, C.: A web based subjective evaluation platform. In: Workshop on Quality of Multimedia Experience, Klagenfurth, AT, July 2013
81. Redi, J., Hoßfeld, T., Korshunov, P., Mazza, F., Povoa, I., Keimel, C.: Crowdsourcing-based multimedia subjective evaluations: A case study on image recognizability and aesthetic appeal. In: Workshop on Crowdsourcing for Multimedia, Barcelona, ES, October 2013
82. Redi, J., Liu, H., Alers, H., Zunino, R., Heynderickx, I.: Comparing subjective image quality measurement methods for the creation of public databases. In: IS&T/SPIE Electronic Imaging, p. 752903. International Society for Optics and Photonics (2010)
83. Redi, J., Povoa, I.: Crowdsourcing for rating image aesthetic appeal: Better a paid or a volunteer crowd? In: 3rd International ACM workshop on Crowdsourcing for Multimedia (CrowdMM 2014), Orlando, FL, USA, November 2014

84. Redi, J., Siahaan, E., Korshunov, P., Habigt, J., Hossfeld, T.: When the crowd challenges the lab: lessons learnt from subjective studies on image aesthetic appeal. In: Proceedings of the Fourth International Workshop on Crowdsourcing for Multimedia, pp. 33–38. ACM (2015)
85. Reichl, P.: From charging for Quality of Aervice to charging for Quality of Experience. Annales des Télécommunications **65**(3–4), 189–199 (2010)
86. Rerabek, M., Yuan, L., Krasula, L., Korshunov, P., Fliegel, K., Ebrahimi, T.: Evaluation of privacy in high dynamic range video sequences. In: SPIE Optical Engineering + Applications. International Society for Optics and Photonics, San Diego, CA, USA, August 2014
87. Ribeiro, F., Florencio, D., Nascimento, V.: Crowdsourcing subjective image quality evaluation. In: Image Processing. Brussels, BE, September 2011
88. Ribeiro, F., Florencio, D., Zhang, C., Seltzer, M.: CrowdMOS: an approach for crowdsourcing mean opinion score studies. In: International Conference on Acoustics, Speech and Signal Processing. Prague, CZ, May 2011
89. de Ridder, H.: Cognitive issues in image quality measurement. J. Electron. Imaging **10**(1), 47–55 (2001)
90. Rossi, P.E., Gilula, Z., Allenby, G.M.: Overcoming scale usage heterogeneity. J. Am. Stat. Assoc. **96**(453), 20–31 (2001). http://www.tandfonline.com/doi/abs/10.1198/016214501750332668
91. Rubino, G.: Quantifying the quality of audio and video transmissions over the internet: the PSQA approach. In: Design and Operations of Communication Networks: A Review of Wired and Wireless Modelling and Management Challenges. Imperial College Press (2005)
92. Sackl, A., Schatz, R.: Evaluating the impact of expectations on end-user quality perception. In: Proceedings of International Workshop Perceptual Quality of Systems (PQS), pp. 122–128 (2013)
93. Sanchez-Iborra, R., JPC Rodrigues, J., Cano, M.D., Moreno-Urrea, S.: QoE measurements and analysis for VoIP services. Emerging Research on Networked Multimedia Communication Systems, p. 285 (2015)
94. Schatz, R., Egger, S.: Vienna surfing - assessing mobile broadband quality in the field. In: Taft, N., Wetherall, D. (eds.) Proceedings of the 1st ACM SIGCOMM Workshop on Measurements Up the STack (W-MUST). ACM (2011)
95. Schatz, R., Hoßfeld, T., Janowski, L., Egger, S.: From packets to people: quality of experience as a new measurement challenge. In: Biersack, E., Callegari, C., Matijasevic, M. (eds.) Data Traffic Monitoring and Analysis. LNCS, vol. 7754, pp. 219–263. Springer, Heidelberg (2013). doi:10.1007/978-3-642-36784-7_10
96. Schnitzer, S., Rensing, C., Schmidt, S., Borchert, K., Hirth, M., Tran-Gia, P.: Demands on task recommendation in crowdsourcing platforms - the workers perspective. In: CrowdRec Workshop, Vienna, Austria (9 2015)
97. Shaikh, J., Fiedler, M., Paul, P., Egger, S., Guyard, F.: Back to normal? Impact of temporally increasing network disturbances on QoE. In: 2013 IEEE Globecom Workshops (GC Workshops), pp. 1186–1191. IEEE (2013)
98. Silverstein, D.A., Farrell, J.E.: Quantifying perceptual image quality. In: PICS, vol. 98, pp. 242–246 (1998)
99. Soldani, D., Li, M., Cuny, R.: QoS and QoE management in UMTS cellular systems. Wiley, West Sussex (2006)
100. Thurstone, L.L.: A law of comparative judgment. Psychol. Rev. **34**(4), 273 (1927)
101. Tsukida, K., Gupta, M.R.: How to analyze paired comparison data. Technical report, DTIC Document (2011)

102. Varela, M., Mäki, T., Skorin-Kapov, L., Hoßfeld, T.: Increasing payments in crowdsourcing: don't look a gift horse in the mouth. In: 4th International Workshop on Perceptual Quality of Systems (PQS 2013), Vienna, Austria (2013)
103. Varela, M., Mäki, T., Skorin-Kapov, L., Hoßfeld, T.: Towards an understanding of visual appeal in website design. In: 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX), pp. 70–75. IEEE (2013)
104. Varela, M., Skorin-Kapov, L., Mäki, T., Hoßfeld, T.: QoE in the web: a dance of design and performance. In: 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX), pp. 1–7. IEEE (2015)
105. Virtanen, M., Gleiss, N., Goldstein, M.: On the use of evaluative category scales in telecommunications. In: Human Factors in Telecommunications (1995)
106. Wickelmaier, F., Schmid, C.: A matlab function to estimate choice model parameters from paired-comparison data. Behav. Res.h Methods Instrum. Comput. **36**(1), 29–40 (2004)
107. Winkler, S., Mohandas, P.: The evolution of video quality measurement: from PSNR to hybrid metrics. IEEE Trans. Broadcast. **54**(3), 660–668 (2008). http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4550731
108. Wolters, M.K., Isaac, K.B., Renals, S.: Evaluating speech synthesis intelligibility using Amazon mechanical turk. In: 7th Speech Synthesis Workshop (2010)
109. Wu, C.C., Chen, K.T., Chang, Y.C., Lei, C.L.: Crowdsourcing multimedia QoE evaluation: a trusted framework. IEEE Trans. Multimedia **15**(5), 1121–1137 (2013)
110. Xu, Q., Huang, Q., Jiang, T., Yan, B., Lin, W., Yao, Y.: HodgeRank on random graphs for subjective video quality assessment. Trans. Multimedia **14**(3), 844–857 (2012)
111. Yu-Chuan, Y., Chu, C.Y., Yeh, S.L., Chu, H.H., Huang, P.: Lab experiment vs. crowdsourcing: a comparative user study on Skype call quality. In Proceedings of the 9th Asian Internet Engineering Conference, pp. 65–72 (2013)
112. Zinner, T., Hirth, M., Fischer, V., Hohlfeld, O.: Erwin - enabling the reproducible investigation of waiting times for arbitrary workflows. In: 8th International Conference on Quality of Multimedia Experiene (QoMEX), Lisbon, Portugal, June 2016