|  |  |  |  |
| --- | --- | --- | --- |
| **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. **Image Set:** A collection images extracted from the original set of test images, treated as its own study set.
9. **Overlapping Image sets:** A number of images that range in quality from best to worst, taken from the original set of images 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.

# Test Methods and experimental design

In order to successfully design a crowdsourcing study, experimenters are encouraged 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. We strongly recommend that experimenters 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 and usage characteristics by submitting study eligibility criteria to the platform (i.e. age, gender, geo locations…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 should 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.

Research question and experimental design goals must be considered heavily when setting exclusion criteria.

## Compensation

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. 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

Similar to laboratory testing, workers should not be exposed to lengthy tests that induce boredom, fatigue, and disinterest. 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.

One moving an experiment from the lab to a crowdsourcing platform, a pilot test of 20-30 participants will give insight to appropriate study length and data accuracy and reliability.

### 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 should 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 should consider their experimental design and research method when deciding which rating scale to use for an IQA. Some of the common rating methods are:

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

## 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 study with 20-30 workers first to test 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.

## Image sets

An image set is a smaller number of images that make up a test, chosen from the original larger pool of content. In order to lessen the amount of time spent on a test, it is recommended that experimenters break a test with a large amount of content into smaller image sets, and allow workers to participate in as many image sets as they would prefer to. Images in each image set are not mutually exclusive; some images could be used in other image sets as test images as well.

## 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

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 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.

## Overlapping warm-up images

Since not participant will see every image, we strongly encourage having a common set of 10 images before every set of images in a series of studies that use the same image set. These images assist experimenters with testing between subject stability and replicability of within subject ratings over time. Overlapping images are chosen from the original pool of images, ranging in quality from low to high quality. Additionally, have the images be representative of the contents of the photos in your set. For example, if 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 and encouraged for IQAs with large number of content images, divided into smaller image 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.

It is also encouraged that the experimenter 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 what will types of responses may result in rejected work, if any. 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 a experiment pays enough. Another consideration is to identify your research group. While remaining anonymous may help to protect the the reputation of the research lab, stating affiliations helps to build trust with the crowdsourcing community.

**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>

**Bibliography**

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