Best Practices and Recommendations for Crowdsourced QoE

Lessons learned from the Qualinet WG2 Task Force "Crowdsourcing"

Tobias Hoßfeld, Matthias Hirth, Judith Redi, Filippo Mazza, Pavel Korshunov, Babak Naderi, Michael Seufert, Bruno Gardlo, Sebastian Egger, Christian Keimel

Whitepaper version: October 4, 2014

COST Action IC1003 European Network on Quality of Experience in Multimedia Systems and Services (QUALINET)

Preface

Crowdsourcing is a popular approach that outsources tasks via the Internet to a large number of users. Commercial crowdsourcing platforms provide a global pool of users employed for performing short and simple online tasks. For quality assessment of multimedia services and applications, crowdsourcing enables new possibilities by moving the subjective test into the crowd resulting in larger diversity of the test subjects, faster turnover of test campaigns, and reduced costs due to low reimbursement costs of the participants. Further, crowdsourcing allows easily addressing additional features like real-life environments.

Crowdsourced quality assessment however is not just a straight-forward implementation of existing subjective testing methodologies in an Internet-based environment. Additional challenges and differences to lab studies occur, in conceptual, technical and motivational areas [9, 25, 26]. For example, the test contents need to be transmitted to the user over the Internet; test users may have low resolution screens influencing the user experience; also users may not understand the test or do not execute the test carefully resulting in unreliable data.

This white paper summarizes the recommendations and best practices for crowdsourced quality assessment of multimedia applications from the Qualinet Task Force on "Crowdsourcing". The European Network on Quality of Experience in Multimedia Systems and Services Qualinet (COST Action IC 1003, see `www.qualinet.eu`) established this task force in 2012 which has more than 30 members. The recommendation paper resulted from the experience in designing, implementing, and conducting crowdsourcing experiments as well as the analysis of the crowdsourced user ratings and context data. For understanding the impact of the crowdsourcing environment on QoE assessment and to derive a methodology and setup for crowdsourced QoE assessment, data from traditional lab experiments were compared with results from crowdsourcing experiments. Within the crowdsourcing task force, several different application domains and scientific questions were considered, among others

- video and image quality in general,
- QoE for HTTP streaming [31, 32] and HTTP adaptive streaming [19, 30],
- selfie portrait images perception in a recruitment context [10],
- privacy in HDR images and video [39, 20, 36],
- Compression of HDR images [37] [38],
- evaluation of 3D video [38],
- image recognizability and aesthetic appeal [13, 13],
- multidimensional modeling of web QoE [14],
- QoE factors of cloud storage services [21].

From a crowdsourcing perspective, the following mechanisms and approaches were investigated which are relevant to understand for crowdsourced quality assessment.

- Motivation and incentives, e.g. extrinsic motivation scale [2]
- Influence of payments [14, 12]
- Impact of task design [1] and affective crowdsourcing [11]
- Impact of crowdsourcing platforms selection [4, 12, 34]
- Reliability methods and screening mechanisms [5, 6, 31, 24], monitoring result quality [15]
- Development of crowdsourcing frameworks and platforms [3, 6, 7, 8]

As an outcome of the task force, scientific papers on best practices and crowdsourcing for QoE assessment in general were published.

- Challenges in Crowd-based Video Quality Assessment [9, 24, 25, 26, 33]
- Crowdsourcing in QoE Evaluation [25] and best practices for QoE crowdtesting [24, 26]
- Survey of web-based crowdsourcing frameworks for subjective quality assessment [27]

The scope of this white paper is to share practical issues and best practices in crowdsourcing experiments, to summarize the major lessons learned and to give key recommendations from the Qualinet task force. Thereby, the authors especially focus on aspects which are highly important in practice, but often not explained or discussed in scientific research papers for various reasons. Due to the large number of contributors, the document will not reflect the opinion of each individual person at all points. Nevertheless, each section of this white paper briefly formulates an important lesson learned.

Contents

1 Use common software without requiring admin installations!

Crowdsourcing test can be basically implemented in any programming language and could require an arbitrarily complex setup from the participant. However, this dramatically decreases the number of participating users and also increases the monetary costs to recruit test participants. In order to find a large number of test participants, it is recommended to use easy-to-use software tools, which do not need any installation, and also reduce the amount of data, which has to be downloaded by the users during the test.

One of the most suitable techniques for developing portable crowdsourcing tests is the implementation as web page or web application. In principle only a browser is required to access and participate in the test. However, in most cases additional constraints exists, e.g. specific a browser or browser version, enabled JavaScript, or installed third party extensions like Adobe Flash. Nevertheless, these constraints can usually be tested automatically and appropriate information messages, e.g. to enable JavaScript, can be provided to the participating users. The centralized character of this implementation approach also enables easy support and changes of the test implementation. A newly version is immediately available to all participants and no previous versions have to be removed/replaced like in decentralized implementations.

Web based test are highly portable but still the amount of transferred data can impose problems due to bandwidth limitations of participants or even unintentionally alter the test stimuli. Considering the evaluation of high definition video content, the size of the videos can result in a large amount of data, which has to be transferred to the participant. This test preparation time has also to be considered in the reward and has to be clearly communicated to the participants before starting the test. When evaluating e.g. stalling pattern of streaming services, the streaming via the Internet can also result in unwanted impairments, namely in additional stallings. Therefore, in some cases it is required to pre-cache all data required during the test at the client side, before starting the actual test [31].

Furthermore, also the server side infrastructure has to be considered during the implementation. Some Crowdsourcing tests can easily be scaled to several hundreds of participants in only a few minutes imposing significant network and computation load to the server and reporting infrastructure. This can be overcome by a careful scaling of the experiment of usage appropriate hardware dimensioning.

2 Simplify your questions!

One of the major differences between web based crowdsourcing experiments and traditional laboratory tests is the unsupervised environment. This includes both, the unknow surrounding conditions of the test participant as well as the lack of information about the participants themselves due to the anonymous recruiting process. Moreover, a direct interaction between the participants and the experimenter is usually not given. This makes it hard for test participants to clarify questions arising during a test and hard for the experimenter to identify misunderstandings of the test instructions. This issue is even amplified by the diversity of the test participants. Compared to laboratory tests, crowdsourcing users differ more in terms of spoken language, cultural background [17], background knowledge, used devices, etc. However, simple guidelines can help to minimize misunderstandings by the test participants.

The usage of simple English or native-language instructions helps non-native speakers to understand the instructions without using additional language resources, e.g. dictionaries. For most crowdsourcing users working on micro-tasks it is not economical to spend a lot of effort on understand the instructions of an individual task. Consequently, the users will either skip the task if they do not understand the instructions or try to complete the task to the best of their knowledge leading to low quality results.

Besides using a simple language it is also necessary to avoid technical or scientific terms. In an initial iteration of the YouTube QoE study presented in [31] the test participants we asked "to rate how the quality of experience of the [presented] video gets worse by stalling". The results obtained from this test indicated that the video quality is independent of the number and length of the stalling events, which is unintuitive. A redesign of the user test, including a reformulation of the task description and a detail explanation of the word stalling resulted in a significant increase of the result quality and a positive user feedback: "Thanks, for including the meaning of Stalling in your survey, as this helped me answer better than previously". A similar issue can be observed in [11]. Here, the change to a more intuitive rating scale improved the quality of the crowdsourcing result significantly.

