Impact of feedback on crowdsourced visual quality assessment with paired comparisons

Mohsen Jenadeleh*, Alexander Heß*, Simon Hviid Del Pin†, Edwin Gamboa‡,
Matthias Hirth§, Dietmar Saupe*
*Department of Computer and Information Science, University of Konstanz, Konstanz, Germany
†Department of Computer Science, Norwegian University of Science and Technology, Gjøvik, Norway
‡ScaleHub GmbH, Germany
§Faculty of Electrical Engineering and Information Technology, Ilmenau University of Technology, Ilmenau, Germany
{mohsen.jenadeleh, alexander.2.hess, dietmar.saupe}@uni-konstanz.de, simon.h.d.pin@ntnu.no,
edwin.gamboa@scalehub.com, matthias.hirth@tu-ilmenau.de

***Abstract*—This paper presents a comprehensive investigation into the effects of immediate feedback on crowdworkers' performance in subjective image quality assessment tasks using paired comparisons. The study is motivated by the need for reliable and efficient crowdsourcing tasks for image quality assessment. A large-scale experiment involving 200 participants was conducted, where participants completed 120 paired comparisons with and without feedback. The feedback informed the workers of the correctness of their responses to comparisons. Almost all of the participants (97%) preferred receiving feedback. The results indicate that feedback reduced response time, improved user experience, and did not cause a bias in the estimation of the just noticeable difference (JND). On the other hand, feedback did not significantly affect accuracy, correlation with the ground truth, or create a learning effect. This study contributes to the field by being one of the first to examine the impact of feedback on crowdworker performance in subjective image quality assessment tasks. The dataset which includes the images and ratings can be accessed at https://database.mmsp-kn.de/feedback-study-dataset.html.**

***Index Terms*—Two-alternative forced choice, psychometric functions, crowdsourcing, subjective image quality assessment, feedback, just noticeable difference**

I. Introduction

Subjective quality assessment of visual media relies on data from human subjects which typically is collected in laboratory or crowdsourcing studies. Crowdsourcing has several advantages over lab studies for perceptual tasks such as image and video quality assessment. In particular, Crowdsourcing offers easy access to diverse populations, scalability enabling very large experiments in a time-efficient manner, flexibility in the duration of experiments, and all of this at relatively low cost.

However, crowdworkers cannot be personally guided and supervised by the experimenter. They may easily get distracted from their work, and their motivation and work quality may decrease, e.g., due to boredom caused by the similarity of the task, waiting times for loading images and videos, or the complexity of cognitive tasks. Also, crowdworkers may just try to maximize their hourly earnings by performing the subjective task as quickly as possible. As a result, they may provide only low-quality subjective data.

Thus, two of the main issues researchers encounter when conducting crowdsourcing experiments is how to ensure people stay attentive during the crowdsourcing experiments and how to ensure the quality of their work. Therefore, it is an overarching goal to develop a reliable crowdsourcing environment to ensure high-quality, accurate, and consistent subjective data [1], [2].

Adding game elements such as competition, collaboration, or reward can increase users' desire to participate in crowdsourcing experiments and motivate them to perform the subjective task accurately [3], [4].

In some cases, it may be enough to ask participants to volunteer out of personal interest. However, in many cases it is more practical to hire freelancers or crowdworkers on platforms such as Amazon Mechanical Turk (MTurk) in exchange for monetary reward. Notably, these platforms typically offer a uniform per-task payment to all crowdworkers who pass quality control, regardless of their task completion reliability.

To further motivate crowdworkers to pay attention and fully engage in the crowdsourcing task, gamification can be a great driver [5]. Our general hypothesis is that adding competitive or reward elements can increase people's desire to contribute to crowdsourcing tasks and also provide more accurate, precise, and reliable subjective data.

In this study we focus on the most simple component that can be considered as an element of gamification, namely immediate feedback to crowdworkers' responses. One can think of many different forms of such feedback, for example, encouragement like "Good job!" or "You just finished Level 2! Keep going!". Another kind of feedback is a progress bar that shows the percentage of the job that has been finished at each time.

However, in a class of crowdsourcing experiments for image/video quality assessment another important kind of feedback can be given, namely for full-reference image/video quality assessment (FR-IQA) with paired comparisons. Source images are distorted by, e.g., compression artefacts, and a distorted stimulus is shown side-by-side with the corresponding source. The ordering is random; the source image may appear

This research is funded by the DFG (German Research Foundation) – Project ID 496858717, titled "JND-based Perceptual Video Quality Analysis and Modeling". D.S. is funded by DFG Project ID 251654672 – TRR 161. S.H.D.P. is supported by the project "Quality and Content: understanding the influence of content on subjective and objective image quality assessment" (grant number 324663, approved on 1 October 2021) from the Research Council of Norway.

