A CROWDSOURCED IMPLEMENTATION OF ITU-T P.910

Babak Naderi\textsuperscript{1}, Ross Cutler\textsuperscript{2}

\textsuperscript{1}Technical University of Berlin, \textsuperscript{2}Microsoft Corporation

ABSTRACT

The gold standard for measuring video quality is the subjective test, and the most prevalent is the ITU-T P.910, a lab-based subjective standard in use for the past two decades. However, in practice using P.910 is slow, expensive, and requires a lab, which all create barriers to usage. As a result, most research papers that need to measure video quality don’t use P.910 but rather metrics that are not well correlated to subjective opinion. We provide an open-source extension of P.910 based on crowdsourcing principles which address the speed, usage cost, and barrier to usage issues. We implement Absolute Category Rating (ACR), ACR with hidden reference (ACR-HR), and Degradation Category Rating (DCR). It includes rater, environment, hardware, and network qualifications, as well as gold and trapping questions to ensure quality. We have validated that the implementation is both accurate and highly reproducible compared to existing P.910 lab studies.

Index Terms: Subjective Quality, Video Quality

1. INTRODUCTION

Measuring the quality of video is an important task in many engineering areas, such as codec development, video enhancement, and video telecommunication systems. As a result, there are many standards and metrics for measuring video quality, though the gold standard is the subjective test done in a controlled lab environment. The most prevalent of these standardized subjective tests is the ITU-T P.910 [1]. However, using P.910 in practice is slow due to the recruitment of test subjects and the limited number of test subjects, and expensive due to paying qualified test subjects and the cost of the test lab. The speed and cost result in the vast majority of research papers not using P.910 but rather objective functions that are not well correlated to subjective opinion.

An alternative to lab-based subjective tests is to crowdsource the testing, and there are several such systems that do this (see Table 1). However, none of these identified systems have the rater, environment conditions, and hardware qualifications that P.910 requires. In addition, these existing systems have not been rigorously validated to show they are accurate compared to a P.910 lab-based study and give reproducible results. We provide a crowdsourced implementation that includes rater, environment, hardware, and network qualifications, as well as gold and trapping questions to ensure quality. We include a validation study that shows it is both accurate and highly reproducible compared to existing P.910 lab studies. The tool is open-sourced and can be used on the Amazon Mechanical Turk platform for wide-scale usage. We recently used the tool in the CVPR 2022 CLIC challenge to provide the challenge metric for ML-based video codecs.

The vast majority of published video compression research papers provide results using the objective metrics Peak signal-to-noise ratio (PSNR) [2], structural similarity index measure (SSIM) [3], multi-scale SSIM (MS-SSIM) [4], and Video Multi-Method Assessment Fusion (VMAF) [5], e.g., [6, 7, 8]. However, it has been shown that PSNR, SSIM, and MS-SSIM are not well correlated to subjective opinion [5, 9]. VMAF has been trained on traditional DSP-based video codecs such as H.264/AVC [10], but has not been shown to correlate with the newer deep learning based codecs. Moreover, VMAF hasn’t been trained for longer term effects like recency, primacy, and rebuffering [11]. Therefore for video codec research there is no alternative to subjective tests to evaluate and compare codecs at this time.

Section 2 provides an overview of related work, Section 3 describes our tool’s implementation, Section 4 gives the tool’s validation, and Section 5 gives conclusions.

2. RELATED WORK

A recent review of subjective quality assessment tools is given in [12]. P.910 [1] provides a general subjective video quality assessment standard for multimedia applications. P.910 includes ACR, ACR-HR, DCR, and paired comparison (PC) methods, as well as rater qualifications, environment conditions, and video playback procedures. P.912 [13] provides a target-specific subjective video quality assessment standard, such as for faces, license plates, etc.

Tominaga et al. [14] conducted a comparison of eight different subjective video quality assessment methods and found that ACR was the most suitable for statistical reliability, assessment time, and ease of evaluation.

Keimel et al. [15] describes an open-source tool Quality-Crowd that supports ACR video quality assessment. Quality-Crowd is extended by [16] to include a Double Stimulus Continuous Quality Scale (DSCQS). Rainer et al. [17] describes
the tool WESP, an open-source tool that supports ACR, ACR-HR, DCR, and PC.

