­

|  |  |  |  |
| --- | --- | --- | --- |
| **Question(s):** |  | **Meeting, date:** |  |
| **Study Group:** |  | **Working Party:** |  | **Intended type of document** (R-C-TD): |  |
| **Source:** | USA |
| **Title:**  | Using Crowdsourcing As A Tool to Perform Subjective Image Quality Assessments  |
| **Author(s) of the Contribution:**

|  |  |
| --- | --- |
| **Name**  | **Email** |
| **Hanan Alnizami, Intel Corporation, USA** | **hanan.alnizami@intel.com** |
| **Philip J Corriveau, Intel Corporation, USA** | **philip.j.corriveau@intel.com** |
| **Patrick E. McKnight, George Mason University** | **pmcknigh@gmu.edu** |
| **Jake Quartuccio, George Mason University** | **jquartuc@gmu.edu** |
| **David Nicholas, Intel Corporation, USA** | **david.g.nicholas@intel.com** |
| **Michele Saad, Intel Corporation, USA** | **michele.a.saad@intel.com** |

Please do not change the structure of this table, just insert the necessary information. |

**Recommendation ITU-T <No.>**

**Using Crowdsourcing As A Tool to Perform Subjective Image Quality Assessments**

# Summary

This Recommendation describes a subjective methodology for assessing image quality using crowdsourcing. The proposed method enables experimenters to collect a large number of image quality ratings in a short period of time. It saves the experimenter the time and effort it takes to recruit participants and run a study. This recommendation outlines a unique study design that includes source stimuli and image sets, overlapping image sets, and validity checks, to ensure data integrity.

# Keywords

Crowdsourcing, Image Quality Assessments, Image sets,

# Scope

This Recommendation describes a subjective methodology for assessing image quality using crowdsourcing. Crowdsourcing taps into the collective intelligence of the public at large to complete image quality tasks that would normally either be performed by image quality experts themselves or outsourced to a third-party provider to perform the assessment.

Standardized laboratory methods were previously executed to gather reliable and accurate perceived image quality ratings from a number of recruited participants in a lengthy, expensive, and controlled experimental setups. The crowdsourcing approach enables experimenters to access a larger pools of participants in order to collect participants’ feedback on topics presented to them via innovative experimental design approaches that improve the quality of the data collected.

# References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-R BT.500] Recommendation ITU-R BT.500 (2012), *Methodology for the subjective assessment of the quality of television pictures*.

[ITU-T P.910] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.

[ITU-T J.140] Recommendation ITU-T J.140 (1998), *Subjective picture quality assessment for digital cable television systems*.

[ITU-R BT.710] Recommendation ITU-R BT.710 (2011), S*ubjective assessment methods for image quality in high-definition television*.

[ITU-R BT.1129] Recommendation ITU-R BT.1129 (2011), *subjective assessment of standard definition digital television (SDTV) systems*.

[ITU-R BT.802] Recommendation ITU-R BT.802 (1994), *Test pictures and sequences for subjective assessments of digital codecs conveying signals produced according to Recommendation ITU-R BT.601*.

[ITU-T BT.1210] Recommendation ITU-T BT.1210 (2004), *Test materials to be used in subjective assessment.*

# Terms and definitions

This Recommendation defines the following terms:

1. **Subjective assessment** [ITU-T J.144]: The determination of the quality or impairment of program-like pictures presented to a panel of human assessors in viewing sessions.
2. **Crowdsourcing** [ITU-T P.912]: Obtaining the needed service by a large group of people, most probably an on-line community.
3. **Workers** [ITU-T P.912]: Members of the crowdsourcing environment signed up to participate in a crowdsourcing test, also referred to as participants throughout this document (Amazon Mechanical Turk refers to participants as workers).
4. **Masters:** Amazon Mechanical Turk refers to masters as members of an elite workers group in crowdsourcing that demonstrate exceptional accuracy when performing a test. A member of such group should continue to demonstrate accurate excellence to maintain this title.
5. **Requester:** The experimenter who submits a test to the crowdsourcing platform seeking workers’ input.
6. **Test** [ITU-T P.912]:The subjective assessment study that a worker is asked to perform in a crowdsourcing environment.
7. **Task** [ITU-T P.912]: Set of actions assigned to a worker to perform in order to complete a test.
8. **Stimuli Set:** A collection images extracted from the original set of test images, treated as its own study set.
9. **Overlapping stimuli sets:** A number of stimuli that range in quality from best to worst, taken from the original set of stimuli studied, used as a control set and applied to every image set.
10. **Stimulus:** An image that users are asked to rate in a test.

# Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

IQA Image Quality Assessment

# Introduction

Image quality assessments (IQA) rely fundamentally on subjective studies in order to capture humans’ perception of image quality. In the past, a number of standardized laboratory studies were conducted simultaneously to evaluate image and video quality, employed via stringent procedures (such as screening for visual acuity and color deficiencies) and in controlled environments (such as the tuning of ambient lighting, noise, chair comfort, viewing distances…etc.). While these methods were proven to collect reliable and valid data, these procedures are expensive and lengthy for both the experimenter and the participant.

Previous literature showed that controlled laboratory conditions are not needed to obtain reproducible subjective ratings; experimenters may capture equivalent data from more flexible experimental designs using online samples. The more flexible approach includes crowdsourcing methods that became popular since the advent of online data collection in the mid 1990’s. A potential downside to his approach, however, is the diminished experimental control in both sampling and study implementation. The benefits outweigh the costs - at least with respect to generalizability. Crowdsourcing experiments provide experimenters access to difficult to recruit participants in a wider array of settings (e.g., image displays, ambient lighting, and participant motivation to name a few). These varied settings allow researchers to test image quality ratings across varying observed conditions and to determine the conditions that may most affect ratings outside the laboratory setting. All these benefits may come at a substantial costs if the crowdsourcing studies fail to abide by some key principles known to affect the results. We outline those principles below.

Crowdsourcing broadens the scope of studies and allows investigating various user-related and context-related influence factor due to the large pool of subjects and realistic test conditions. However, crowdsourcing faces several challenges, e.g. conceptual challenges in the test design, unreliability of users, but also incentives and payment schemes to motivate users, hidden influence factors in the uncontrolled environment, and statistical analysis of the results.

This standard considers crowdsourcing for quality assessments in general. To make the recommendations clearer, as an example image quality assessment is used throughout this document.

# Test Methods and experimental design­

In order to successfully design a crowdsourcing study, experimenters need to focus on the hypothesis depicted by their research to achieve the desired results. Experimenters are out to consider the following factors as dictated by their experimental design, as the design choices made will highly influence the accuracy and reliability of the data collected:

## Sampling subjects and compensation

Crowdsourcing environments contain a massive pool of participant workers who are willing to partake in tasks submitted by requesters for a compensation. Crowdsourcing platforms commonly have controls for screening participants based upon their geographical location and researcher ratings. The experimenters need to keep careful documentation of all decisions made in participant recruitment, and report results in light of those decisions.

### Screening before a test

In order to target the right audience, crowdsourcing platforms allow requesters the ability to screen for certain worker demographics, socio-economic background, usage characteristics, the physical and mental constitution by submitting study eligibility criteria to the platform (i.e. age, gender, geo locations, occupation…etc.). Workers are requested to disclose demographic information when creating an account, making it possible to screen them against tests’ eligibility criteria. This will expose the right pool of participants to tests they are eligible to partake in.

### Screening when signing up to participate in a test

In addition to the previous screening method, crowdsourcing environments allow requesters to use additional screening methods to ensure workers eligibility:

1. Requesters can begin the test with questions that pertain to exclusion criteria. For example, if interested in recruiting participants who aren’t color blind, a requester could ask participants if they are color blind, then decide upon their answer whether they can participate in the test. In a situation where participants are not eligible but have answered the eligibility questions, requesters must inform participants that they have not met the eligibility criteria and must compensate workers for the time they took to answer the eligibility questions and then excuse them from the study.
2. Crowdsourcing platforms with APIs allow researchers to write software that collects information about the user to determine eligibility. For example, if an experimenter is interested in workers who are viewing tests in a specific browser or ones who are using a specific operating system, or specific device form factors, researchers can write a web program on their own servers to request the information with common programming languages. In addition, researchers can use scripts to record the ID that the crowdsourcing platform assigns to the participant. Researchers can use this information to determine participants’ eligibility for future studies. In the case that the worker is not eligible to participate based upon the answer to a question, researchers can either screen participants during the data analysis or inform participants that they have not met the eligibility criteria while compensating them for the time they took to answer the eligibility questions and excusing them from the study.