979-8-3503-6157-5/24/$31.00 ©2024 IEEE

on the left side or the right side. Subjects are asked to detect the distortion and respond to the question on which side the stimulus with the better quality (i.e., the source) is, left or right.

This experimental procedure enables important applications. From many FR-IQA comparisons the distorted stimuli can be scaled to quantify the perceived visual quality. Also it is possible to extract the Just Noticeable Difference (JND), i.e., the minimal distortion level for which at least half of the observers can detect a distortion relative to the source stimulus [6], [7].

In FR-IQA with paired comparisons it is plausible to assume that the perceived visual quality of each stimulus is a monotonic function of the degree of distortion.[1] Therefore, although a pair comparison response is a subjective judgment, it is clear to us whether it identifies the source stimulus correctly. Thus, for each comparison, feedback about the correctness of the response of the crowdworker can be given. This is the setting for our feedback study. In the following, *feedback* shall refer to informing participants about the correctness of their answers to such comparison tasks.

Let us assume we have collected two sets of responses to the same set of comparisons (questions) in two conditions, one with and one without feedback. At a first level of statistical data analysis we compare the two conditions based on different metrics.

First, we evaluate the *Accuracy* of the responses, using the percentage of correct responses for each set and the Kendall rank order correlation (KRCC) of the reconstructed percentages with the ground truth given by distortion magnitudes.

As second metric, we consider the *Response time*. Feedback may increase the confidence of subjects in their ratings, such that the responses can be given more quickly.

Next, we evaluate the *Learning effect*. When subjects perform a difficult perceptual task there may be a learning effect. Then, the proportion of correct responses increases during an experiment. With feedback about the correctness of responses this learning effect may be accelerated and reaching farther.

Finally, we consider the *User experience*. In particular, we are interested if the feedback improves the user experience of the crowdsourcing subject.

A second level of data analysis follows after the scaling of the response data, which yields reconstructed perceptual quality values for each stimulus in the form of a psychometric function for each combination of source stimulus and distortion type. The visual qualities of the source stimuli are anchored at zero.

There are several types of psychometric functions that can be applied. For our purposes, we selected the Weibull cumulative probability density function (cdf). It has only two parameters, the scale and the shape parameters. The median is interpreted as the JND and the variance governs the slope of the function at the JND, and, therefore is a measure of the precision of the JND assessment. This leads up to the following two main hypotheses.

[1]There are counterexamples to this assumption. For example, some image/video compression techniques may also have a denoising effect. This may invoke an increase of perceptual quality at high quality parameter settings.

Hypothesis 1 *Feedback does not influence the JND (null-hypothesis).*

Hypothesis 2 *Feedback increases the precision of JND assessment.*

For the latter, we also consider the corresponding null-hypothesis.

The contributions of our work can be summarized as follows:

- We have developed a framework for the experimental and statistical analysis of the impact of real-time feedback for crowdworkers in FR-IQA with paired comparisons.
- We have shown the benefits of feedback about the correctness of paired comparisons in reduced response times and showed that feedback did not cause any bias in the JND estimation.
- The user experience was reported as positive; crowdworkers appreciated the feedback.

To the best of our knowledge, this is the first study to investigate the impact of feedback on crowdworkers' performance and experience in subjective image quality assessment.

II. Related work

In 1952, Blackwell proposed and discussed several procedural variables for psychophysical testing, such as JND assessment, including the amount of feedback provided to subjects about the correctness of their responses [8]. His recommendation was that feedback should be provided, and for many years feedback has been routinely applied in various psychophysical tasks [9].

However, we have found only a couple of works in the field of image quality assessment that made use of such feedback in paired comparisons. In 1989, a study tested subjective image fidelity of compressed images for several source images and codecs using the 2-alternative-forced-choice (2AFC) approach very much like we have done in our experiment [10]. For each response of a subject, a voice synthesizer responded with a feedback "correct" or "wrong" appropriately. Similarly, in 2018, the study in [11] used 2AFC with a flicker test to detect the distorted high dynamic range (HDR) image in a pair. Only the incorrect responses were indicated by an auditory feedback of a 100 ms long tone. Neither one of these contributions analysed or discussed the benefits and limitations of the feedback.