Jung et al. [18] provide a methodology how to conduct remote subjective video quality assessment studies in which the raters download videos and view them manually on their devices. While the methodology suggests the rater checks visual acuity and color blindness, it doesn’t provide any such tests. There are also no tests for environmental conditions or hardware setup, which we will show is essential.

| Tool       | Measures          | Rater qual | Viewing cond | HW | Network | Accur | Repro |
|------------|-------------------|------------|--------------|----|---------|-------|-------|
| QualityCrowd [15, 16] | ACR, ESSQ, ESQS | N          | N            | N  | N       | Y     | N     |
| WESP [17]   | ACR, ACR-HR, DCR, PC | N          | N            | N  | N       | N     | N     |
| avrateNG [19] | ACR              | N          | N            | N  | N       | Y     | N     |
| Ours        | ACR, ACR-HR, DCR  | Y          | Y            | Y  | Y       | Y     | Y     |

Table 1: Open-source crowdsourcing video quality assessment systems

3. IMPLEMENTATION

Our implementation\(^1\) can be used either as an integrated survey in Amazon Mechanical Turk (AMT) or as an external survey deployed on a dedicated Web server\(^2\). Consequently, AMT or any other crowdsourcing platform can be used for recruiting and paying test participants. We also provide a lightweight container-based web application that can serve the experiment. The web application can easily be deployed on any Linux virtual machine. As a result, this implementation can be used for both crowdsourcing tests or remote testing with a dedicated panel of participants.

The open-source implementation includes the Absolute Category Rating (ACR), ACR with hidden reference (ACR-HR), and Degradation Category Rating (DCR). All methods can be used with either a five or nine point discrete Likert-scale. We also followed and extend best practices on video and speech quality assessment in crowdsourcing [20, 21] in our implementation.

3.1. Tools

We provide a set of program scripts to ease the interaction with the system and avoid operation errors. The scripts are used to prepare the subjective test (e.g., create customized trapping sequences), process the submitted answers (i.e., data cleansing, aggregating ratings, and generating reports), and interact with AMT (e.g., sending bonuses to workers).

In the first step (see Figure 1), trapping clips customized to the dataset under the test should be created using the corresponding script. They will be created by adding a text message, in the middle of selected existing video sequences, asking participants to select a specific score when rating this trapping clip. Next, test configuration and URLs for test sequences (in case of DCR and ACR-HR also reference sequences), trapping clips, gold, and training clips should be provided to the master script. It creates the HTML template of the test, a list of variables (dynamic content), and a configuration file for the result parser. Gold clips are video sequences whose quality is known to the experimenter (either excellent or bad) [22]. A test can be created in the HIT App Server by providing the HTML template and a list of variables generated in the last step. Finally, a generic project description and list of URLs can be downloaded and used to create a new test in AMT (or any other crowdsourcing platform).

When the test is finished, the submitted answers to AMT, and HIT App Server should be provided to the result parser script in addition to the configuration file which was created in step 2 (see Figure 1). It will perform the data cleansing process, and aggregate the valid and reliable ratings over test sequence and over Hypothetical Reference Circuits (HRCs) if exist. In addition, reports on the list of bonus assignments, and lists of the submissions to be accepted/rejected or extended will be generated.

3.2. Test components

The subjective test is organized in different sections from the participant’s perspective (see Figure 2). Each section is designed to instruct the test participant, qualify the participant, their environment, and their setup, and collect their votes. The instructions and ratings sections are included in all tests, whereas the qualification, and calibration only need to be performed once per test. The setup and training sections are shown periodically. It is recommended to keep the crowdsourcing test session short (i.e., less than 15 minutes), therefore we present a limited set of test sequences in the rating

---

\(^1\)https://github.com/microsoft/P.910

\(^2\)Currently the integrated surveys within AMT are not able to playback videos in full-screen mode, we strongly recommend to deploy the test as an external survey.
section. We encourage participants to take part in more than one session by providing monetary incentives and dynamic sections (i.e., the next test session will be significantly shorter than the first one).