In order to test the suitability of a task design it is recommended to apply preliminary tests with a small number of known test users. These testers should not be familiar with the research topic or the purpose of the test in order to exhibit similar background knowledge as a regular crowdsourcing user. One possibility to acquire such test users is asking via social networks.

3 Include proper training sessions!

The conceptual differences between crowdsourced QoE studies and studies conducted in a laboratory environment arise mainly from two issues: firstly, crowdsourcing tasks are usually much shorter (5-15 minutes) than comparable tests in a laboratory and secondly, the crowdsourcing environment lacks a test supervisor to provide direct and immediate feedback to the test subjects. The test subjects are only guided by a web interface through the tests that provide an explanation of the test, what to evaluate and how to express their opinion and/or ratings. The training of the subjects is mostly conducted by means of qualification tests that on the one hand enable the subjects to practice, but on the other hand also allow the test supervisor to assess to a certain extent the subjects' understanding of the test setup. Nevertheless, if problems due to a lack of understanding of the test procedures by the test subjects occur e.g. uncertainty about rating scales, appropriate mechanisms or statistical methods have to be applied. In particular, it is more difficult to ensure a proper training of the subjects as no direct feedback between supervisors and subjects during the training phase of the test is possible and thus potential misunderstandings are neither recognised by the test supervisor nor can any clarification questions of the subjects be addressed. Therefore it is essential to address any known issues from similar laboratory-based tests in the training phase with a particular consideration of the fact that no direct feedback is possible e.g. for example by providing short and illustrated (or even animated) explanations of the procedures considering questions commonly encountered in a similar laboratory setup. Still, the already short task duration in crowdsourcing and the fact that the qualification phase for familiarization of the subjects with the test structure and procedures is not included in the analysis, can decrease the efficiency of a test and increase the costs.

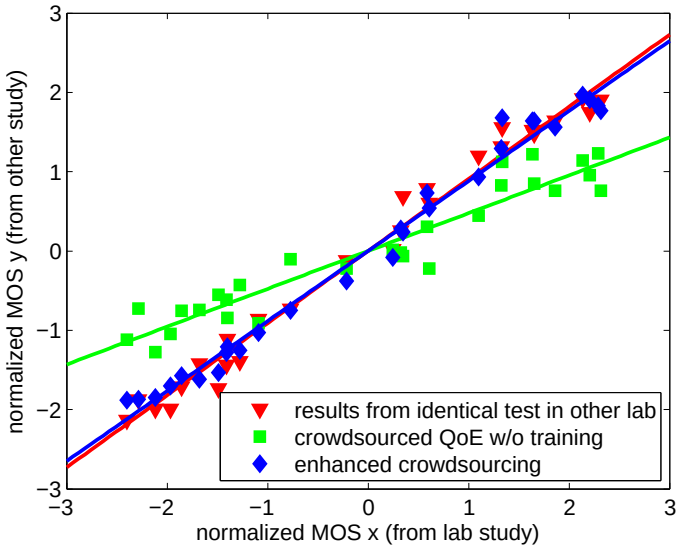


Figure 1: Experiments on video quality were conducted in two different laboratories with identical test setup. The study was crowdsourced a) without training and reliability mechanisms and b) with proper test instructions and a training mode as well as basic reliability mechanisms.

Without any worker training and reliability questions the results are significantly different than in a traditional laboratory environment or when using an advanced crowdsourcing design [25]. Figure 1 shows the results for video quality assessment conducted in two different test laboratories. The obtained mean opinion scores are normalized by the average of all user ratings. It can be seen that the results from the two laboratories provide the same user ratings. The experiments were also conducted in a crowdsourcing environment. The first experiments did not include proper training phases and it can be seen that the results strongly disagree from the tests conducted in the laboratory. However, after including training phases and basic reliability checks in order to confirm if the user understood the test properly, a very good match can be observed which is similar to the agreement between the results from the two laboratories. A conclusion of this observation is that training phases must be included in the task design!

4 Integrate a feedback channel!

Even with simplified language, proper instructions, and training sessions test participants might face issues while participating in tests. These issues might either occur due to misunderstandings or other unclear points, but also due to hard- and software issues. Therefore, it is important to provide a feedback channel to the users to contact the experimenter. This feedback channel has to implement three main properties: *accordance with the platform's terms of use (TOS)*, *robustness*, and a *permanent availability* throughout the test.

The *accordance with the TOS* of the crowd-providing platform can be challenging, because some providers do not allow contact between workers and employers outside the provided system.

In this case the platform might provide a messaging system or another platform specific solutions has to be found. It no feedback possibility is given, the platform provider should be changed. If custom feedback solutions are possible, they should be implemented in a *robust* way. This means the feedback channel has to be available, even if the actual task setup is broken, e.g. due to server issues, or the lack of technical requirements at the worker side. Otherwise, workers facing these very issues are not able to communicate them and the issues remain undetected. Moreover, the feedback channel should be accessible throughout the whole task, not only at the end of it. Questions might occur at an intermediate step or specific technical issues that hinder the worker to complete the task. If here a feedback is only available at the end of the tasks these issues also remain undetected.

Multiple possibilities exist to implement feedback channel for a crowdsourcing task. The most easiest option is adding a feedback form. However, this is usually only added at the end of a test and because of this neither robust nor always accessible. Moreover it is only a one-way communication from worker to employer. External, interactive feedback channels, which are independent of the task execution are more suitable here. These feedback channels could include e.g. live chat or email support for small test or forum threads for larger test groups. The contact details, respectively the links to the forum threads can easily be integrated in every step of the task enabling both robustness and availability of the channel.

In general, all questions from the users should be answered. An intensive discussion with the workers and a reasonable support helps to improve the task design and respectively the task results. Moreover, it helps to increase the employers reputation, as workers tend to gather in virtual communities and share their experiences with certain employers and tasks.

5 Use event logging!

Besides pure user based reliability test, additional automatic and semi automatic measures can be integrated in task. Analysing user interactions with the test is a recommended way to find out what really happened during the test execution. Thereby, a participating users actions and test conditions are be monitored and logged including their respective timestamps. User actions/interactions that are part of the test (e.g., clicking behavior) are as important as user actions which modify the test environment (e.g., window focus, window resize, page reload, switching of tab). The typical web browsing behavior of switching between different tabs could for example lead to the miss of a test condition because the user was watching another tab.

Besides estimating the reliability of a users, event logging can also help to debug a crowdsourcing tasks. Tests are usually implemented as web applications and are consequently executed in a very diverse software and hardware environment. Different browsers might render a test page differently which might result in unexpected test conditions. Moreover, tests might require a third party plugin (e.g., Adobe flash player for video tests) which could work in some browsers only, or which could be available in several versions which work slightly differently. The resulting test conditions should be logged, include overall test conditions like page load times, but also specific conditions, e.g., start of video playback in a video test. Also extraordinary events, which can alter or disturb a test condition, have to be logged, e.g., stalling events in a video test.