## Compensation and Incentives

Similar to compensations given to participants in a laboratory setting, compensation in a crowdsourcing environment is determined by the requester. In crowdsourcing, compensation is used as a way to tempt participants to partake in a test and the value must be relevant to the amount of time it takes the worker to complete a test. The greatest benefit to paying participants more is that the data collection happens quicker. As other research has shown, there does not seem to be a quality increase with higher amounts of pay. If payments are high enough, it only takes a few hours to get a few hundred participants. One potential limitation to high pay rates is the allure to “game” the system.

Although the amount of payment may influence speed of test, it might also cause bias by higher user ratings [‎1, ‎2]. An explicit test instruction, to make the worker clear that is not necessary to “please” the employer, will overcome this issue.

We do not know the optimal pay but experienced higher than normal random responses when the pay rate exceeded the usual rates for most crowdsourcing studies.

## Handling human subjects

Similar to regulations of handling human subjects in a laboratory study, experimenters must be aware of one’s own country and institutional regulation. In the United States, some of these regulations include:

1. Obtaining informed consent from participants
2. Maintaining data confidentiality
3. Participants’ right to opt out of the study at any time

Experimenters must obtain research approval and must abide by country laws for handling human subjects online. In crowdsourcing, an experimenter must provide information that describes the study, the potential risks and benefits, and compensation. Consent can be obtained actively by requiring the worker to agree to the terms and conditions of the study and to click a button to navigate to the next page of the test.

If possible, it is also good to not ask for email addresses or other social media identifiers as it undermines participant confidentiality. Most crowdsourcing platforms have a mechanism for communicating with participants.

## Test length

*The length of the test strongly correlates to several parameters, namely the*

* *overall enjoyability of the whole experiment,*
* *complexity of the task,*
* *user interface,*
* *amount of reward,*
* *user’s ability to understand the task.*

*Hence, the overall length of the experiment is a combination of the parameters above. Generally, the more the user enjoys the task, the longer the test can last. This is strongly related to the key issue of incentives and payment schemes.*

On the other hand, similar to laboratory testing, workers should not be exposed to lengthy tests that induce boredom, fatigue, and disinterest. Short durations will favor engagement of the workers with the tasks, thereby favoring reliable executions and commitment. Hour-long crowd-based experiments would most likely result in poorly reliable executions and, therefore, outcomes.

An experimenter should expect data accuracy to decrease with lengthy tests. Ideally, tests ought to be short - 5-10 minutes. Participants are much less likely to correctly answer a validity check at the end of the experiment than at the beginning. Since the enrollment for online crowdsourcing experiments is fast, it is ideal to launch a series of smaller studies. The studies could be launched over a series of hours or days. An experimenter can adopt different techniques to design their experiment such that test length is minimized.

## Pilot testing

When designing and executing research, experimenters must perform pilot testing to ensure that the experimental design is answering the questions posed by the researcher. As such, experimenters must pilot their crowdsourcing study with 20-30 workers first to get an insight about the appropriate study length, as well as data accuracy and reliability. If the accuracy of data decreases with time, experimenters must lessen the number of images rated per image set, shortening the length of the test.

### Source Stimuli

Content chosen for evaluating image quality via crowdsourcing must cover a wide range of stimuli. Experimenters are encouraged to include a large number of images to test with in order to minimize effects of any confounding variables. This large collection of images must then be divided into smaller image sets in order to reduce fatigue and boredom. Each smaller image set is treated as its own study, reducing the duration of the test and maximizing reliability and quality of data captured. The stimuli must be selected in accordance with to the goal of the test and the hypothesis established by the experimenters. The deciding factor to the number of images displayed per test must be decided upon based on pilot testing and pilot data analysis to determine data accuracy. Also, the short study design lessens the opportunities for technical failures, such as internet connectivity. It is recommended that the number of images per test not exceed anywhere between 50-70 images in an IQA study, including validity checks and the overlapping image set. However, a pilot is needed to ensure the accurate number of images per test.

## Rating scales

Experimenters must consider their experimental design and research method when deciding which testing method and rating scale to use. The list of standardized subjective test methods and rating scales has been provided in ‎[8] among which, the following two rating methods are recommended.