![Flowchart](Image)

**Fig. 2:** The crowdsourcing test from the participant’s perspective.

A set of automatic measurements are performed on the test’s loading time. The experimenter can restrict test participants to specific viewing devices (i.e., mobile, PC, or both), minimum screen refresh rate, and minimum resolution.

### 3.2.1. Video playback

We developed a video playback component based on HTML5. To avoid the confounding effect of network connection, it fully downloads all the video materials on the browser’s local storage. The videos are also played in full-screen mode, and the participant must watch the video until the end before casting their votes. The player records the playback duration of videos which will be used later in the data cleansing process. The experimenter can decide whether the videos should be presented in their original size or scaled to fill the participant’s viewing device. In the case of the DCR test, the reference and processed video sequence will be shown in a sequence with a gray screen in between for one second.

### 3.2.2. Qualification

Within the qualification section, the eligibility of test participants is evaluated. According to the ITU-T Rec. P.910, participants should be screened for normal color vision and normal or corrected-to-normal visual acuity (i.e., no error on the 20/30 line of a standard eye chart) [23]. The standard Ishihara test for color vision includes 15 plates in the normal and 6 plates in the short version, which are both too long for a crowdsourcing test. In a prestudy, we invited 300 participants from AMT (34% Male) and 191 participants (91% Male) from two Internet communities dealing with color vision deficiencies. Both groups participated in the full Ishihara test in which 6% of participants from AMT and 96% from online forums were detected as color blind. Applying a decision tree classifier with entropy as the criterion revealed that only using Plates 3 and 4 reached 98% accuracy (sensitivity 0.996, specificity 0.95). Consequently, we only use these two plates in the qualification section.

For the visual acuity test, we use the Landolt ring optotypes as recommended by ISO 8596:2017 [24]. In this test, participants are presented with broken rings (like the letter C). The gap could be in 8 directions (every 45 degrees). The diameter of the ring is 5 times the size of the gap. The visual acuity is the inverse value of the gap size (in arc minutes) of the smallest identified Landolt ring [24]. In each row (i.e., ring size) 5 samples should be presented and the participant should answer 3 or more correctly to pass that size. Our implementation of the visual acuity test consists of two steps: 1) setup, 2) answer to up to 5 Landlot rings on the specific size. In the setup section, the size of a pixel in the user’s screen is estimated by asking them to adapt the size of a given picture (here a credit card) on their screen by clicking on +/- buttons to reach the specified size. In addition, we asked them to sit in the range of 50 to 75cm from the screen during the test. Although the maximum viewing distance of participants is limited to a distance in which they can easily interact with their device, we cannot enforce a minimum distance. Given the pixel size, the corresponding Landloth ring size is calculated in which correctly answering them in 50 cm distance shows a minimum of 20/30 visual acuity. The participant passes the visual acuity test if they correctly select the direction of three out of the five Landlort rings. Finally, participants only can continue to the next sections, if they successfully pass the qualification.

### 3.2.3. Display calibration and instructions

Participants are asked to set the resolution of their device to default or recommended value suggested by their operating system. We also ask them to perform display color calibration using methods provided by their operating system. This section provides information on how to perform these tasks for Windows and Mac operating systems and will only be shown once during the test.

In the instruction section, a sample video for some of the perceptual quality dimensions [25, 26] (i.e., fragmentation, discontinuity, and uncleanness) are presented to the participants. They are also informed that impairments can happen in a specific area or time within a video clip and the impairments are not limited to the presented samples.

### 3.2.4. Setup

In the setup section, the viewing conditions including the brightness of the screen, background room light, and viewing distance are evaluated. The ITU-T Rec. P.910 provides a detailed specification for viewing conditions that can only be measured in a controlled laboratory environment using professional devices. We created a brightness/light calibration task in which participants should count the number of geometric shapes presented in a picture. They can insert their answers and verify them. Given the correct answer, they can continue to the next steps whereas on mismatch they are encouraged to modify the brightness of their screen, lighting of the environment, or repeat the display calibration task until they can insert the correct number for each picture. The presented picture is a matrix of 4x4 squares. Each square has a
different level of a gray background and may contain a triangle or circle in different sizes and locations. The foreground color of the shapes is close to the background color of the square. A pretest, with a limited group of experts, showed that the calibration task performs best when the difference between foreground and background colors is four integer points in RGB24 color space. Figure 3 illustrates an example matrix picture used in our tests. Within the setup section, a second matrix is also presented for which no feedback is given. In post-processing, submissions will be filtered based on their answers to the second matrix question.