The automatic evaluation of the event logs during the test can then be used to give immediate feedback to the worker. For example an error message can be displayed if a missing plugin is

detect. Moreover, a user can also be warned if he did not watch the video as expected but switched to another tab instead [31]. During the analysis of the test results, the logged events can be used to apply filter rules [24]. Here it can be checked whether the desired test conditions actually were perceived by the test user. Furthermore, more sophisticated reliability checks can be performed to find out whether the test users conducted the test in an expected way [15]. In case the event logs show an unusual test or user behavior, the results for this user should be excluded from the overall analyses. Moreover, abnormalities in the event logs provide valuable insights for further tests. If a systematic error of test execution is visible, e.g., because of browser compatibility, the test can be improved. If user behavior is consistently unexpected for some parts or even the whole test, the test design has to be improved. For example, long response times for a specific questions could indicate that the question was too hard and should be reformulated.

Extensive user monitoring can provide valuable insights in the task interactions of the participants. However, it can also be considered as intrusive as a large amount of data about the user is gathered, sometimes including personal data, e.g. the IP address. Even if the implementation of the monitoring in a web application guarantees a certain amount of privacy, because only the interactions with the test page can be observed, the gathered data should be handled carefully.

6 Include reliability checks! In the test design...

The remoteness of the test participants does not only impose challenges, because of misunderstanding and technical problems. The anonymity also encourages some participants to work sloppy or to cheat in order to increase their income, by maximizing the number of completed task per time. Numerous approaches already exist to identify these user, however, most of them require objective tasks, like text transcription. The results of these tasks can easily be categorized in either 'correct' or 'incorrect'. In contrast, QoE evaluations are highly subjective and may differ significantly among the participants. Consequently, it is impossible to identify 'correct' subjective ratings.

To overcome this issue, reliability checks have to be added to a task in order to estimate the trustworthiness or reliability of a user. This in turn can be used to estimate the reliability of the ratings given by the very user. Reliability tests may include consistency checks, content questions, gold standard approaches or tests verifying the users attention or familiarity with the test. During or after the test, the results from those checks and additional questions are then analyzed in order to identify unreliable users. In contrast to outlier detection which aims at identifying individual outliers, i.e. individual user ratings are rejected, reliability checks aim at identifying unreliable users and the rejection of all ratings from those unreliable users. This in turn requires that the identification of unreliable users is only based on information which is not related to the user ratings. In particular, the following elements may be added in the test design to check the reliability of the users. Combining these elements also leads to an improved reliability of the results [24].

1. Verification tests [41, 42], including captchas or computation of simple text equations: "two plus 3=?", "Which of these countries contains a major city called Cairo? (Brazil, Canada, Egypt, Japan)". Captchas and the computation of simple equations help to identify scripts and bot automatically submitting results. Further verification questions might be added as indicators for sloppy workers or random clickers.

2. Consistency tests: At the beginning of the test, the user is asked "In which country do you live?". Later after several steps in the test, the user is asked "In which continent do you live?". Here it is important that the user cannot look up the previous answer. This approach helps to estimate the validity of the users answers, as random clickers might not remember their first anwer.
3. Content questions about the test: "Which animal did you see in the video? (Lion, Bird, Rabbit, Fish)". This type of question can be used as an indicator of the participants attention during the test. The content questions should be rather easy so that misunderstandings or language issues are avoided, but still should not be answerable with pure guessing.
4. Gold standard data [43]: "Did you notice any stops to the video you just watched? (Yes, No)", when the actual test video played without stalling. Usually gold standard data refers to tasks for which the correct result is known in advance, e.g. the correct transcription of a text on a picture. However, as discussed above it is not meaningful to define e.g. a gold standard MOS rating for a given impairment as this is a subjective rating. In contrast, gold standard data in QoE evaluation should aim at checking obvious impairments like the number of stallings or the presence of (significant) graphical distortions, not the resulting ratings.
5. Repetition of test conditions 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. Repetition of test conditions might be used to minimize the need of additional consistency questions but during the evaluation familiarization and memory effects have to be considered.

The important thing to keep in mind is not to add too many reliability items and questions, as otherwise the assessment task will become too lengthy. Further, too many of these questions may give a signal of distrust to the users. As a result, users may abort the survey. In general, incentives and proper payment schemes depending on the actual work effort are the key to high quality work. Application-layer monitoring or event logging is also helpful to detect any problems, e.g. monitoring and analyzing response times of users.

7 Include reliability checks! ... during the test ...!

The majority of the crowdsourcing assessments rely on the reliability checking and examining of the results after the campaign is finished, i.e. a posteriori. However with such an approach, reliable users are only discovered after the assessment is finished, diminishing the advantage to engage reliable users with more tasks directly, while they are still in the testing application. Alternative approach has been proposed in [6], where the reliability profile of the user is being built continuously, while the user is executing the task assigned to him. Such 'in momento' reliability checking offers quicker campaigns execution, while also increasing the reliability of the results.

While a posteriori designs often rely on repetitive hiring schemes, where reliable users are invited to participate in further campaign of the employer, it has been shown that this might also lead to the exhaustion of the crowd at the risk of declining motivation and poor rating performance. In momento approach successfully addresses this problem, since it avoids repetitive

hiring of the users, utilizing the advantages of huge workforces available nowadays on crowd-providing platforms. The campaign is offered to thousands of users in a single take, and reliable users are directly engaged by additional tasks for additional reward.

Additionally, reliability checks during the test and their direct evaluation reduce the administrative overhead introduced by traditional a posteriori approaches that require extensive data cleaning and group generation with repeated campaign runs. In addition, in momento approach allows for building a rapid feedback component for better communication with test participants. Such component is used to directly communicate any suspicious behavior to the user. It enables users to reflect on their performance and to choose whether to stop or to continue the testing process.

In momento reliability checking utilizes the multi-stage design of the crowdsourcing assessment and relies on numerous reliability checks during the test.

1. Reversed scales order for questions not related to the test content [6]: For better engagement of the users attention, simple test asks about maximum and minimum visible number on the picture. The scales used for providing the answers are presented in the reversed order (lowest to highest, highest to lowest), and also includes false answers, which are not presented on the test picture at all. Random clickers and users not focusing enough on the test are quickly discovered and penalized.
2. Ask about artifacts, which are not discoverable by a reliable user: User is asked to mark several visible images on the screen. Images presented to the user includes also invisible shapes, which are also markable, adding additional reliability check for random clickers (reliable users would never found these). Clicks on different area that the actual visible patterns or high number of clicks, which are typical for random clickers and users trying to found all visible images are directly analyzed and suspicious behavior is recorded into the user's reliability profile.
3. Altering test patterns for each test execution: Random movement of the testing shapes and numbers prevents cheating by sharing of correct locations between participants.
4. Checking execution and focus time. Each stage of the test has a certain defined minimal duration, i.e. time needed to finish the test. If the time spent on the page is very short, or several orders higher than usual, user is considered unreliable.
5. Additional test related to the content being tested. Apart from the test listed above, in video quality testing one can rely also for the checks related to the playback of the video. It is necessary to check if the video was played to the end, whether the user paused the playback, if the sound is presented, check also the volume settings, etc.