1. Absolute category rating (ACR) method; also known as the single stimuli method. This method is a method where a single stimulus is presented to the worker to collect a user rating. Stimuli are rated independently from each other. The commonly used scale is a five-points rating scale:

 5 Excellent

 4 Good

 3 Fair

 2 Poor

 1 Bad

1. Degradation category rating (DCR) method; also known as the double stimulus impairment scale (DSIS) method. This method is one where a pair of stimuli is presented to the worker at once; one being the reference image, and the other being the test image. Workers are asked to rate impairments of an image in relation to the other using the following scale five-point scale:

 5 Imperceptible

 4 Perceptible but not annoying

 3 Slightly annoying

 2 Annoying

 1 Very annoying
The other test methods, rating scales and allowed changes addressed in clause 7 of [‎8] are also appropriate to use in crowdsourcing tests.

## Anchors

When having a single stimulus test setup such as in ACR method [‎8], users are entailed to visualize the stimulus once and quantify its quality/level of impairment on a discrete scale, along which qualitative labels (adjectives) are reported. Participants are required to indicate which of these five adjectives better expresses the quality level of the stimulus.
Although such a direct scaling fits perfectly many of the requirements of crowdtesting (ease of implementation, task simplicity and fast completion), it is important that the test designer takes into account one of its major drawbacks: the risk of returning scores suffering from context effects [‎8, ‎9]. 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. Considering a stimuli 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 MOS close to ’Bad’ (although in reality it is not as bad as other stimulus in the first set, with a true quality value < A/2). The re-alignment of stimulus is a solution to this issue [‎10], though, is time consuming.

In order to overcome context effects derived from the fragmentation of QoE evaluations in crowdsourcing, it is recommended to introduce in each evaluation campaign a small number of stimuli, kept equal for all sub-tasks, spanning a wide range of quality [‎13, ‎14]. These stimuli, named Anchors, have the purpose of limiting context effects by fixing the extreme values of aesthetic appeal to be seen in each sub-task. 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, e.g., from a small pilot study).

Limiting the context effects using anchors was shown in [‎14]. 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, added to each campaign 5 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 lab experiment (see Figure 1). In analyzing the data, the authors performed a re-alignment of the image MOS across campaigns, which was concluded to be unnecessary (i.e. MOS and their ordering would not change significantly after realignment), thereby proving the effectiveness of the anchors.



Figure 1 Anchors used in the crowdsourcing-based image aesthetic quality assessments reported in ‎[14]

Another approach to overcome the context effects when direct scaling is to use comparison rating procedures instead [‎11, ‎12].

## Overlapping warm-up stimuli set

Since not every participant will see all stimuli, a common set of 10 stimuli must be included in all series of studies. This common set assists experimenters with testing between subject stability and replicability of within subject ratings over time. Overlapping stimuli are chosen from the original pool of stimuli, ranging in quality from low to high quality. Additionally, this set should be representative of the contents of the stimuli in the whole set. For example, if in a test image set, some photos show balls or cars, ensure that the warm up images also show those images. This can also be applied to the lighting or other features of the stimuli. These images serve as a training set placed in a random order in the beginning of each image set, and ratings for these images should be discarded from the analysis. However, the same images should also be incorporated randomly in each image set as a test image as well.

## Test design

Because of the nature of a crowdsourcing study, a within subject study design is recommended with a large number of test stimuli divided into smaller sets that become their own tests. Workers must be allowed to participate in as many tests as they desire because experimenters must aim to approximate as much of a fully-crossed study as possible.

The experimenter must present test images to workers such that images are shown in full screen mode to ensure clarity of image shown especially if the test aims to discover perception of subtle changes in an image.

Experimenters must be aware of the following practices that help strengthen their test design, yielding more reliable data collected from workers in a crowdsourcing study:

1. The use of clear instructions. Unlike in a laboratory setting, an experimenter does not directly interact with participants. As a result, instructions given to participants must be specific and unambiguous, leaving no chance for misinterpretation.
2. Allowing email communication between workers and experimenters to discover any encountered problems. This holds great importance as it provides the experimenter with valuable information during the analysis. For example, in an IQA, a worker could contact the experimenter to inform them of any technical difficulties faced while performing the test, such as the content not loading to rate, causing the worker to give low ratings.

