https://doi.org/10.2352/EI.2024.36.11.HVEI-222

# Study on Effect of Display Layout for Web Conferencing Services on Subjective Evaluation Stability

*Kimiko Kawashima, Yuichiro Urata, Noritsugu Egi, Kazuhisa Yamagishi; Nippon Telegraph and Telephone Corporation (NTT); Tokyo, Japan*

## Abstract

*The quality of web conferencing services often degrades by network quality. Parametric video quality-estimation techniques are essential to detect quality degradation because they can estimate the quality of videos displayed to each user by using information about video encoding. To study these techniques, we have to consider display layouts such as single and grid views, which are layouts unique to web conferencing videos. Therefore, we investigated the subjective evaluation stability of both views. Then, we showed the evaluation of grid views is less stable because of the smaller size of face images comprising the display and wider quality distributions.*

## Introduction

The use of web conferencing services has increased dramatically due to the COVID-19 Pandemic [1, 2]. Since the quality of web conferencing services is affected by network quality (e.g., throughput and packet loss), which varies depending on the time of day and the line in use, perceived quality is degraded. To detect such problems, parametric quality-estimation techniques are essential because they can assess video quality by using information about video encoding, such as bitrate and resolution, that can be acquired during service provision. To establish a quality-estimation technique, subjective quality evaluation data need to be derived by using a large number of experimental videos. To derive subjective evaluation data, the quality of web conferencing videos needs to be stably evaluated.

Several subjective quality evaluation methods exist, such as the absolute category rating (ACR, [3]), the degradation category rating (DCR, [3, 4]), the double stimulus continuous quality-scale [4], and the pair comparison methods [3]. These methods have different characteristics, and the experimenters select the method on the basis of the evaluation purpose. A five-point ACR method (5: Excellent, 4: Good, 3: Fair, 2: Poor, 1: Bad) is often used in constructing parametric quality-estimation techniques [5, 6, 7, 8].

It is important to assess the effect of display layout on the quality for web conferencing services because web conferencing videos (i.e., their face images) can be displayed as not only a single view but also a grid view [9, 10, 11]. In the single view, one person's face image is displayed; in the grid view, multiple people's face images are reduced in size and integrated into a single display. Therefore, quality degrades uniformly across the entire display in a single view. On the other hand, in the grid view, degradation does not occur uniformly across the entire display because the quality of each participant's face image differs depending on each participant's network environment. This means that factors such as display patterns (i.e., the face image size, the number of faces, and the placement of the face images) and quality distributions (i.e., quality differences in each face image comprising the display) affect subjective quality.

Among subjective quality, we focused on stability, which means the wideness of confidence intervals of subjects' scores. Different participants may pay attention to different parts of the display. In that case, some may notice quality deterioration while others do not. In addition, some may evaluate the quality of web conferencing videos on average, including good- and bad-quality parts of the display, some may evaluate the quality as higher than average due to good-quality parts, and some may evaluate the quality as lower than average due to bad-quality parts. Thus, participants may perceive the quality differently, even when they see the same image. This means that the confidence intervals of subjects' scores may be affected. Therefore, we focused on the stability of subjective evaluation for grid views.

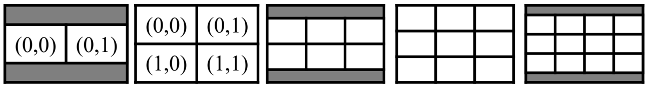
However, no study has investigated the stability of subjective evaluation for grid views. Although grid views showing three face images [12] and four face images [13] have been evaluated, stability has not been discussed. Single views have also been evaluated [14, 15]. Still, there have been no studies on the evaluation of single and grid views in the same experiments, and the difference in stability between the two has not been elucidated.

Therefore, this paper clarifies the stability of subjective evaluation for single and grid views. Then, we describe the effect of display patterns and quality distributions on the evaluation stability for grid views.