8 Include reliability checks! ...after the test!

Reliability checks can be performed also based on the outcomes of the test. A typical approach, developed already for lab-based testing, is the detection of outlier participants, which provided evaluation that significantly depart from the average evaluations expressed by the crowd, and in a non-systematic way (i.e., their evaluation are not systematically above or below the average, which may simply indicate a different usage of the scoring method rather than a poor understanding of instruction or a sloppy performance of the task). To detect scoring outliers when using ACR or interval scales, the procedure proposed in the ITU BT.500 [44] is most suitable,

and widely adopted also for lab participant screening. For paired-comparison based experiments, instead, it was proposed by [45, 46] to check for inversions in the judgment of pairs involving the same images.

Outlier detection should be also deployed to control for a second essential indicator of reliable task completion, which is task execution time. Whereas in Lab-based experiments the presence of an experimenter forces the participant to be fully committed to the task, dedicating to it the proper attention, in CS it is impossible to control for this. It may therefore happen that:

1. workers skip across stimuli as fast as possible, without taking the time to properly evaluate them
2. workers intertwine the execution of the task with other tasks, e.g. surfing the web or playing a game, thus being constantly distracted and taking longer in evaluating stimuli
3. workers get distracted at a specific point in time by a different task, such as talking on the phone or preparing a coffee, thereby taking an unusually long time for evaluating one stimulus in particular (during which they interrupted their evaluation task to perform the distracting task).

Should any of these three scenarios be verified, the corresponding evaluations cannot be trusted to be reliable. To identify scenario (1) and (2), the outlier detection procedure advised in ITU BT.500 can be applied using as a dependent variable the scoring time per participant and per stimulus. Participants identified to repeatedly score in an amount of time which is significantly lower or higher than average, can then be deemed unreliable and excluded from further analysis. To identify scenario (3), it is necessary to capture unusually high evaluation times for a single stimulus. In [13] and [12] it was proposed to observe this through the standard deviation of the time taken to evaluate each stimulus. In [13] for example, a test was run in three different regions (Europe, US and Asia), consisting in the evaluation of the aesthetic appeal of 20 images on an ACR scale. The authors observed that the Asian participants took on average much longer than the other participants to complete the entire test. Taking a closer look at the mean test completion time, authors found that for European and US participants the median test completion time was close to the mean completion time, but in the case of asian participants the mean was significantly larger than the median. This hinted the authors that some of the asian participants may have had with very large evaluation times for some images, as measured from the moment when the image was displayed until the time the evaluation was inserted is given. Such differences were captured by computing the standard deviation of the image evaluation time, as shown in Figure 2. In order to filter out participants being distracted during the subjective test, all participants with a standard deviation of the image observation time larger than 20 s were rejected. This value was determined empirically and chosen to accommodate possible variations in download speeds of different users but reject users with significantly high variations in completion times.

Some other techniques for filtering out the unreliable users can rely on such metrics like time it took for a worker to complete the overall test and the mean time spent on each stimuli as discussed in [20]. Similarly, to the standard deviation in time per stimuli, the empirical thresholds can be found for these additional metrics and workers that do not fit into these thresholds can be filtered out.

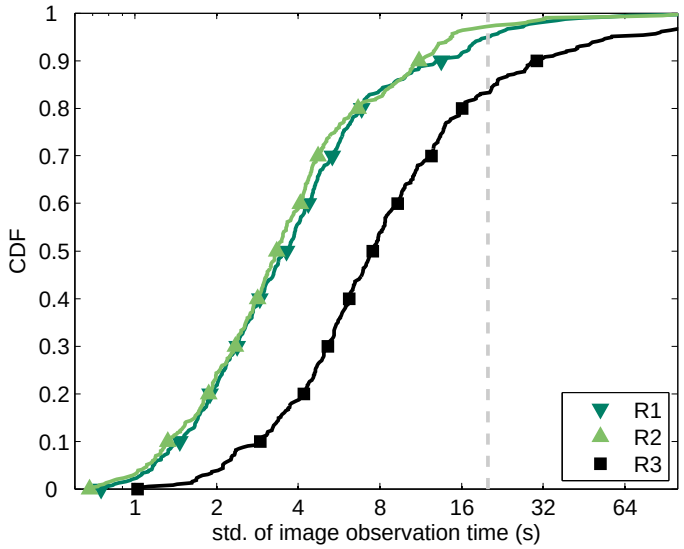


Figure 2: Cumulative distribution of the standard deviation of the image evaluation time in [13]. It can be seen that for several asian users (R3 curve) such variability in time taken to evaluate a single image becomes quite pronounced. In fact, some users were found to take up to 5 minutes to score a single image, where on average they took about three seconds to score the remaining images in the test. In [13], this procedure was found to exclude from the analysis about 13 % of the participants that completed the whole test.

9 Look a gift horse in the mouth!

The use of a crowdsourcing-based approach for testing the visual appeal of websites [14] resulted in a number of interesting lessons learned. Firstly, it became clear that crowdsourcing does provide a valuable mechanism for quickly and cheaply conducting these types of experiments while still obtaining meaningful results. In that sense, the results obtained are encouraging.

On the other hand, a number of issues were also noticed. Firstly, and contrary to possible expectations, an increase in payments will not necessarily lead to better results. In fact, in our results, it led to an increase in the number of unreliable users, most likely due to increased financial incentive to participate. This effect is illustrated by the table below where in the campaign shown the right column workers were paid three times more than in the campaign on the left. In the same time, the average ratio of reliable users is less in the more expensive campaign. Taking this into account, it is clear that additional incentives (e.g., gamification) and careful statistical analysis are required to avoid poor quality results.

Another apparent impact of the increased payments was the much faster completion of the test campaign. While this is in some cases desirable, it also results in a narrower variety of users in terms of demographics (due, for example, to the influence of time-zones). It might be worth taking this into account when proposing the campaigns, and possibly throttling their execution in order to obtain more representative population samples. The effects of time-zone differences also affects the reproducibility of the results, as it is hard, if not impossible to obtain similar

demographics distributions in different test runs.

In terms of the actual scores we notice that while payment level influences absolute MOS values for given assessment tests, it does not influence qualitative relations (i.e., main effects, interactions, shape of curves). Thus there does not appear to be a severe impact on models built from the campaign data (if such models exist). However, user ratings may have to be normalized to cope with the payment effect and to merge data from different studies (with different payments).

Table 1: Two identical crowdsourcing campaigns on web QoE asssement were conducted which only differ in the reward to the participants. Subjects completing campaign C_2 earned three times more money than the participants in campaign C_1.

Measure	C_1 with payment P_1	C_2 with payment $P_2 = 3P_1$
Number of countries	45	30
Ratio of completed tests	90.26 %	89.34 %
Campaign completion time	173.05 h	2.74 h
Avg. #correct content questions	8.27	7.48
Ratio of reliable users	71.54 %	66.10 %
Mean user rating	3.60	3.81

Since monetary incentive can have a negative effect on the reliability of the crowdsourcing results, it is possible to attract online workers through social networks. The disadvantage of using social networks is the limited access to the crowd and the fact that more effort is needed to pursue people from social networks to participate in a crowdsourcing experiment. Once example of using social network in the crowdsourcing is presented in [39], where Facebook users were used in the evaluation of different privacy filters applied on video.

A Facebook application was build and the call was to participate in the subjective test via such social networks like Facebook, Twitter, and LinkedIn, as well as various research mailing lists. With an estimated outreach to more than 1,500, some 120 among them used the application and submitted subjective scores. The resulted scores were compared (see figure below) with the results from a similar evaluation conducted by a conventional approach in a designated research test laboratory. The results demonstrated high correlation with only some minor differences favoring the crowdsourcing method, which means that it can be considered as a reliable and effective approach for subjective evaluation of visual privacy filters.