###  Experiment description for participants

Clearly state eligibility requirements, give a time estimate on how long it takes to complete the experiment, and describe which type of responses may result in rejected work, if any. As mentioned before, it is very important to pilot the online experiment in house to give a fair time estimate to the participants. A time estimate helps participants to determine if an experiment pays enough. Another consideration is to identify your research group. While remaining anonymous may help to protect the reputation of the research lab, stating affiliations helps to build trust with the crowdsourcing community.

## Validity checks

Experimenters are encouraged to incorporate validity checks randomly into each image set. These serve as attention checks to ensure workers full attentiveness to the task at hand. It is recommended to either 1) use open ended questions where a user is meant to answer by typing instead of selecting an answer from multiple choices, or 2) instruct the user to select a specific answer from a multiple choice question. Workers who fail one or more of the validity check questions must be removed from the analysis as the data collected is identified as invalid. However, we suggest compensating regardless of accuracy for the sake of time efficiency. Participants who do not get compensated may email the researchers and it can be time consuming to arrange for payments outside of the crowdsourcing platform.
When analyzing crowdsourced data, the greatest concern is the data validity. There are a few different reasons why data may not be valid.

1. Participants did not pay attention to the stimuli. In other words, they responded “carelessly” to the stimuli.
2. Participants have a variety of personal devices from which they view the screen. Therefore, the images might have seen a decrement in quality

Verification tests [‎5, ‎6] help in identifying automatization in the form of scripts, but can also be an indicator for sloppy workers and random clickers. 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, so-called gold standard data [91] can be utilized: 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, also the general interactions of the user with the task interface can be monitored, e.g., to measure the focus time of a video clip, i.e. time interval during which browser focus is on the website belonging to user test, or the time it takes the users to answer questions. In order to increase the number of valid results, a warning message can be displayed if the worker, e.g. did not watch more than 70% of the video, so that the user can decide to watch the video again or to continue the test. When workers become aware of this control mechanism, the percentage of completely watched videos, thus the number of workers considered as reliable will be increased.

Based on preliminary tests about how trustworthy users behave, an additional reliability score is derived. These reliability tests may either be employed a-posteriori after the test or alternatively already during the test. This in momento reliability checking [‎7] also allows to identify reliable workers during the test, which allows to engage reliable users with more tasks directly in the current test.

One way to screen for participants who are not paying attention is to remove them if they have failed attention checks. The merit to this approach is that you may reduce the amount of noise in your data. The downside is that you may be removing participants based on a trait of low consciousness or ability to pay attention. It is important for researchers to set up a criteria before the study to decide the threshold for removing participant’s data, if they even decide to remove data from careless responders. Sometimes researchers chose to not pay participants who fail attention checks. Other times, researchers pay participants, but remove their data from the analysis.

**Statistical analysis of the crowdsourcing results**

To evaluate the crowdsourcing results, first we need to identify the reliability of the user ratings [‎4]. This can be done by considering hidden influence factors in addition to the variability of subjects’ sensitivities to different artefacts. In previous sections, several screening techniques to identify unreliable users were presented. In addition, reliability measures such as inter- and intra-rater reliability should be also stated for QoE crowdtesting studies, where high values show reliable user ratings, but low values imply the presence of unreliable users or hidden influence factors in the QoE crowdtesting campaign. More detailed approaches and examples for statistical analysis of the test results, independent of any hidden influence factor and the actual user rating are presented in appendix I.

**Data Analysis**

When first establishing a crowdsource study’s software and methodology, correlate the data with lab “ground truth” data and visualize the bivariate relationship. Ideally, there will be a rectilinear relationship between the controlled lab data and the crowdsourced data. If the relationship is not rectilinear, it may mean that the correlation between lab and crowdsourced data changes across either axis.

Before analyzing a completed crowdsourced dataset, decide how to remove “careless responders”. Does one careless response or multiple careless responses make the participant eligible for removal from the data set? Also, evaluate the types of devices that participants used in the study. Is the screen size and device type adequate for comparisons? If the same participants evaluated multiple image sets, it adds to the strength of the study so report the total number of unique participants. Finally, consider evaluating the “warm up” images to see if participants are responding similar to all of them.

Other than the points listed above, continue to analyze crowdsourced data the same way as a normal dataset.

**Annex A

<Annex Title>**

(This annex forms an integral part of this Recommendation.)