The crowdsourcing test also includes a short paired-comparison task to evaluate participants’ viewing distance. The task is inspired by [27] which is the recommended method to evaluate crowd workers’ setup and environment in a speech quality assessment test [22]. In this task, three pairs of images are presented to the participants who are asked to sit at a distance of 1.5 times their screen height. In each pair, one of the images is distorted with a blur effect. Participants are asked to select the image of better quality. They can choose between either of the images or select that both have the same quality. In a separate experiment, we used blur effects with different radiiuses from 1 pixel to 9. A small group of participants took part in six Just-Noticeable Difference (JND) quality tests (adaptive staircase method 2 down–1 up) [28] with two different screen sizes and three different viewing distances (too close, expected, too far). Using the results, we selected three pairs for this task: 1) the difference is mostly recognized if the participant is too close to the screen, 2) the difference is correctly recognized if the participant is seated at the expected distance, 3) the difference is still recognized even they are too far from their screen. Given their response, there might be a feedback message asking them to adjust their seating distance to 1.5 times their screen height. An example from the paired-comparison questions is given in Figure 4.

3.2.5. Training and rating

In the training section, participants should watch the videos from the training set and rate their quality. The training set should represent the dataset under the test and cover the entire range of the rating scale. We also add a trapping clip to the training set, which instructs the participant to make a specific selection with text in the video. In case participants provide a wrong answer, they will be alerted and asked to watch the video again. Finally, the training section is shown periodically (every 60 minutes) to keep its anchoring effect following best practices in the speech quality domain [29]. Within the rating section, participants rate the quality of ten video clips, one trapping, and one gold clip. The trapping and gold clip will be inserted automatically and will be used in the data cleansing step in post-processing.

4. VALIDATION

We used the VQEG HDTV datasets [30] to evaluate the validity and reliability of our implementation. The datasets contain coding only and coding plus transmission error impairments. They were created to validate objective video quality models that predicted the quality of High Definition Television (HDTV). The video materials and subjective data from experiments VQEGHD1-5 (in a laboratory) are made publicly available in the Consumer Digital Video Library⁴. Each experiment includes 168 video clips, with 24 shared between all experiments. Each experiment includes its specific dataset where 21 HRCs were applied to 9 source videos (i.e., 144 processed sequences). We used the datasets from experiments VQEGHD3 and VQEGHD5 for which 40” (native resolution 1920x1080) and 24” (native resolution 1920x1200) displays were used in the laboratory viewing sessions by 24 test participants, respectively. The test sequences have various bit rates; 1-15 Mbps for VQEGHD3 and 2-16 Mbps for VQEGHD5. The test sequences created from two source clips in the dataset VQEGHD5 were not included in the published package, leading to 136 sequences. Results of laboratory-based subjective tests were published in [30] by VQEG.

To prepare the video clips for our crowdsourcing experiments, we re-encode the video sequences using FFmpeg with the H.264/AVC codec keeping the original bit rate, framerate, and resolution with a GOP of one second and Constant Rate Factor of 17.

We conducted six crowdsourcing studies with the two above-mentioned datasets. In one of the tests, the VQEGHD5 dataset was used. In the rest of the five tests, we used VQEGHD3 dataset on five separate days, each with unique raters, to evaluate the reproducibility of our implementation.

---

3For instance, if one answers (2) wrongly but (3) correctly, it will be categorized as sitting too far from their screen.

⁴https://www.cdvl.org/
In all tests, participants with a minimum display resolution of 1280x720 and a refresh rate of 30Hz were allowed.

4.1. Accuracy

In each experiment, 10 test sequences, one gold clip, and one trapping clip were presented in one session and we aimed to collect 30 ratings per clip. In the five experiments using the VQEGHD3 dataset on average 71% of submissions passed the data cleansing step; the rest were not used due to the rater providing a wrong answer to the gold question, longer playback duration, failure on the second brightness check, low variance in ratings, or wrong verification code. On average we have 21 accepted votes per test sequence (with a minimum of 15 ratings).