## Related Works

To enable the acquisition of stable subjective evaluation data, evaluation characteristics called subject's bias [16, 17] and inconsistency [3, 18] have been widely investigated. A subject's bias means the overall shift between a subject's scores and the actual value. Previous studies [16, 17] have shown that increasing the number of participants minimizes the influence of the subject's bias and makes it possible to obtain highly stable evaluation data. An inconsistency means a subject cannot consistently evaluate videos on the basis of this evaluation standard. ITU-T Recommendation P.910 [3] provides a technique to reject these two factors. This technique can minimize the influence of outliers by changing the weight of inattentive subjects' scores [3, 18].

However, there has been no study on the subjective evaluation stability for single and grid views of web conferencing videos, as described in Section I, because the previous studies [12, 13, 14, 15] focused on constructing parametric quality-estimation techniques. Therefore, some issues need to be addressed to acquire stable subjective evaluation data of grid views, because different participants may pay attention to different parts of the display on the grid view.



(a) 2 faces (b) 4 faces (c) 6 faces (d) 9 faces (e) 12 faces

***Figure 1.*** *Display patterns of the number of face images.*

First, to show the effect that multiple people's face images are reduced in size and integrated into a single display, the stability of subjective evaluation needs to be clarified for single and grid views. Second, to identify factors related to the display layouts unique to web conferencing videos, the effect of display patterns and quality distributions on the evaluation stability for grid views needs to be described. Finally, to acquire stable subjective evaluation data, the number of participants required for stable evaluation needs to be estimated while considering the effect of display patterns and quality distributions.

## Subjective Evaluation Experiment

As described above, to clarify the stability of subjective evaluation for single and grid views, we conducted a subjective evaluation experiment. To solve the issues described above, we prepared evaluation videos with different display patterns and quality distributions for grid views. Display patterns mean various sizes, numbers, placement patterns, and combination patterns of face images. We set various sizes, numbers, and placement patterns to clarify the effect on evaluation stability when multiple people's face images are reduced in size and integrated into a single display. We also set combination patterns of face images to account for variations in the difficulty of encoding. Quality distributions mean quality differences in each face image comprising the display. We set quality distributions to clarify the effect on evaluation stability when the quality does not degrade uniformly across the entire display

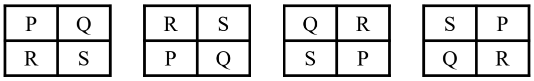
### *Source videos and test conditions*

For both views, we prepared source videos, also called source reference circuits (SRCs). We used face images of web conferencing (1 minute, 1920×1080, 30 fps, I420 yuv format, 8-bit depth). To consider the effect of encoding difficulty, we used various variations in clothing and background (white ("W") and gray or brown wood-grained ("GB")).

In addition, we prepared test conditions, also called hypothetical reference circuits (HRCs). We used four quality states (a stable-quality state ((A)stable) and three fluctuating-quality states) to reflect events that occur in web conferencing services due to variable bitrate encoding and network conditions. We prepared quality fluctuations such as variations when encoding begins ((B)start), when the bitrate increases ((C)up) and when the bitrate decreases ((D)down). For (A) and (B), ten bitrate patterns were prepared: (HRC_(AB)), i.e., 64k, 128k, 256k, 384k, 512k, 640k, 768k, 896k, 1024k, and 2048kbps. For (C) and (D), we set the different bitrates to 0 and 20 seconds. For (C), we prepared three patterns (HRC_(C)), i.e., 384k→896kbps, 256k→768kbps, and 128k→512kbps. For (D), we prepared three patterns (HRC_(D)), i.e., 1024k→128kbps, 768k→128kbps, and 512k→128kbps.