<Body of annex A>

**Appendix I

<Appendix Title>**

(This appendix does not form an integral part of this Recommendation.)

<Body of appendix I>

### <Monitoring hidden influence factors>

Hidden influence factors may also cause the test to fail if they are dominating, e.g. screens of users do not allow to discriminate certain test stimuli. Less severe hidden influence factors may cause higher variances of test results, which means that more users are required to obtain the same statistical significance than in the lab. In the following, we highlight some examples for best practices to cope with the unknown context and the resulting hidden influence factors.

1. Monitoring of the Workers’ Environmental Conditions and Context: The environment in which the workers evaluate the stimuli in QoE crowdtesting may impact the overall QoE.
	1. The general viewing conditions represented by the background illumination or the screen resolution itself 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, a basic test pattern similar to the test patterns used for calibration of the monitor contrast and illumination in a professional environment can be utilized to quantify the user's viewing conditions, for example by asking how many grey steps on a grayscale step-wedge are visible. Moreover, such patterns can also be used to instruct workers how to calibrate their display.
	2. To prevent an undesirable context from the technical perspective, for example for video QoE assessment, the test videos should be pre-loaded in the remote browser, so that additional distortions introduced by the transmission do not affect the playback.
2. Expectation of Users: A hidden influence factor on the user level can be the users’ expectations: those used to lower quality will rate differently than those typically consuming higher quality. The expectation level may be closely related to the country of the subject and the typical quality of the provided content there. There are two options to cope with expectations: either quantifying the degree of expectations or reducing the expectations by instructing the test user accordingly. One option to quantify the expectations is to group users according to their expectations by asking them about their habits and typical use of a service, for example, “How often do you watch Internet videos?” and “Do you watch low or high resolution in YouTube?”, respectively, where the assumption is that subjects who do not use video streaming services often may be more tolerant to worse quality.

In the rating task, a user may additionally be asked to rate on an extra expectation category scale that is better aligned with the actual user’s expectations. The subjects then rate the quality with, for example, five levels of expectations: (-2) Much worse than I expected, (-1) Worse than I expected, (0) Just as I expected, (1) Better than I expected, (2) Much better than I expected. This rating scale is accompanied with a question regarding the perceived quality, e.g., “Please indicate to which degree the overall quality of this video was in line or not in line with your expectations? The overall video quality was...”. Still, the quantification of expectation remains a topic for future work.

1. Demographics and User Impairments: There are several options for measuring demographic data that may have an impact on the QoE results and should therefore be statistically analyzed.
	1. Surveys, although the user may not give correct answers.
	2. Extraction of data from social networks, but information could be also not reliable.
	3. Consistency tests to derive relevant information, but only a subset of data can be retrieved in order to avoid overusing consistency questions about demographics.
	4. Get the information from crowdsourcing platform, if available.

Furthermore, hidden influence factors on the QoE results may be caused by physical impairments of the subjects themselves. For visual stimuli, for example, a test for color blindness may be necessary to confirm normal color vision.

1. Hard-and Software Environment: QoE crowdtesting are subjective tests conducted in a heterogeneous and therefore partly uncontrolled environment. Thus, monitoring on the system level is required to analyze hidden influence factors on a system level. Due to bottlenecks at the end user devices in terms of CPU, memory, or network bandwidth, additional artefacts may arise and affect the user rating accordingly. For example, the user’s Internet access bandwidth may not be large enough to conduct a video quality test without stalling. However, those stalling events and the corresponding freezing of the video will impact the QoE. To overcome the impact of the network delay due to Internet delivery of data, the test application and data may be completely downloaded before the actual user test starts. Even 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 level is required. As a possible solution, download speed and latency may also be measured before the actual test, and then only users are selected with suitable connection speed and latency.

**<Recommendation: Two-Stage QoE Crowdsourcing Design>**

In the following, a general recommendation including two stages for crowdsourcing campaign is presented (Figure 2).



Figure 2 A recommendation for QoE crowdtesting design

The first stage represents a very simple and easy to do task, which:

1. tests the reliability of the users,
2. gathers a huge panel of users,
3. gathers information about the users in the crowd,
4. has a duration of less than a minute and has low pay ($0.10),
5. can perform context monitoring: hardware or software, or perform user’s training.