Visual quality assessment and detection of distortions in an image or video relative to a reference may improve by perceptual learning during the assessment task. It is plausible to assume that feedback about the correctness of responses will improve the learning rate. However, research findings have been mixed regarding support for this hypothesis. In [12], Vernier acuity was studied in which very small misalignment between two parallel vertical lines needs to be detected. In this task, performance improves with practice. However, there was no clear evidence that feedback significantly boosted the speed of learning. Similarly, mixed results were obtained in a study on the effect of feedback for direction-specific motion discrimination [13]. In contrast, for pattern recognition tasks like 10-

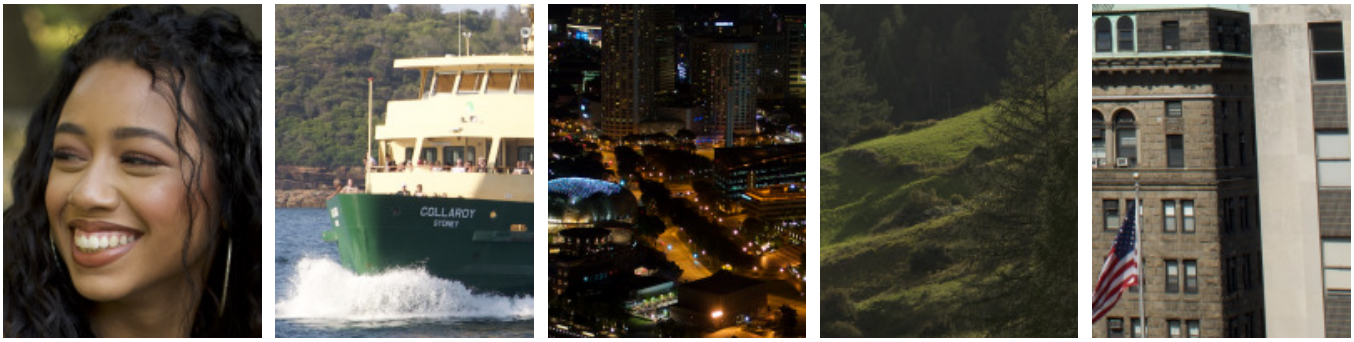


Fig. 1: Source images used in our study. Image names, from left to right, are 00002, 00006, 00007, 00009, and 00010.

alternative forced choice letter identification, the learning effect was much stronger with feedback, see, e.g., [14].

III. Experimental design and methods

A. Source and test images

1) Test images: JPEG AIC-3 [15], [16] provided 10 source images, numbered from 00001 to 00010, with content from different categories, along with compressed versions at 10 distortion levels for each of six selected codecs (JPEG, JPEG 2000, HEVC Intra, VVC Intra, JPEG XL, and AVIF). The distortion levels are approximately at equidistant points on the scale of perceived quality, ranging from 0 to 2.5 JND. These distortion levels were selected based on a pilot study conducted in [16]. For crowdsourced quality assessment, the images had been cropped to a resolution of 620×800 pixels.

For our study, we used crops derived from the five source images shown in Fig. 1 that were compressed by two codecs, VVC Intra and JPEG XL. For ease of notation, we refer to the five image crops as *sources* in the following. This resulted in 10 sequences of 11 cropped images each, starting with the source at distortion level 0 and increasing levels from 1 to 10.

B. Paired comparisons in study questions

We used 2AFC paired comparisons for subjective quality assessment. For each source image and each codec, there is a sequence of 11 images at distortion levels 0 to 10, where 0 refers to the source image. Each compressed image is shown side-by-side with its corresponding source, in random position, left or right. For each comparison, the subject is asked to identify the side of the source image. Altogether, for 5 sources and 2 codecs, we have 100 paired comparisons. We refer to these paired comparisons as study *questions*.

C. Training and trap questions

Participants were trained by 6 comparisons between a compressed image and the corresponding source. The perceptual difference in a pair was imperceptible, noticeable but not strong, and very large in two pairs each, respectively.

To filter out unreliable participants like random clickers, we included 20 so-called *trap questions*. These questions asked to distinguish compressed images having strong distortion from the corresponding sources.