**Table 1: PVSs of grid views**

| Display patterns | | | | Quality distributions | | |
|---|---|---|---|---|---|---|
| Size | Num | Place | Combination | QD1 | QD2 | QD3 |
| 1/4 | 2 | 1 | 4 (W,GB: 2) | 1 | 2 | 0 |
| | 4 | 1 | 1 (W: 1) | 10 | 6 | 6 |
| | | 1 | 1 (GB: 1) | $18^{*s}$ | 32 | 8 |
| | | $4^{*p}$ | 1 (W: 1) | 0 | 0 | $2^{*q}$ |
| | | $4^{*p}$ | 1 (GB: 1) | 0 | $2^{*r}$ | 0 |
| 1/9 | 6 | 1 | 1 (GB: 1) | 4 | 8 | 0 |
| | 9 | 1 | 1 (GB: 1) | 10 | 18 | 8 |
| 1/16 | 12 | 1 | 1 (GB: 1) | 10 | 18 | 8 |



(a) pl1 (b) pl2 (c) pl3 (d) pl4

***Figure 2.*** *Example of placement patterns.*

### *Processed video sequences*

Participants evaluated 208 processed video sequences (PVSs). The 208 PVSs were composed of 16 PVSs of single views and 192 PVSs of grid views in Table 1.

For single views, we used eight SRCs (four people (two males; two females) in two states (speaking and listening)) of "W" and "GB", respectively. For each SRC, we set one HRC in HRC_(AB).

For grid views, each PVS (PVS*i*) consists of small PVSs (PVS*i*(*r*, *c*)) where *i* means the index of PVS, *r* means the row number, and *c* means the column number of the face images in Fig. 1. PVS*i*(*r*, *c*) is created by encoding SRC*i*(*r*, *c*) with HRC*i*(*r*, *c*). The sizes of PVS*i*(*r*, *c*) are 960×540 (size 1/4), 640×360 (1/9), and 480×270 (1/16) in Table 1. There were five face display patterns, as shown in Fig. 1. In each display pattern, one PVS*i*(*r*, *c*) is speaking, and the others are listening. To clarify the effect of placements, we prepared four placement patterns pl1-pl4 in Fig.2 (**p* in Table 1) and set four conditions (**q* and **r* in Table 1) for each placement pattern. These four placement patterns consist of the same four small PVSs, as shown in Fig. 2, and this means each small PVS appears once in all four regions. In addition, to avoid the influence of the combination of face images, we prepared different combination patterns of SRCs for two and four face images.

For quality distributions, we set three ways: QD1, QD2, and QD3. QD1 means that the quality differs slightly among all parts of the display. QD2 means that the quality differs only in one part of the display. QD3 means that the quality differs depending on each part of the display. In addition, we set the quality state for each quality distribution. For QD1, we mainly set QD1_(A) (the quality distribution is QD1, and the quality state is (A)), although the column **s* in Table 1 consisted of 10 patterns of QD1_(A) and 8 patterns of QD1_(B). For QD2, the detailed conditions are shown in Table 2. In Table 2, QD2_(A)_bad and QD2_(B)_bad mean that the quality of only one PVS*i*(*r*, *c*) is bad. In addition, QD2_(A)_good and QD2_(B)_good mean that the quality of only one PVS*i*(*r*, *c*) is good. We used QD2_(C) only where there were 4 face images, because the quality change in a short period (i.e., 10

**Table 2: Detailed conditions of QD2**

| Display patterns | | | Quality States | | | | | |
|---|---|---|---|---|---|---|---|---|
| Num | Place | Combination | QD2_(A)_bad | QD2_(A)_good | QD2_(B)_bad | QD2_(B)_good | QD2_(C) | QD2_(D) |
| 2 | 1 | 4 (W,GB: 2) | 2 | | 0 | | 0 | 0 |
| 4 | 1 | 1 (W: 1) | 3 | 3 | 0 | 0 | 0 | 0 |
| | 1 | 1 (GB: 1) | 2 | 2 | 2 | 2 | 12 | 12 |
| | 4 | 1 (W: 1) | 0 | 0 | 0 | 0 | 0 | 0 |
| | 4 | 1 (GB: 1) | 1 | 1 | 0 | 0 | 0 | 0 |
| 6 | 1 | 1 (GB: 1) | 4 | 4 | 0 | 0 | 0 | 0 |
| 9 | 1 | 1 (GB: 1) | 3 | 3 | 0 | 0 | 0 | 12 |
| 12 | 1 | 1 (GB: 1) | 3 | 3 | 0 | 0 | 0 | 12 |