In the sixth experiment with the dataset VQEGHD5, 78.8% of submissions passed the data cleansing step; the rest were not used due to similar issues. On average we have 25 accepted votes per test sequence (with a minimum of 20 ratings).

The results are presented in Table 2 per test sequences and in Table 8 per HRCs. We observed a strong correlation between laboratory and crowdsourcing subjective tests ($\tau = 0.952$ per test sequence and $0.964$ per HRC). The distribution of MOS values from crowdsourcing and laboratory tests, before mapping, are illustrated in Figure 5.

4.2. Reproducibility

On average 63 unique workers participated in each run. We observed $PCC = 0.971$ and $SRCC = 0.95$ on average between MOS values of sequences in the five different runs. The correlation coefficients between the runs are reported in Table 3. We also fitted a linear mixed-effects model (LMEs) with random intercept in which test sequences and runs are considered as fixed factors and participants as a random factor. The result shows there was no significant effect of run on the subjective ratings ($\chi^2(4) = 2.691, p = 0.611$).

4.3. Ablation study

We evaluated the effect of different components on the performance of the crowdsourcing test. In pretests, we observed that percentage of correct answers to brightness check increases by 19% when participants are asked to perform screen cali-
Fig. 5: Distribution of MOS values between laboratory-based and crowdsourcing-based experiments (a-b) and two crowdsourcing runs (c). (a) Dataset VQEG HD3-run1 and (b) Dataset VQEG HD5.

Table 3: Correlation coefficients between five runs of the VQEGHD3 dataset. Pearson correlation coefficient on upper triangle and Spearman’s rank correlation coefficient on lower triangle.

|           | Run 1       | Run 2       | Run 3       | Run 4       | Run 5       |
|-----------|-------------|-------------|-------------|-------------|-------------|
| Run 1     | 0.984       | 0.987       | 0.957       | 0.977       |             |
| Run 2     | 0.959       | 0.985       | 0.957       | 0.977       |             |
| Run 3     | 0.974       | 0.969       | 0.952       | 0.972       |             |
| Run 4     | 0.943       | 0.941       | 0.942       | 0.956       |             |
| Run 5     | 0.954       | 0.947       | 0.942       | 0.933       |             |

Table 4: Effect of integrated filtering criteria on validity of crowdsourcing test.

| Case                      | Avg.# Ratings | PCC  | SRCC | RMSE  | RMSE FOM |
|---------------------------|---------------|------|------|-------|----------|
| All passed                | 104           | 0.96 | 0.96 | 0.62  | 0.31     |
| Gold clips failed         | 17            | 0.57 | 0.53 | 1.02  | 0.93     |
| Playback duration failed  | 8             | 0.62 | 0.57 | 1.12  | 0.89     |
| Brightness check failed   | 17            | 0.89 | 0.88 | 0.84  | 0.52     |
| Straightliners            | 6             | 0.29 | 0.30 | 1.53  | 1.09     |

All statistics are significantly different from their counterpart in All passed at $\alpha = .05$.

4.3.1. Viewing distance

We conducted a crowdsourcing experiment with two settings in which participants were asked 1) to sit at 1.5 times the height of their screen and 2) to sit as far as they could but still be able to interact with their device and not be closer than three times of the height of their screen. The setup section (including viewing distance check) was forced to be shown in all instances for both settings. Despite the original viewing distance test, no feedback to participants was provided upon their answers. Submissions that pass all the data cleansing criteria are aggregated based on their performance in the viewing distance test. Results are presented in Table 5. The correlation coefficient between ratings from submissions that passed the viewing distance and laboratory scores is significantly higher than those that failed the test ($z = 4.435$, $p < .001$).