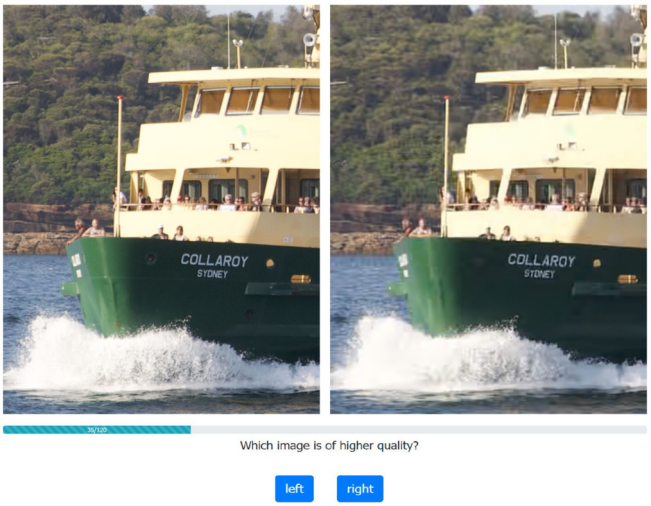


Fig. 2: User interface of the feedback study.

D. Two test conditions with and without feedback

Our study was carried out following the within-subject design. To each of the participants, all of the 126 training, study, and trap questions were presented twice; with and without feedback. Let the test condition with feedback be abbreviated by AFCF, and the other without feedback by AFCN. Note that we provided feedback for the training questions also in condition AFCN.

The feedback was in the form of a text message and an audible beep. For example, if the subject correctly identified the better quality image on the right side, the following text message was displayed: “Yes! The right image is of higher quality.” For the wrong answer the message would have been “No! The right image is of higher quality.” Additionally, one of two distinct beeping sounds accompanied each correct and incorrect response, respectively.

E. Questionnaires

After finishing the responses for each test condition, workers answered a questionnaire consisting of seven items addressing their perceived preference resp. lack of preference for feedback, task clarity, confidence, mental demand, temporal demand,

performance, and frustration. The questionnaire was designed very similar to the NASA-TLX task load questionnaire [17].

F. Crowdsourcing study

We conducted our experiment on the MTurk platform. We posted one human intelligence task (HIT) with 200 assignments for 200 unique crowdworkers. In each assignment, besides the training questions, a worker had to answer 120 study and trap questions in random order, once with feedback and then again in different random order without feedback. The order of the test conditions was also randomized among workers. Requirements for workers to participate were:

- completion at least 500 HITs in previous work on MTurk with a 99% approval rate,
- usage of PC or laptop with screen resolution of at least 1980×1080 pixels to properly fit the web interface with images rendered at 1:1 logical pixel ratio, and
- usage of the Google Chrome browser.
- Stimuli are presented in a 1:1 logical resolution.

At the beginning of the experiment, a brief instruction was shown to the workers, explaining the subjective task and how to submit responses, as well as the payment conditions. Then, the workers had to accept a consent form to continue the experiment. Finally, detailed instructions were provided to the workers for each test condition, with examples of paired comparisons, followed by an explanation of the task load questionnaire. After that, they were allowed to start the experiment with the training session.

For each paired comparison, the two images were displayed for 8 seconds, during which workers could respond at any time to determine which image was of higher quality by pressing either the "left" or "right" button (Fig. 2). If they did not respond within the 8-second display time, the image would be hidden, and a gray page would appear, giving the workers an additional 3 seconds to answer. If the crowdworker failed to answer, the response would be labeled as "undecided". The assignment was accepted and paid for if the number of correctly answered trap questions was at least 32 out of 40.

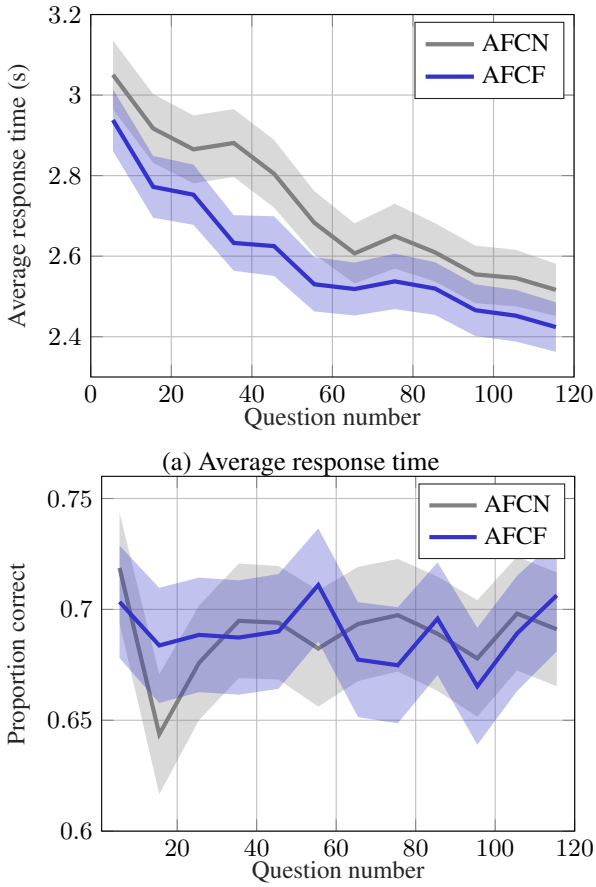
The experimental procedures and protocols used in this study were ethically approved by the Institutional Review Board of the local university.