Note that no reliability checks or any filtering was applied on the workers from Facebook, while the results appeared to be highly reliable with high correlation to the lab-based evaluation. Such high reliability is due to the fact that Facebook users are generally verified individuals with very little number of them with fake IDs. Since the application was disseminated only to either friends, family, or friends of friends, it was in a way, propagated in a trusted way (since friends can be trusted). Such measures insured significantly more reliable results of the subjective tests, compared to a classical crowdsourcing scenario.

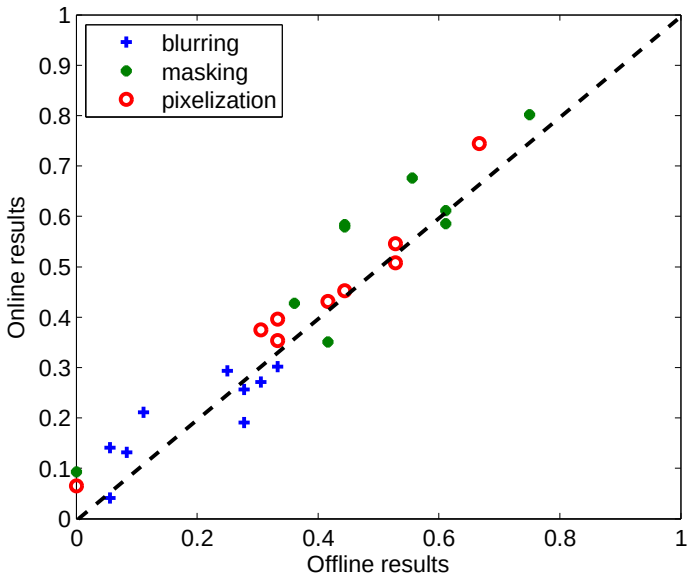


Figure 3: Privacy of Online vs. Offline Evaluations in Facebook-based crowdsourcing.

10 The crowd consists of human beings!

The question arises why we should care about ethics in the context of crowdsourcing.[1] We are actors in crowdsourcing systems to directly benefit from an environment that respects ethics. In general, long run altruism beats greed. An ethical system is more sustainable (everyone is empowered to reach full capacity). Also, according to most of the state laws, including European Directive 95/46/EC [47], governing processing of personal data, one has to be very careful about handling subjective information provided by the online workers in an ethical and privacy respecting way.

The next question is when we should care about ethics. The answer is simple always. In cases where crowdwork income is not essential: What are people not doing when they are crowdsourcing? Does crowdsourcing have an addictive side? In cases where people are subsisting on crowdwork income: Is the crowdwork paying only money, without building a longer term future? Is the money enough? Is the timing of the payment ok?

Followings are list of suggestion for taskmakers to take ethics into consideration:

General

- Try some crowdwork.
- Do not reveal identity of the crowdworkers. Abide the law on processing and storage of personal data: anonymize the workers, restrict access to any personal data, and the delete such data upon completion of the related research project.

[1] In the Dagstuhl seminar 13361 "Crowdsourcing: From Theory to Practice and Long-Term Perspectives", the ethical aspects in crowdsourcing were discussed in a special session which are summarized in [35]. Some results, suggestions and opinions from [35] are included in this section.

Job Design

- Use contracts to clearly define your relationships.
- Validate the functionality of the test implementation carefully, to make sure the worker can complete the task. Be available during the test by email.
- Dont engage in the race of bottom and pay appropriately for requested work. Be honest about the time your job takes, and have realistic expectations.
- Be very specific about your needs, use clear questions, and make the instruction with clear explanation.
- Avoid test designs that have too many complex details, keep them short.
- Tell to the workers what's required up front.
- In case of any sensitive issue or check you may require for your experiments, explain them to workers and allow them to opt out of the experiment early if they do not want to engage in the test.

Checking Responses

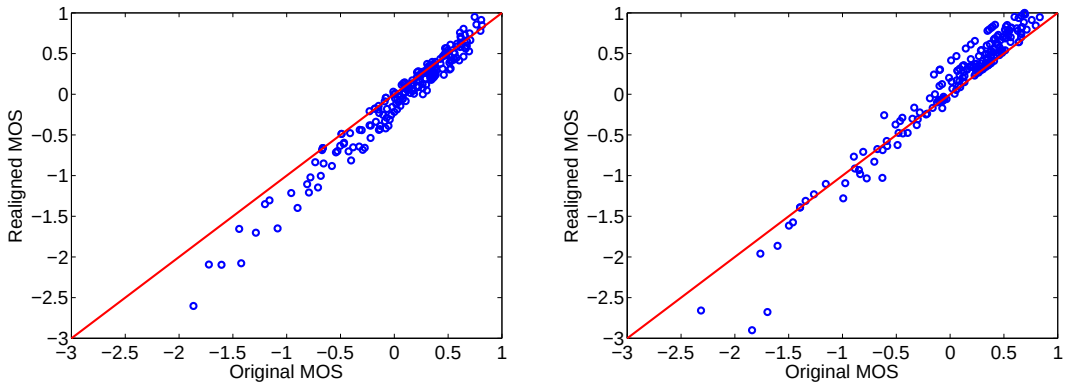
- Review work promptly. Realize there is an investment of time and not only money.
- If you are in doubt of paying the worker or not, pay the worker.
- Address crowdworkers politely (and try to apologize if you goof up).
- Respond to inquiries (even the ones with horrible grammar).
- Take a look at crowdworker's communities like turkopticon [48]. See their concerns and how do they rate you.

11 Lessons learnt from lab test. Use them in Crowdsourcing too!

Although performed out in the open, by thousands of different users across the world, crowdsourcing-based QoE test still remain psychometric experiments, which aim at quantifying user perceptions and preferences. To this purpose, there is a large body of literature that can be accessed to tackle the challenges that crowdsourcing-based testing poses. One such challenges is that of keeping the tests to a minimum duration (see also lesson 13). Whereas in lab-based experiments participants can take over one hour to perform the rating of a full stimulus set, in CS it is recommended to keep the duration much shorter, limiting the number of stimuli that can be evaluated by the same user. To be able to still obtain scores for a large set of stimuli, researchers typically segment the set of stimuli in subsets, to be scored by different users in different campaigns. This practice is prone to a major drawback: especially when using direct scaling (e.g. the most popular Single Stimulus, ACR stimulus scoring [44], users tend to use the whole scale for scoring, independent on the absolute quality of the stimuli visualized. In other words, if campaign A covers a large range of quality, and campaigncampaing B a smaller one, clustered towards the top of the quality range of A, it is quite likely that the worst stimuli in B, still being of much higher quality than the stimuli in A, will be scored equally low. This phenomenon is commonly known as "context" or "range" effect [49], and is usually countered by re-aligning MOS scores a posteriori. This can be done by means of the known scores of a subset of stimuli either kept constant (anchors)

throughout all campaigns, or collected from all campaigns and then re-evaluated in an extra campaign altogether [50].