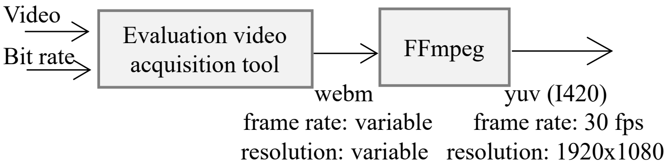


***Figure 3.*** *Flow of acquiring web conferencing video.*

seconds) is difficult to perceive as the speed of increase in quality according to the increase in bitrate is slower than that of the decrease in quality according to the decrease in bitrate. For QD3, we only set QD3_(A), because multiple quality states and quality differences are too difficult to perceive simultaneously.

To represent the quality distributions, HRC was set as follows. For QD1_(A) and QD1_(B), all HRC*i*(*r*, *c*) are the same and were selected from HRC_(AB). For QD2_(A)_bad, QD2_(A)_good, QD2_(B)_bad, and QD2_(B)_good, only one of HRC*i*(*r*, *c*) is different from others. All HRC*i*(*r*, *c*) were selected from HRC_(AB). For HRC*i*(*r*, *c*) of QD2_(C), the HRC for one SRC (i.e., SRC_(C)) was selected from HRC_(C), and the other HRCs were selected from HRC_(AB). Similarly, for HRC*i*(*r*, *c*) of QD2_(D), the HRC for one SRC (i.e., SRC_(D)) was selected from HRC_(D), and the other HRCs were selected from HRC_(AB). Finally, for QD3_(A), three patterns are selected from HRC_(AB).

To summarize the above, each PVS (PVS*i*) consists of small PVSs (PVS*i*(*r*, *c*)). The combination of small PVSs determined the condition of each PVS, that is, numbers, placement patterns, combination patterns of face images, and quality distributions. Therefore, we express the condition of each PVS as con(*i*, pl*j*, *k*, QD*l*_(*m*)_*n*_*o*). Here, the index *i* means the number of face images (*i*=2, 4, 6, 9, 12). The index *j* means the placement patterns (*j*=1, 2, 3, 4) as shown in Fig. 2. The index *k* means combination patterns of face images (*k*=W, GB). The index *l* means the quality distributions (*l*=1, 2, 3). The index *m* means the quality states (*m*=A, B, C, D). The index *n* (*n*=bad, good) indicates the condition in which only one part of the display quality is different from others. The index *o* indicates the number of conditions of quality distributions under the condition that *i*, *j*, *k*, *l*, *m*, and *n* are identical. For example, as shown in Table 2, when *i*=4, *j*=1, *k*=W, *l*=2, *m*=A, and *n*=bad, we express con(4, pl1, W, QD2_(A)_bad_*o*) (*o*=1,2,3).

### *Method to create PVSs*

To reproduce web conferencing videos, we developed an evaluation video acquisition tool using the web conferencing system called Janus WebRTC Server [19] and the Google Chrome browser. Janus WebRTC Server controlled codecs and selected vp8 or vp9. Since the video files saved by Janus WebRTC Server have variable frame rates and resolutions, the frame rate was fixed at 30 fps and the resolution at 1920 × 1080 when decoding as shown in Fig. 3. Pixels from the previous frame were copied using FFmpeg (v4.3.1) for missing frames due to quality fluctuations. When the resolution was reduced due to quality fluctuations, considering that the monitor size is automatically enlarged to 1920 × 1080 when the web conferencing service is actually used, we used FFmpeg to upsample the videos using the bicubic method.

First, we created a PVS*i*(*r*, *c*). We input SRC*i*(*r*, *c*) and HRC*i*(*r*, *c*) into the tool and acquired the single view (PVS*i*(*r*, *c*)_s). Then, we cropped a 10-second segment. For QD1,2,3_(A), we selected the segment that did not contain quality fluctuations in the speaker's PVS*i*(*r*, *c*)_s. For QD1,2_(B), we set the first 10 seconds. For QD2_(C)/(D), we determined the segment when the bitrate increased/decreased in PVS*i*(*r*, *c*)_s of SRC_(C)/(D). Then, we cropped all PVS*i*(*r*, *c*)_s in the same 10-second segment. After that, PVS*i*(*r*, *c*)_s was resized using FFmpeg to match the size of PVS*i*(*r*, *c*), and all of PVS*i*(*r*, *c*) was integrated into PVS*i*.