G. Statistical model for data analysis

We applied maximum likelihood estimation (MLE) to fit a psychometric function $\psi(x)$ to the proportions of correct paired comparison responses. For this purpose we selected the Weibull cumulative distribution function (cdf) $1 - e^{-(x/\lambda)^k}$. The root-mean-square error (RMSE) in a compressed image is taken as the independent variable x. As the RMSE tends to zero, subjects can only guess the answer correctly with a probability of one half. Therefore, the function is scaled to ensure that for the resulting psychometric function, we have $\psi(0; \lambda, k) = 0.5$:

$$\psi(x; \lambda, k) = \frac{1}{2} + \frac{1}{2}\left(1 - e^{-(x/\lambda)^k}\right). \quad (1)$$

As in common practice, the JND is reported at the value x for which the model yields a 75% proportion correct, i.e., $\psi(x; \lambda, k) = 0.75$ [7], [18], [19]. The standard deviation σ of the Weibull function serves as an estimate of the precision of the JND assessment.



(a) Average response time

(b) Proportion of correct responses

Fig. 3: The plot shows (a) the average response time and (b) the proportion of correct paired comparisons plotted against their order in the subjective experiment. The shaded areas indicate the 95% confidence intervals. The averages shown here were computed over all subjects and over 10 successive paired comparisons each, with trap questions removed.

To determine the confidence intervals (CI) of the JND estimate, its precision, and other features, non-parametric bootstrapping with 1000 trials was used; the 149 subjects were sampled with replacement. In figures we show the 95% confidence intervals from bootstrapping.

IV. Experimental results

In both test conditions together, the crowdworkers responded to 50,400 paired comparisons and provided 2,800 answers to the task load questions. The response time for each paired comparison was also recorded. On average, each worker spent 20 minutes to complete the experiment. Approximately 9 minutes were allocated to answer the 252 train, study and trap questions. The rest of the time was spent reading the instructions and answering the questionnaires. Of 200 participants, 55.5% (111

participants) were male, and 44.5% (89 participants) were female. No country filtering was applied, and most of the workers were from the US. This section presents the filtering of unreliable participants and the data analysis for our hypotheses.

A. Data cleansing

We used two criteria to discard data from unreliable participants. First, the responses of participants who answered incorrectly more than eight trap questions (more than 20% of the trap questions) were discarded. Second, the responses of participants who skipped four or more questions from the 240 (study and trap) questions in both test conditions were disregarded. Consequently, 51 participants were excluded, and their responses were not considered in the analysis. The analysis was conducted based on the responses of the remaining 149 participants. In order to maintain a strict within-subject design, if a participant did not answer a study question in one test condition, the corresponding question from the other test condition was excluded.

B. Accuracy

1) Proportion of correct responses: The proportion was computed for all paired comparisons. Their averages and confidence intervals with and without feedback are:

AFCF: 0.688	CI: [0.668, 0.707]
AFCN: 0.689	CI: [0.670, 0.707]

This result indicates no significant difference in the proportion of correct responses between the two test conditions.

2) Correlation with ground truth: For each source, both codecs and both test conditions, we computed the KRCC between the proportions correct and the RMSE of the corresponding compressed images. With larger RMSE, distortions are more easily detected, and so the proportion of correct responses in a pair comparison should be larger as well. Their averages and confidence intervals with and without feedback are:

AFCF: 0.699	CI [0.634, 0.761]
AFCN: 0.705	CI [0.642, 0.770]

This and the results in the Fig 5 (d) indicate no significant difference in the correlation with distortion magnitude between the two test conditions.