The intention of this stage is to create a pseudo-reliable group of users, who will be later invited to the actual crowdtesting task. An example of such an application is a simple screen quality test, where the user has to select visible pictures from a group of difficult-to-see or invisible images on a low quality screen. The task is easy to do, fast to finish and has low pay, so within a short period of time and with low costs it is possible to create a reliable panel of users. This stage significantly improves the overall efficiency of the whole campaign. It has been observed that creating this pseudo-reliable panel would increases the overall efficiency by more than 60% [‎4].

The actual QoE test is then conducted in the second stage, only with invited reliable users from a previous campaign. However, it is important to test if the same hardware or software is being used as in the first stage, but also to test the users’ reliability, for example, with content questions, demographic questions, or repetitive presentation of tested content. Note, that the use of hidden reference methods e.g., in ITU-R BS.1116 [‎15] can be considered as consistency questions as suggested in the two-stage design. This stage also requires higher reward for the workers. In the notion of ITU-R BS.116 [‎15], we also apply a pre-screening and a post-screening technique. The major argument for introducing the pre-screening, i.e., the first stage, is to reduce costs of the overall campaign and to get a pseudo-reliable crowd, while the post-screening in the second stage is required to ensure a reliable data set. Although not necessary, it is sensible that the task required of the workers in the first stage is related to the task in second stage, for example, if the main task in second stage consists of a visual quality assessment, the first stage should also consist of a task including visual stimuli as in a screen quality test mentioned in the example above. Moreover, this can avoid any disappointment by the workers, resulting in decreased reputation of the employer, if the tasks in the two stages are very different. Also not every worker passing the first stage may be willing to participate in the second stage. In [‎16], for example, up to 75% of the workers passing the first stage declined to participate in the second stage.

**<Crowdsourcing Frameworks>**

Web-based crowdsourcing frameworks for QoE assessment represent an approach with programming tools to develop subjective studies that can be executed in a web browser. Such frameworks allow multimedia content to be displayed in a browser for test subjects to evaluate the quality using web forms. The test logic can be implemented at the client-side, e.g., javascript or at the server-side, e.g., PHP. Such frameworks allow to enable the execution of the experiments utilizing typical crowd-provider platforms. Functionality of each framework would include (1) the creation of the test (by supporting common standardized testing methodologies like ACR, DCR, PC), (2) test execution including training, task design, task order, and screening, and (3) the storage and access to the result data. Hoßfeld et al. provided a survey of widely used frameworks specifically developed for QoE experiment [‎3]. The survey is structured along specific criteria such as the experiment design, the applied test methodology, the type of media to evaluate, and the available hard- and software environment. In Table 1 a comparison of these frameworks is presented.

Table 1 Comparison of existing Crowdsourcing Frameworks for QoE Assessment [‎3]



**<Statistical Analysis of the campaign's test results ‎[4]>**

The first step to determine the unreliable user ratings is to consider the ratings provided for content questions, consistency questions, and gold data, in addition to video focus time as described in Section Validity check. It is quite common to consider the confidence intervals of the cumulative distribution function (CDF) of the users’ rating to quantify their reliability of the user ratings. However, this often mislead us, since a 95% confidence interval for a MOS value only shows that the mean rating including the unreliable user ratings lies within the confidence interval with a probability of 95%. On the other hand, confidence intervals *I* may even decrease in the presence of unreliable user ratings due to the increased number *N* of ratings in total as *I* $∼$ *1/*$\sqrt{N}$*.*

Therefore, the reliability of the users has to be quantified by appropriate means. In literature, there exist two overall categories of screening mechanisms: firstly, filtering of users based on the actual user ratings and secondly, screening of users, independently of the ratings, but with additional reliability mechanism e.g., consistency tests. In this contribution user rating based screening mechanism is abbreviated as URS and additional reliability mechanisms with ARM. The ARM approach leads to extra effort in the implementation and in the analysis, however, unreliable user can be clearly identified.

The URS screening methods can be roughly separated into at least two classes: one is based on inconsistencies compared with the mean result and relies on the ability of the subject to make correct identifications, the second class primarily eliminates subjects who cannot make the appropriate discriminations. Considering the variability of subjects’ sensitivities to different artefacts, however, the URS screening mechanisms are not sufficient for QoE crowdtesting and thus ARM is necessary for unreliable user identification. More details on URS screening methods are presented in [‎4].