In their aesthetic appeal assessment experiments, Redi et al. [13] adopted the method of using anchor stimuli common to all campaigns. They chose 5 anchor images so that they would space the entire quality range of the test stimuli, and so that they would span it in a uniform way (based on previous evaluations run in the lab), and they added them to each of the 13 campaign of 15 stimuli each making up their 200-image test set. This practice turned out to keep context effects to a minimum. When attempting at re-aligning the MOS of images of all campaigns to the scale of a reference campaign c*, authors of [13] found re-alignment almost useless. Figure 4 shows the aesthetic and recognizability scores obtained throughout the 13 campaigns of [13] against their realigned values with respect to campaign c*, selected as the one with stimuli spanning the widest quality range. As it can be seen, the effect of re-aligning is quite limited, indicating a strong robustness of the original MOS to context effects.



(a) Aesthetic quality **(b)** Recognizability

Figure 4: Original (from 13 different campaigns) and re-aligned (a) aesthetic quality and (b) recognizability MOS for experiment [13]. Range effects seem minimal in both cases.

Another consideration that is useful to make when designing crowdsourcing experiment is that, by dividing the traditional design over many users and campaigns, we intrinsically generate mixed-subjects designs. So, it may be the case that the measurements themselves loose in accuracy, because one cannot fully exploit within-subjects variance. An interesting solution to that has been recently proposed in [45], which involves randomized paired comparison [51] to accommodate incomplete and imbalanced data. Paired Comparison has been found to be an effective methodology for measuring QoE via crowdsourcing, due to the simplicity of the task and the availability of tools for the analysis of incomplete preference matrices. Furthermore, it allows easy embedding of worker reliability checks [45, 46] (see lesson 8). On the other hand, for scaling large sets of stimuli (in the order of hundreds), the applicability of paired comparison is still limited, as the number of pairs to be judged may be intractable also for such a far-outreaching methodology.

12 Use the appropriate scale for your problem!

A common issue in subjective measurement tasks (even beyond QoE) assessments are scale usage heterogeneity problems for absolute category rating (ACR) scales [52]. On the one hand users tend to avoid to use the both end of the scales, thus the votes tend to saturate before reaching the end points as shown in [40, 9], and on the other hand language and cultural differences regarding the distance between scale labels for a given ITU scale as reported in [53, 54] make it difficult to compare results across cultural or international boundaries. Typically, these problems are tackled by either ensuring that scale labels and designs are clearly understood in the respective language or culture (cf. ACR scale labels for different languages as described in [23]), [SE1] or by using extensive training sessions that educate the subjects regarding proper scale usage.

However, both of these solutions cannot be directly applied to crowdsourcing campaigns. Training sessions can be used (cf. lesson 3), but not as elaborate as in related laboratory trials due to the limited number of video sequences that can be used, in order not to lose crowd workers attention and ensure reliable results [8]. Another possible solution would be paired testing for eliminating offsets between different CS campaigns and laboratory tests as proposed by [46, 55]. Although this minimizes offsets between different test campaigns, it only provides relative ratings instead of absolute category ratings. This is useful for comparing different implementations of algorithms or codecs but provides less insight in the actually perceived quality of the customer. Therefore, industry and research are interested in absolute category rating (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 the customer's interaction with the company [52, 40].

In terms of language and scale design CS workers are quite heterogeneous regarding their native language and their cultural background. Therefore, they often receive instructions and scale descriptors (often) different not in 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 CS campaigns.

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

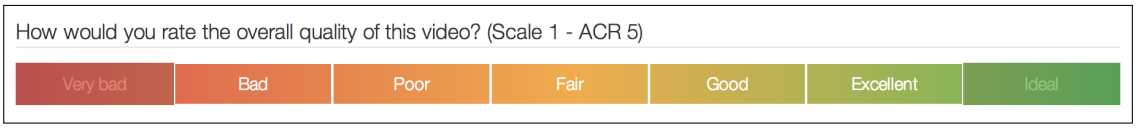


Figure 5: ACR-5 scale with non clickable anchor points and a traffic-light semaphore design. The scale designs is available under Creative Commons Attribution 3.0 Austria License at `https://github.com/St1c/ratings`.

13 Take the right duration for your experiment!

Participants in crowdsourcing are much less committed than normal in-lab participants and they can withdraw from works at any time. This fact depends from multiple elements. First of all the experimenter is a complete stranger for participants so they feel much less compelled to do something for him. Secondly a large number of other possible crowdsourcing works are present online, so if the ratio job difficulty / price paid is not high enough they simply give up and turn to another job. These elements do not reflect only on willingness to complete the whole test, but also in reliability as stated before: even if participants finish a long experiment they may end up working poorly or fastly just to finish as soon as possible.

As experiment should be kept as easy as possible, the other factors to be considered are experiment length and price paid. However, price paid is usually minimized considering other comparable jobs available online (and sometimes suggested/imposed by platforms while posting a job). It is then fundamental to shorten the experiment duration. Different works in literature underlined participants withdrawal and that a good rule of thumb is to keep duration under five minutes [9, 10, 13]. To achieve this, job can be split into multiple sessions to be assigned to different users. Moreover really committed participants can eventually participate multiple sessions.

Some online crowdsourcing platforms monitor participants behaviours, penalizing those who accept a work from which later withdraw. Mostly this metric appears in participants' statistics (i.e. Microworkers). However this metric alone is not enough to discourage withdrawals behaviour due to the large availability of small jobs easier to accomplish. Experiment duration must be correctly estimated and stated in jobs offers. Workers have the opportunity to report problems and a duration incompatible with declared job can be seen as an employer misconduct, leading to suspension of service from the platform.

14 Motivate your users!

Often requesters in crowdsourcing micro-task platforms assume workers to be motivated to complete tasks by small monetary rewards. However, previous studies show that workers are at least partly motivated by aspects other than monetary rewards such as killing time or having fun [56, 57].

In [12] Redi and Povoa showed how monetary reward can have a multifaceted impact on the quality of the data collected. They Set up a Facebook app, Phototo, for users to rate the aesthetic appeal of images processed with Instagram-like filters. Scores were expressed on a scale from 1 to 5 through a playful, star-based rating system. The App was launched through the Facebook networks of the experimenters. Users would then access it on a voluntary basis, out of curiosity or personal invitation from the experimenter, and without receiving any monetary rewards for performing the experimental task. Simultaneously, a Microworkers campaign was also launched, through which crowdsourcing workers could access Phototo and perform the experiment under a 0.30 $ compensation. Table 2 shows some striking differences between the behavior of users receiving a monetary compensation and volunteer users. A wide majority of the latter did not get to perform the experimental task (i.e., did not complete the training phase). About half of those did not even access the experiment introduction, possibly because they decided not to

grant Phototo the permission to access their personal data. Microworkers users, instead, were more than twice as likely to complete the first experimental task (67 %). On the other hand, when running a set of reliability checks to screen the remaining users, a higher percentage of paid users was ruled out, with respect to facebook volunteer users (see Tabl 2, last row). From this analysis, the authors of [12] conclude that paid users are more likely to commit to the execution of a crowdsourcing task; however, they may not perform it as reliably as volunteer users may do, driven by their intrinsic motivation. Furthermore, the authors found, as already noticed in [4] and lesson 9, a strong bias of paid users to rate image towards the top-end of the quality scale (see figure xx). This was less true for volunteer users.