C. Response time

Fig. 3(a) illustrates how the average response time varies as subjects progress through the 120 randomized study and trap questions, with or without feedback. The response time is calculated only for the study questions. We note that the response time decreases by about 0.5 seconds in both conditions. The average response time is 2.598 s with feedback and 2.723 s without. The confidence intervals overlap only little, and the p-value is 0.0005 indicating the response time with feedback is significantly smaller. However, the effect size is small (Cohen's $d = 0.1$).

D. Learning effect

Similarly, Fig. 3(b) shows how the proportion of correct responses develops over time in the experiment. In this comparison, no significant difference between the conditions with and without feedback can be seen. There is no learning effect in either case. The figure indicates a lower proportion of correct responses for batch 2 compared to others. The two-proportion Z-test was conducted to compare these proportions, yielding a Z-score of 0.45 and a p-value of 0.65, suggesting no significant difference.

E. User experience

We used a Wilcoxon matched-pairs signed ranks test to analyze the questionnaire scores for each sub-scale. Only the analysis of the "confidence" sub-scale revealed a statistically significant difference between AFCN and AFCF ($z = -1.6672$, $p < .05$), indicating higher median values for confidence in AFCN (59.57) compared to AFCF (56.74). The question for this subscale was: "How did your confidence change in making correct image quality judgments as the experiment progressed?". The response was on a continuous analog scale ranging from -100 (strongly decreased) to 100 (strongly increased). The smaller confidence with feedback may be caused by often having to guess the correct response and then receiving negative feedback half of the time (in AFCF).

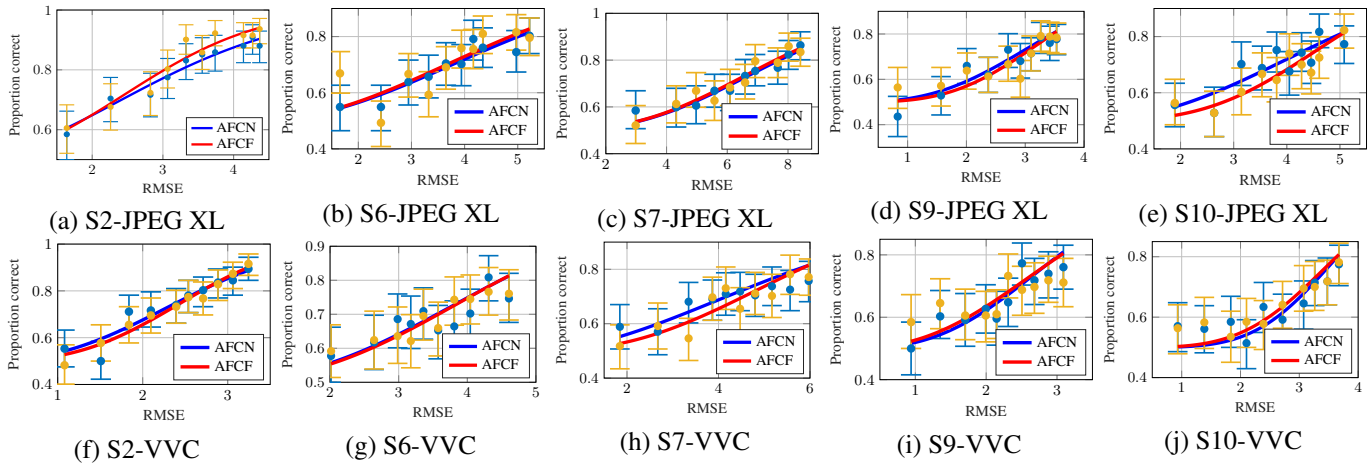
In a multiple-choice question at the end of the HIT, crowdworkers informed us about which feedback they preferred. Of the 149 accepted workers, 90 selected "Only text message", 39 opted for "Text message and audible (beep)", 15 chose "Only audible (beep)", and only 5 crowdworkers selected the response option "None of them". Hence, 97% of the workers preferred receiving feedback.

F. Goodness of fit

The goodness of fit assesses the accuracy of predicting the proportion of correct responses by the fitted psychometric for the function of each model. These functions are shown in Fig. 4. A measure of their accuracy is the negative log-likelihood (NLL). The NLL averaged over all ten models for the AFCN test condition is 5.897, and for the AFCF test condition, it is 5.931, almost the same. Fig. 5 (a) gives the NLL and corresponding CIs for the ten models with and without feedback. The CIs of the NLL for one condition include the mean NLL for the other condition, except for the first model (S2_JPEG XL).