### *Experimental details*

PVSs were presented on a 23.8-inch PC monitor. One participant evaluated each monitor at a viewing distance of 3H (H: display height). The room luminance was 200 lux.

The experiment was conducted by ITU-T Recommendation P.910 [3]. First, color vision and visual acuity tests were conducted to confirm that all participants had the appropriate color vision and visual acuity for the evaluation. Next, we explained the contents of the evaluation experiment (object to be evaluated, viewpoint of evaluation, input scale of evaluation, and posture during video viewing) to the participants. In the evaluation process, participants were asked to watch a 10-second video and then to rate the video within 5 seconds by using the 5-point ACR method. During the practice session, the participants watched 16 10-second videos and input their ratings to familiarize themselves with the evaluation procedure. In the main test, 4 sessions of 48 PVSs (approximately 20 minutes) were conducted, and 1 session of 16 PVSs (around 5 minutes) was conducted at the end, for a total of 208 PVS. A five-minute break was taken between sessions, and the total time from the time the participants entered the laboratory to the end of the session was approximately two hours. To eliminate the influence of the order of PVS presentation on the evaluation results, we created eight different randomized patterns of the PVS presentation order, and four participants evaluated each randomized pattern.

The participants were 16 males and 16 females aged 18-24, none of whom had professional experience in video quality evaluation. To reduce the impact of the difference in age of the participants on the stability of the evaluation, the participants were recruited within a narrow age range.

## Results

In this section, we show the results of the stability of subjective evaluation for single and grid views, the effect of display patterns and quality distributions on the evaluation stability for grid views, and the number of participants required for stable evaluation considering the effect of display patterns and quality distributions, as described in Sect. II.

### Stability of subjective evaluation for single and grid views

First, the correlation between the Mean Opinion Score (MOS) and each participant's score was calculated to screen the participants. When we set the screening threshold to 0.75, 3 participants were excluded. To use the participants' scores as they are to clarify the stability (i.e., 95% confidence intervals (CIs) of the participants' scores), we determined the screening threshold on the basis of ITU-T Recommendation P.913 Annex A.1 (Screen by PVS) [20]. To clarify the stability, Fig. 4 shows 95% CIs of the 29 participants' scores. The maximum and mean 95% CIs were 0.40 and 0.26, respectively, indicating that the stability is similar to that in previous video quality evaluation experiments for streaming [21].

Next, to compare the stability of the single view and grid view, the relationship between MOS and 95% CIs is shown for each number of face images on a display (Fig. 4). Differences in the plot symbols and approximate curves indicate differences in the number of face images on a display. Fig. 4 shows that the grid views have larger 95% CIs than the single views when the MOS is almost the same, indicating that the evaluation is less stable. To confirm these results, we conducted statistical analysis. We conducted a t-test between the 95% CIs for the single view and the 95% CIs for the grid view with MOS ranging from 2 to 4 because the 95% CIs bear a non-linear relationship with the MOS. As a result, at the significance level of 5%, there was a significant difference between the single view and the grid view with 4 face images, between the single view and the grid view with 9 face images, and between the single view and the grid view with 12 faces images, respectively. These results statistically confirm our result that the grid view is less stable than the single view shown in Fig. 4. Therefore, to clarify the factors that affect the stability of grid views, we analyzed the effects of display patterns and quality distributions.