4.3.2. Visual acuity test

In a new crowdsourcing test, the task was modified so that the Visual Acuity Test (VAT) did not give any feedback on its finish, and despite its results, participants could continue to the rating section. Meanwhile, the qualification section (which contains the VAT) was included in all tasks. To increase the
Table 5: Effect of the viewing distance (VD) component on the validity of the crowdsourcing test.

| Case                  | PCC  | SRCC | RMSE | RMSE FOM |
|-----------------------|------|------|------|----------|
| VD - Passed           | 0.93 | 0.92 | 0.76 | 0.41     |
| VD - Failed           | 0.83 | 0.78 | 0.95 | 0.62     |
| VD - Failed (too far) | 0.87 | 0.84 | 0.81 | 0.58     |
| VD - Failed (unknown category) | 0.79 | 0.72 | 1.07 | 0.70     |

Beside this value, all statistics are significantly different from their counterpart in VD- passed at $\alpha = 0.05$.

Table 6: Effect of the Visual acuity test (VAT) component on the validity of the crowdsourcing test.

| Case                          | PCC  | SRCC | RMSE | RMSE FOM |
|-------------------------------|------|------|------|----------|
| VAT Passed & All criteria passed | 0.91 | 0.90 | 0.88 | 0.47     |
| VAT Failed & All criteria passed | 0.87 | 0.89 | 0.96 | 0.59     |
| VAT Passed *                  | 0.91 | 0.90 | 0.97 | 0.48     |
| VAT Failed *                  | 0.63 | 0.59 | 1.41 | 0.89     |

* No other criterion was considered.

Table 7: Effect of calibration task and trapping clips.

| Case                          | PCC  | SRCC | RMSE | RMSE FOM | sig. diff. to lab |
|-------------------------------|------|------|------|----------|------------------|
| Complete test                 | 0.95 | 0.95 | 0.72 | 0.34     | %16              |
| No Calibration                | 0.91 | 0.89 | 0.80 | 0.47     | %23              |
| No Trapping clip              | 0.92 | 0.92 | 0.88 | 0.46     | %29              |

Beside this value, all statistics are significantly different compared to the Complete test at $\alpha = 0.05$.

We anticipate the difference is due to the fact that the laboratory test was conducted using a 40” screen, whereas crowd workers used their PC and laptop screens (typically 11 to 24”).

5. CONCLUSIONS

We have created the first crowd-sourced implementation of P.910 that is shown to be as accurate as a lab study and is highly reproducible. In a set of experiments, we showed that each test component significantly improves the validity of collected ratings.

It should be noted that the impact of the test components depends on the reliability of crowd workers (and their environment and device) and the test sequences under the study. If a dedicated group of workers performs the test in a suitable environment using proper devices, those components might look redundant. However, they strongly increase the validity of collected scores despite the group of test participants and make large-scale data collection possible.

We plan to include the simultaneous presentation of sequence pairs for future work and implement P.910 Annex E: An advanced data analysis technique for tests under challenging conditions.

6. REFERENCES

[1] “ITU-T Recommendation P.910 Subjective video quality assessment methods for multimedia applications,” International Telecommunication Union, Tech. Rep., 2021.

[2] R. Gonzalez and R. Woods, Digital image processing, 3rd ed. Prentice Hall, 2006.

[3] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[4] Z. Wang, E. Simoncelli, and A. Bovik, “Multiscale structural similarity for image quality assessment,” in The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. Pacific Grove, CA, USA: IEEE, 2003, pp. 1398–1402.
[5] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, “Toward A Practical Perceptual Video Quality Metric,” Tech. Rep., 2016. [Online]. Available: http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html.

[6] J. Lin, D. Liu, H. Li, and F. Wu, “M-LVC: Multiple Frames Prediction for Learned Video Compression,” in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, Jun. 2020, pp. 3543–3551.

[7] Z. Hu, G. Lu, and D. Xu, “FVC: A New Framework towards Deep Video Compression in Feature Space,” in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, Jun. 2021, pp. 1502–1511.

[8] O. Rippel, A. G. Anderson, K. Tatwawadi, S. Nair, C. Lytle, and L. Bourdev, “ELF-VC: Efficient Learned Flexible-Rate Video Coding,” in 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, Oct. 2021, pp. 14 459–14 468.

[9] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, “A Subjective Study to Evaluate Video Quality Assessment Algorithms,” 2010.

[10] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, “Overview of the H.264 / AVC Video Coding Standard,” IEEE Transactions On Circuits And Systems For Video Technology, p. 19, 2003.