Table 2: Two identical crowdsourcing campaigns on image quality were conducted with paid users (Microworkers) and volunteers (Facebook).

Measure	Facebook recruited participants (volunteers)	Microworkers recruited participants (paid)
No. registered users	414	258
Did not complete the training	63 %	31 %
Finished scoring the first set of images	32 %	67 %
No. users considered in the analysis	133	172
No. users considered after the filtering	96	113

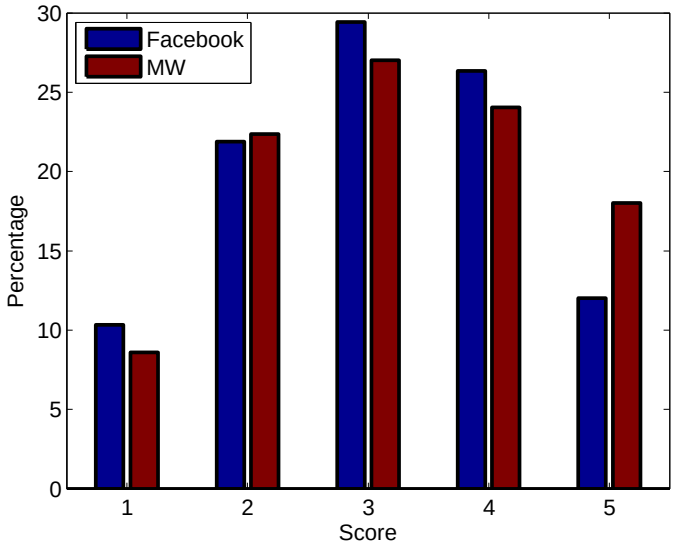


Figure 6: Strong bias of paid users (labeled 'MW') to rate images towards the top-end of the quality scale in contrast to voluntary users ('Facebook').

Naderi et al. [2] introduce a scale for measuring continuum of worker motivation from Amotivation to internic motivation based on Self-determination theory (SDT). The SDT distinguishes one's motivation based on its locus of causality and how far it is internalized by one meaning

"taking in" underlying regulation and the value of the activity. Moreover, SDT address the relation between work motivation i.e. degree of internalization and the quality of outcome. As a result, the more internalized motivation, the higher quality has the outcome.

We recommend to motivate workers by providing more insight about what they are doing and how important their performance and response quality are for you and for community.

Publications from Qualinet Crowdsourcing Task Force

[1] Andreas Sackl, Michael Seufert, and Tobias Hoßfeld. "Asking costs little? The impact of tasks in video QoE studies on user behavior and user ratings". In: *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. Vienna, Austria, Sept. 2013.

[2] Babak Naderi et al. "Asking costs little? The impact of tasks in video QoE studies on user behavior and user ratings". In: *3rd International ACM workshop on Crowdsourcing for Multimedia (CrowdMM 2014)*. Orlando, FL, USA, Nov. 2014.

[3] Babak Naderi et al. "Crowdee: Mobile Crowdsourcing Micro-Task Platform for Celebrating the Diversity of Languages". In: *Fifteenth Annual Conference of the International Speech Communication Association (Interspeech 2014)*. Singapore, Sept. 2014.

[4] Bruno Gardlo et al. "Microworkers vs. facebook: The impact of crowdsourcing platform choice on experimental results". In: *Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012)*. IEEE. Yarra Valley, Australia, July 2012.

[5] Bruno Gardlo, Michal Ries, and Tobias Hoßfeld. "Impact of screening technique on crowdsourcing QoE assessments". In: *22nd International Conference Radioelektronika*. IEEE. Brno, Czech Republic, Apr. 2012.

[6] Bruno Gardlo et al. "Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing". In: *IEEE International Conference on Communications (ICC 2014)*. IEEE. June 2014.

[7] Christian Keimel et al. "Qualitycrowda framework for crowd-based quality evaluation". In: *Picture Coding Symposium (PCS 2012)*. IEEE. Krakow, Poland, May 2012.

[8] Christian Keimel et al. "Video quality evaluation in the cloud". In: *19th International Packet Video Workshop (PV 2012)*. IEEE. Munich, Germany, May 2012.

[9] Christian Keimel, Julian Habigt, and Klaus Diepold. "Challenges in crowd-based video quality assessment". In: *Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012)*. IEEE. Yarra Valley, Australia, July 2012.

[10] Filippo Mazza, Matthieu Perreira Da Silva, and Patrick Le Callet. "Would you hire me? Selfie portrait images perception in a recruitment context." In: *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics. San Francisco, CA, USA, Feb. 2014.

[11] Isabelle Hupont et al. "Is affective crowdsourcing reliable?" In: *5th International Conference on Communications and Electronics (ICCE 2014)*. Da Nang, Vietnam, July 2014.

[12] Judith Redi and Isabel Povoa. “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, Nov. 2014.

[13] Judith Redi et al. “Crowdsourcing-based multimedia subjective evaluations: a case study on image recognizability and aesthetic appeal”. In: *2nd International ACM workshop on Crowdsourcing for Multimedia (CrowdMM 2013)*. Barcelona, Spain, Oct. 2013.

[14] Martin Varela et al. “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, Sept. 2013.

[15] Matthias Hirth et al. “Predicting Result Quality in Crowdsourcing Using Application Layer Monitoring”. In: *5th International Conference on Communications and Electronics (ICCE 2014)*. Da Nang, Vietnam, July 2014.

[16] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. “Analyzing Costs and Accuracy of Validation Mechanisms for Crowdsourcing Platforms”. In: *Mathematical and Computer Modelling* 57.11 (2013), pp. 2918–2932.

[17] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. “Anatomy of a crowdsourcing platform-using the example of microworkers. com”. In: *Fifth International Conference On Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS 2011)*. IEEE. Seoul, Korea, June 2011.

[18] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. “Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms”. In: *Fifth International Conference On Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS 2011)*. IEEE. Seoul, Korea, June 2011.

[19] Michael Seufert et al. ““To pool or not to pool”: A comparison of temporal pooling methods for HTTP adaptive video streaming”. In: *Fifth International Workshop on Quality of Multimedia Experience (QoMEX 2013)*. IEEE. Klagenfurt, Austria, July 2013.

[20] Pavel Korshunov et al. “The effect of HDR images on privacy: crowdsourcing evaluation”. In: *SPIE Photonics Europe 2014, Optics, Photonics and Digital Technologies for Multimedia Applications*. Brussels, Belgium, Apr. 2014.

[21] Philipp Amrehn et al. “Need for speed? On quality of experience for file storage services”. In: *4th International Workshop on Perceptual Quality of Systems (PQS 2013)*. Vienna, Austria, Sept. 2013.

[22] Phuoc Tran-Gia et al. “Crowdsourcing and its Impact on Future Internet Usage”. In: *it-Information Technology Methoden und innovative Anwendungen der Informatik und Informationstechnik* 55.4 (2013), pp. 139–145.

[23] Simon Oechsner et al. “Visions and Challenges for Sensor Network Collaboration in the Cloud”. In: *Eigth International Conference On Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS 2014)*. IEEE. Birmingham, UK, July 2011.

[24] Tobias Hoßfeld et al. “Best Practices for QoE Crowdtesting: QoE Assessment with Crowdsourcing”. In: *IEEE Transactions on Multimedia* 16.2 (Feb. 2014), pp. 541–558.