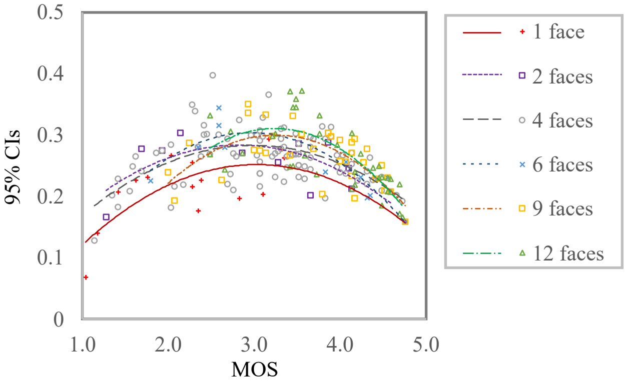


***Figure 4.*** *Comparison of quality distributions of the number of face images on display.*

### Effect of display patterns

#### *Effect of the face image size and the number of faces*

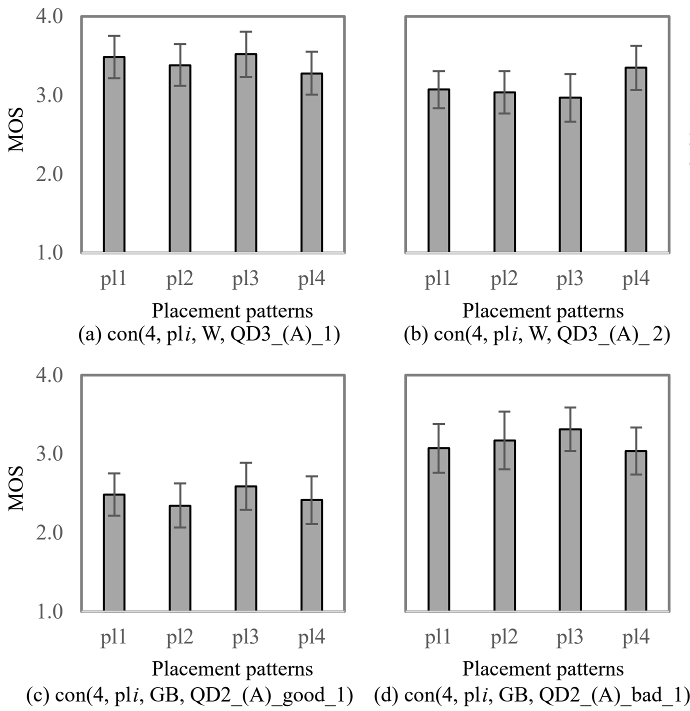
First, we analyzed the effect of the face image size and the number of faces. Fig. 4 shows that the approximate curves for two and four face images on a display almost overlap. This result is thought to be caused by the fact that the face images' size is the same regardless of the number of face images. On the other hand, the stability is lower for 6, 9, and 12 face images on a display than for 2 and 4 face images. This is thought to be because distortion is more difficult to recognize when there are smaller face images and the difference increases between those who notice distortion and those who do not. However, since we used only 12 PVSs for 2 and 6 face images on a display, the effect when the number of face images on the display differs even when the face images' size is the same needs further examination.

#### *Effect of placements*

Second, we analyzed the effect of placements. We analyzed MOS obtained for four conditions (**q* and **r* in Table 1) consisting of the same four small PVSs but with four different placements, as shown in Fig. 2. **q* is consisted of con(4, pl*i*, W, QD3_(A)_1) and con(4, pl*i*, W, QD3_(A)_2) (*i*=1,2,3,4). **r* is consisted of con(4, pl*i*, GB, QD2_(A)_good_1) and con(4, pl*i*, GB, QD2_(A)_bad_1) (*i*=1,2,3,4). Fig. 5 shows the results of four placement patterns for these four conditions. In a variance analysis, the results showed no significant difference between the four different placements in all four conditions at the significance level of 5%, as shown in Fig. 5. This indicates that differences in placement do not affect subjective evaluation in this experiment.