[11] Z. Li, C. Bampis, J. Novak, A. Aaron, K. Swanson, A. Moorthy, and J. De Cock, “VMAF: The Journey Continues,” Tech. Rep., 2018. [Online]. Available: https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12

[12] M. Testolina and T. Ebrahimi, “Review of subjective quality assessment methodologies and standards for compressed images evaluation,” in Applications of Digital Image Processing XLIV. SPIE, Aug. 2021, pp. 302–315.

[13] “Recommendation ITU-T P.912: Subjective video quality assessment methods for recognition tasks,” International Telecommunication Union, Tech. Rep., 2016.

[14] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, “Performance comparisons of subjective quality assessment methods for mobile video,” in 2010 Second International Workshop on Quality of Multimedia Experience (QoMEX). Jun. 2010, pp. 82–87.

[15] C. Keimel, J. Habig, C. Horch, and K. Diepold, “QualityCrowd-A Framework for Crowd-based Quality Evaluation,” 2012.

[16] “Large-Scale Crowdsourcing Subjective Quality Evaluation of Learning-Based Image Coding,” IEEE Visual Communications and Image Processing (VCIP 2021), 2021.

[17] B. Rainer, M. Waltl, and C. Timmerer, “A web based subjective evaluation platform,” in 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX), Jul. 2013, pp. 24–25.

[18] J. Jung, M. Wien, and V. Baroncini, “ISO/IEC JTC 1/SC 29/AG 5: Guidelines for remote experts viewing sessions,” Nov. 2021.

[19] R. R. R. Rao, S. Göring, and A. Raake, “Towards High Resolution Video Quality Assessment in the Crowd,” in 2021 13th International Conference on Quality of Multimedia Experience (QoMEX), Jun. 2021, pp. 1–6, ISSN: 2472-7814.

[20] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, “Best practices and recommendations for crowdsourced qoe-lessons learned from the qualinet task force” crowdsourcing”, 2014.

[21] ITU-T PSTR-CROWDS, Subjective evaluation of media quality using a crowdsourcing approach. Geneva: International Telecommunication Union, 2018.

[22] ITU-T Recommendation P.808, Subjective evaluation of speech quality with a crowdsourcing approach. Geneva: International Telecommunication Union, 2021.

[23] ITU-T Recommendation P.910, Subjective video quality assessment methods for multimedia applications. Geneva: International Telecommunication Union, 2021.

[24] “ophthalmic optics — visual acuity testing — standard and clinical optotypes and their presentation,” Geneva, CH, Standard.

[25] F. Schiffner and S. Möller, “Defining the relevant perceptual quality space for video and video-telephony,” in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017, pp. 1–3.

[26] ITU-T Recommendation P918, Dimension-based subjective quality evaluation for video content. Geneva: International Telecommunication Union, 2021.

[27] B. Naderi and S. Möller, “Application of just-noticeable difference in quality as environment suitability test for crowdsourcing speech quality assessment task,” in 2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2020, pp. 1–6.

[28] B. Treutwein, “Adaptive psychophysical procedures,” Vision research, vol. 35, no. 17, pp. 2503–2522, 1995.
[29] B. Naderi and R. Cutler, “An Open Source Implementation of ITU-T Recommendation P.808 with Validation,” *INTERSPEECH*, pp. 2862–2866, Oct. 2020.

[30] Video Quality Experts Group, “Report on the validation of video quality models for high definition video content,” 2010. [Online]. Available: https://www.its.bldrdoc.gov/media/4212/vqeg_hdtv_final_report_version_2.0.zip

7. APPENDIX

**Table 8**: Comparison between laboratory and crowdsourcing tests in HRC level

| Dataset       | PCC | MOS | RMSE | RMSE FOM |
|---------------|-----|-----|------|----------|
| VQEG HDTV3 -run1 | 0.967 | 0.980 | 0.655 | 0.248 |
| VQEG HDTV3 -run2 | 0.977 | 0.982 | 0.618 | 0.211 |
| VQEG HDTV3 -run3 | 0.968 | 0.981 | 0.577 | 0.245 |
| VQEG HDTV3 -run4 | 0.940 | 0.975 | 0.706 | 0.333 |
| VQEG HDTV3 -run5 | 0.965 | 0.972 | 0.671 | 0.257 |