[25] Tobias Hoßfeld and Christian Keimel. “Crowdsourcing in QoE Evaluation”. In: *Quality of Experience: Advanced Concepts, Applications and Methods*. Ed. by Alexander Raake Sebastian Mller. Springer: T-Labs Series in Telecommunication Services, ISBN 978-3-319-02680-0, Mar. 2014.

[26] Tobias Hoßfeld and Christian Timmerer. “Quality of Experience Assessment using Crowdsourcing ”. In: *IEEE COMSOC MMTC R-Letter* 5 (June 2014).

[27] Tobias Hoßfeld et al. “Survey of Web-based Crowdsourcing Frameworks for Subjective Quality Assessment”. In: *16th International Workshop on Multimedia Signal Processing (MMSP 2014)*. Jakarta, Indonesia, Sept. 2014.

[28] Tobias Hoßfeld, Matthias Hirth, and Phuoc Tran-Gia. “Aktuelles Schlagwort: Crowdsourcing”. In: *Informatik Spektrum* 35 (Apr. 2012).

[29] Tobias Hoßfeld, Matthias Hirth, and Phuoc Tran-Gia. “Modeling of crowdsourcing platforms and granularity of work organization in future internet”. In: *23rd International Teletraffic Congress (ITC 2011)*. International Teletraffic Congress. San Francisco, CA, USA, Sept. 2011.

[30] Tobias Hoßfeld et al. “Assessing Effect Sizes of Influence Factors Towards a QoE Model for HTTP Adaptive Streaming”. In: *6th International Workshop on Quality of Multimedia Experience (QoMEX 2014)*. Singapore, Sept. 2014.

[31] Tobias Hoßfeld et al. “Quantification of YouTube QoE via crowdsourcing”. In: *IEEE International Symposium on Multimedia (ISM 2011)*. IEEE. Dana Point, CA, USA, Dec. 2011.

[32] Tobias Hoßfeld et al. “Initial delay vs. interruptions: between the devil and the deep blue sea”. In: *Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012)*. IEEE. Yarra Valley, Australia, July 2012.

[33] Tobias Hoßfeld. “On Training the Crowd for Subjective Quality Studies”. In: *VQEG eLetter* 1 (Mar. 2014).

[34] Valentin Burger et al. “Increasing the Coverage of Vantage Points in Distributed Active Network Measurements by Crowdsourcing”. In: *Measurement, Modelling, and Evaluation of Computing Systems and Dependability and Fault Tolerance (MMB & DFT 2014)*. Bamberg, Germany: Springer, Mar. 2014.

[35] Tobias Hoßfeld, Phuoc Tran-Gia, and Maja Vucovic. “Crowdsourcing: From Theory to Practice and Long-Term Perspectives (Dagstuhl Seminar 13361)”. In: *Dagstuhl Reports* 3.9 (2013). Ed. by Tobias Hoßfeld, Phuoc Tran-Gia, and Maja Vukovic, pp. 1–33. ISSN: 2192-5283. DOI: http://dx.doi.org/10.4230/DagRep.3.9.1.

[36] Martin Rerabek et al. “Evaluation of privacy in high dynamic range video sequences”. In: *SPIE Optical Engineering + Applications*. International Society for Optics and Photonics. San Diego, CA, USA, Aug. 2014.

[37] Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi. “Crowdsourcing evaluation of high dynamic range image compression”. In: *SPIE Optical Engineering + Applications*. International Society for Optics and Photonics. San Diego, CA, USA, Aug. 2014.

[38] Philippe Hanhart, Pavel Korshunov, and Touradj Ebrahimi. "Crowd-based quality assessment of multiview video plus depth coding". In: *IEEE International Conference on Image Processing*. Paris, France, Oct. 2014.

[39] Pavel Korshunov, Shuting Cai, and Touradj Ebrahimi. "Crowdsourcing approach for evaluation of privacy filters in video surveillance". In: *1st International ACM workshop on Crowdsourcing for Multimedia (CrowdMM 2012)*. ACM. Nara, Japan, Oct. 2012.

[40] Bruno Gardlo. "Quality of experience evaluation methodology via crowdsourcing". PhD thesis. Zilina, Slovakia: University of Zilina, 2012.

Other References

[41] Omar Alonso, Daniel E Rose, and Benjamin Stewart. "Crowdsourcing for relevance evaluation". In: *ACM SigIR Forum*. Vol. 42. 2. ACM. 2008, pp. 9–15.

[42] Julie S Downs et al. "Are your participants gaming the system?: screening mechanical turk workers". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2010, pp. 2399–2402.

[43] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. "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*. Association for Computational Linguistics. 2009, pp. 27–35.

[44] BT.500-13 (01/2012). "Methodology for the subjective assessment of the quality of television pictures". In: *Recommendation ITU-R* (2012).

[45] Qianqian Xu et al. "HodgeRank on random graphs for subjective video quality assessment". In: *IEEE Transactions on Multimedia* 14.3 (2012), pp. 844–857.

[46] Kuan-Ta Chen et al. "A crowdsourceable QoE evaluation framework for multimedia content". In: *17th ACM International Conference on Multimedia*. ACM. Beijing, China, Oct. 2009.

[47] European Parliament Directive. "95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data". In: *Official Journal of the EC* 23.6 (1995).

[48] Lilly C Irani and M Silberman. "Turkopticon: Interrupting worker invisibility in amazon mechanical turk". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. Paris, France, Apr. 2013.

[49] Huib de Ridder. "Cognitive issues in image quality measurement". In: *Journal of Electronic Imaging* 10.1 (2001), pp. 47–55.

[50] Yohann Pitrey et al. "Aligning subjective tests using a low cost common set". In: *QoE for Multimedia Content Sharing* (2011).

[51] Alexander Eichhorn, Pengpeng Ni, and Ragnhild Eg. "Randomised pair comparison: an economic and robust method for audiovisual quality assessment". In: *20th international workshop on Network and operating systems support for digital audio and video*. ACM. Amsterdam, The Netherlands, June 2010.

[52] Peter E Rossi, Zvi Gilula, and Greg M Allenby. "Overcoming scale usage heterogeneity: A Bayesian hierarchical approach". In: *Journal of the American Statistical Association* 96.453 (2001), pp. 20–31.

[53] Bronwen L Jones and Pamela R McManus. "Graphic scaling of qualitative terms". In: *SMPTE journal* 95.11 (1986), pp. 1166–1171.

[54] MT Virtanen, N Gleiss, and M Goldstein. "On the use of evaluative category scales in telecommunications". In: *Human Factors in Telecommunications*. Melbourne, Australia, Mar. 1995.

[55] Chen-Chi Wu et al. "Crowdsourcing multimedia QoE evaluation: A trusted framework". In: *IEEE transactions on multimedia* 15.5 (2013), pp. 1121–1137.

[56] Judd Antin and Aaron Shaw. "Social desirability bias and self-reports of motivation: a study of amazon mechanical turk in the US and India". In: *ACM SIGCHI Conference on Human Factors in Computing Systems*. ACM. Austin, TX, USA, 2012.

[57] Panagiotis G Ipeirotis. *Demographics of mechanical turk*. Tech. rep. http://hdl.handle.net/2451/29585. New York University, Mar. 2010.