 Summarizing, the presented URS approaches cannot be used alone to clearly identify unreliable users. Hidden influence factors or the variability of subjects’ sensitivities to different artefacts are not determined by those URS approaches. Although a combination of them may be interesting to improve screening, such as combining the random clicker approach and ITU-R BT.500, this remains a topic for future work. Hence, the screening of subjects should be done based on ARM methods as proposed in the two-stage design recommendation above, which clearly identifies unreliable users independent of any hidden influence factor and the actual user rating.

* **Reliability metrics to evaluate the crowdsourcing results**

For any subjective user study, specially for QoE crowdtesting, two reliability metrics for both inter-rate and intra-rater should be specified.

*Inter-rater reliability* describes the degree of agreement among raters. For a QoE crowdtesting this can be defined as the absolute value of Spearman rank-order correlation coefficient between all user ratings and the corresponding test conditions. Result of such a metric for a crowdsourcing campaign is a single value between 0 and 1. The low values of these metrics may not be necessarily caused by unreliable users, but may also be an indicator for hidden influence factors. Only high reliability values imply that the test subjects are reliable and that no hidden influence factors exist.

*Intra-rater reliability* however, determines to which extent the ratings of an individual user are consistent. Again, Spearman rank-order correlation coefficient can be used for ordinal data between the user ratings and the test condition.

**Bibliography**

[b-ITU-T X.yyy] Recommendation ITU-T X.yyy (date), *Title*.

1. J. Redi and I. 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.
2. M. Varela, T.M¨ aki, L. Skorin-Kapov, and T. Hoßfeld, “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
3. Tobias Hoßfeld, Matthias Hirth, Pavel Korshunov, Philippe Hanhart, Bruno Gardlo, Christian Keimel, and Christian Timmerer. 2014a. Survey of Web-based crowdsourcing frameworks for subjective quality assessment. In Proceedings of the 2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP’14). 1--6.
4. T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, “Best practices for qoe crowdtesting: QoE assessment with crowdsourcing,” Transactions on Multimedia, vol. 16, no. 2, Feb 2014.
5. [O. Alonso, D. E. Rose, and B. Stewart, “Crowdsourcing for relevance evaluation,” in ACM SigIR Forum, vol. 42, no. 2. ACM, 2008, pp. 9–15.
6. J. S. Downs, M. B. Holbrook, S. Sheng, and L. F. Cranor, “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.
7. B. Gardlo, S. Egger, M. Seufert, and R. Schatz, “Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing,” in International Conference on Communications, Sydney, AU, Jun. 2014.
8. ITU-T. Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in any Environment. ITU-T Recommendation P.913, Jan 2014.
9. P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach, “All subjective scales are not created equal: The effects of context on different scales,” Signal processing, vol. 77, no. 1, pp. 1–9, 1999
10. H. de Ridder, “Cognitive issues in image quality measurement,” Journal of Electronic Imaging, vol. 10, no. 1, pp. 47–55, 2001.
11. Y. Pitrey, U. Engelke, M. Barkowsky, R. Pepion, and P. Le Callet, “Aligning subjective tests using a low cost common set,” in Euro ITV, 2011, pp. irccyn contribution
12. C. Wu, K. Chen, Y. Chang, and C. Lei, “Crowdsourcing multimedia QoE evaluation: A trusted framework,” Transactions on Multimedia, vol. 15, no. 99, Jul. 2013. 63.
13. K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei, “A crowdsourceable QoE evaluation framework for multimedia content,” in Proceedings of the 17th ACM international conference on Multimedia, ser. MM ’09. ACM, 2009, pp. 491–500.
14. T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, “Best practices for qoe crowdtesting: QoE assessment with crowdsourcing,” Transactions on Multimedia, vol. 16, no. 2, Feb 2014.
15. J. Redi, E. Siahaan, P. Korshunov, J. Habigt, and T. Hossfeld, “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. ACM, 2015, pp. 33–38.
16. I. R. Assembly, ITU-R BS.1116-1, Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multichannel Sound Systems, 1997
17. M. Soleymani and M. Larson, “Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus,” in Proc. ACM SIGIR 2010 Workshop Crowdsourcing for Search Evaluation (CSE 2010), Jul. 2010, pp. 4–8.