G. Assessment of JND

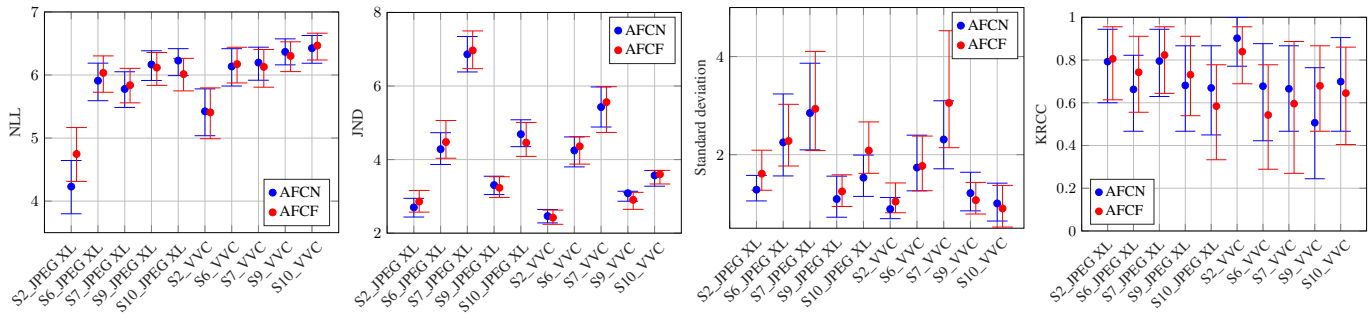
From all fitted Weibull distributions we extracted the JND as explained in Subsec. III-G; see Fig. 5(b) for the results, averaged over 1000 bootstrap samples and with 95% CIs. Clearly, the CIs of the JNDs for AFCN and AFCF largely overlap, indicating that feedback does not influence the JND assessment. Thus, our Null Hypothesis 1 is not rejected by means of the data from our experiment.



(a) S2-JPEG XL (b) S6-JPEG XL (c) S7-JPEG XL (d) S9-JPEG XL (e) S10-JPEG XL

(f) S2-VVC (g) S6-VVC (h) S7-VVC (i) S9-VVC (j) S10-VVC

Fig. 4: Shows the Weibull psychometric functions fitted to the proportion correct of subject responses for each combination of image source and codec. The proportion of correct responses is shown with 95% CIs calculated over 100 bootstrapping runs.



(a) Negative-log-likelihood (b) JND (c) Standard deviation (d) KRCC

Fig. 5: Shows (a) the negative-log-likelihood (NLL), (b) the JND, and (c) the standard deviation of the estimated parameters of the psychometric functions with 95% CI. In (d) the KRCC values between the proportions correct and RMSE values are plotted.

H. Precision of JND assessment

The standard deviations (std) of the Weibull distributions in the fitted psychometric functions serve as a proxy for the precision of the corresponding JND estimates. Fig. 5(c) shows the averages across the same 1000 bootstrap samples. For some of the models, the precision is better with feedback, and for the rest, it is better without. In addition, their CIs strongly overlap, indicating no significant difference in the precision of the fitted psychometric function under the two test conditions. Thus, our Hypothesis 2 has to be rejected; feedback did not increase the precision of the JND assessment.

V. Conclusions and Future Work

This study examined the effects of including feedback to crowdworkers about the correctness of their responses for paired comparisons in full-reference image quality assessment. From the data obtained from 149 crowdworkers that were accepted as reliable, we conclude that the feedback reduced response times and did not cause any bias in the JND estimation. Moreover, almost all crowdworkers confirmed in writing that they appreciated having feedback. The data did not reveal any benefit or harm of feedback regarding the accuracy of responses, increased learning effect for the task, goodness of fit of psychometric models, and precision of JND estimation.

In our future work, an analysis on the impact of the order of the test conditions will be done. Also, we will implement incentive mechanisms, including gamification to keep crowdworkers happy, improve productivity, and reduce stress. The results may help large-scale crowdsourcing studies to better assess JND-based image and video quality.

Future research could also extend this to other subjective quality assessment tasks such as single stimulus, multiple stimulus, or no-reference methods. This could add complexities, however. For example, single stimulus or no-reference methods may have a less clear feedback basis.

We provided feedback via text and sound. Exploring more engaging feedback forms, such as visual, emotional, or gamified feedback, could enhance crowdworker motivation and task performance.

Lastly, individual differences in feedback reception could be considered. Feedback may influence different individuals variably. Understanding these differences could enable tailored feedback strategies, potentially improving performance and user experience.

REFERENCES

[1] Tobias Hoßfeld, Christian Keimel, Matthias Hirth, Bruno Gardlo, Julian Habigt, Klaus Diepold, and Phuoc Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[2] Flávio Ribeiro, Dinei Florencio, and Vítor Nascimento, "Crowdsourcing subjective image quality evaluation," in *IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 3097–3100.

[3] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic, "An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets," in *Proceedings International AAAI Conference on Web and Social Media*, 2011, vol. 5, pp. 321–328.

[4] Ali Ak, Andreas Pastor, and Patrick Le Callet, "From just noticeable differences to image quality," in *Proceedings 2nd Workshop on Quality of Experience in Visual Multimedia Applications*, 2022, pp. 23–28.

[5] Volker Walter, Michael Kölle, and David Collmar, "A gamification approach for the improvement of paid crowd-based labelling of geospatial data," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-4-2022, pp. 13–120, 2022.

[6] Joe Yuchieh Lin, Lina Jin, Sudeng Hu, Ioannis Katsavounidis, Zhi Li, Anne Aaron, and C-C Jay Kuo, "Experimental design and analysis of JND test on coded image/video," in *Applications of Digital Image Processing XXXVIII*. SPIE, 2015, vol. 9599, pp. 324–334.

[7] Mohsen Jenadeleh, Raouf Hamzaoui, Ulf-Dietrich Reips, and Dietmar Saupe, "Crowdsourced estimation of collective just noticeable difference for compressed video with the flicker test and QUEST+," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–17, 2024, Early Access.

[8] H Richard Blackwell, "Studies of psychophysical methods for measuring visual thresholds," *Journal of the Optical Society of America*, vol. 42, no. 9, pp. 606–616, 1952.

[9] Donald G Jamieson and William M Petrusic, "On a bias induced by the provision of feedback in psychophysical experiments," *Acta Psychologica*, vol. 40, no. 3, pp. 199–206, 1976.

[10] Charles S Stein, Andrew B Watson, and Lewis E Hitchner, "Psychophysical rating of image compression techniques," in *Applied Vision*. Optica Publishing Group, 1989, pp. 76–80.

[11] Aishwarya Sudhama, Matthew D Cutone, Yuqian Hou, James Goel, Dale Stolitzka, Natan Jacobson, Robert S Allison, and Laurie M Wilcox, "85-1: Visually lossless compression of high dynamic range images: A large-scale evaluation," in *SID Symposium Digest of Technical Papers*. Wiley Online Library, 2018, vol. 49, pp. 1151–1154.

[12] Manfred Fahle and Shimon Edelman, "Long-term learning in vernier acuity: Effects of stimulus orientation, range and of feedback," *Vision Research*, vol. 33, no. 3, pp. 397–412, 1993.

[13] Karlene Ball and Robert Sekuler, "Direction-specific improvement in motion discrimination," *Vision Research*, vol. 27, no. 6, pp. 953–965, 1987.

[14] Jiajuan Liu, Zhong-lin Lu, and Barbara Dosher, "Similar perceptual learning in 10-alternative letter identification in external noise with and without feedback supervision," *Journal of Vision*, vol. 20, no. 11, pp. 1237–1237, 2020.

[15] ISO/IEC JTC1/SC29/WG1 N100311, REQ, "Common test conditions on subjective image quality qssessment," 2022, https://jpeg.org/aic/documentation.html.

[16] Michela Testolina, Vlad Hosu, Mohsen Jenadeleh, Davi Lazzarotto, Dietmar Saupe, and Touradj Ebrahimi, "JPEG AIC-3 Dataset: Towards defining the high quality to nearly visually lossless quality range," in *15th International Conference on Quality of Multimedia Experience (QoMEX)*, 2023, pp. 55–60.

[17] Sandra G Hart and Lowell E Staveland, "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Advances in Psychology*, vol. 52, pp. 139–183. Elsevier, 1988.

[18] Andrew B Watson, "QUEST+: A general multidimensional bayesian adaptive psychometric method," *Journal of Vision*, vol. 17, no. 3, pp. 1–27, 2017.

[19] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, et al., "Videoset: A large-scale compressed video quality dataset based on jnd measurement," *Journal of Visual Communication and Image Representation*, vol. 46, pp. 292–302, 2017.