### Effect of quality distributions

We compared the approximate curves of various quality distributions as shown in Fig. 6, although the number of PVSs of each quality distribution was limited, and the plots have variance. In Fig. 6, differences in the plot symbols and approximate curves indicate differences in the quality distributions. For four face images on a display (Fig. 6(a)), comparing QD2_(A)_bad in which one PVS*i*(*r*, *c*) is bad quality and QD2_(A)_good where one PVS*i*(*r*, *c*) is good quality, the approximate curves of QD2_(A)_bad tend to be larger than those of QD2_(A)_good. Comparing QD2_(D) in which one HRC*i*(*r*, *c*) has a decrease in bitrate and QD2_(C) in which one HRC*i*(*r*,



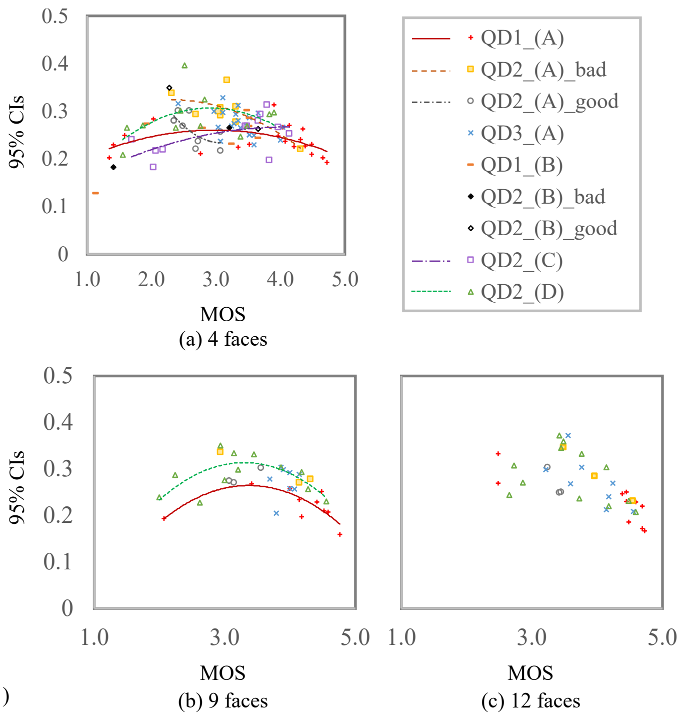
(a) con(4, pl*i*, W, QD3_(A)_1)

(b) con(4, pl*i*, W, QD3_(A)_2)

(c) con(4, pl*i*, GB, QD2_(A)_good_1)

(d) con(4, pl*i*, GB, QD2_(A)_bad_1)

***Figure 5.*** *Effect of the difference in placement patterns of face images on display.*



(a) 4 faces

(b) 9 faces

(c) 12 faces

***Figure 6.*** *Comparison of quality distributions.*

*c*) has an increase in bitrate, the approximate curves of QD2_(D) tend to be larger than those of QD2_(C). In addition, the approximate curves of QD2_(A)_bad and QD2_(D) tend to be larger than those of QD1_(A), in which all HRC*i*(*r*, *c*) are the same. To confirm our results, we conducted a t-test between the 95% CIs for QD2_(A)_bad and the 95% CIs for QD2_(A)_good at the significance level of 5%. There was a significant difference between the two at the significance level of 5%. In addition, there was a significant difference between QD2_(C) and QD2_(D) at the significance level of 5%. This suggests that the stability of the evaluation tends to decrease when quality degrades in one part of the display. Then, for nine face images on a display (Fig. 6(b)), 95% CIs of QD2_(A)_bad tend to be wider than those of QD2_(A)_good around MOS 3.0, although the number of PVSs is limited. The approximate curves of QD2_(D) are larger than those of QD1_(A). These are the same as the results of the four face images. Finally, for 12 face images on a display (Fig. 6(c)), 95% CIs of QD2_(A)_bad tend to be wider than those of QD2_(A)_good around MOS 3.5, although the number of PVSs is limited. However, we cannot compare the approximate curves of QD2_(D) and QD1_(A) because there are no plots from MOS 2.5 to 4.5 in QD1_(A). As described above, we clarified that the stability of the evaluation tends to decrease when quality degrades in one part of the display. However, verification with more PVSs is needed when the number of faces in an image is large; that is, the face images are small.

### *Estimation of the number of participants*

Finally, we estimated the number of participants required for stable evaluation, which will lead to the acquisition of stable evaluation data in the future. For conventional video quality evaluation [22], Kłoda and Ostaszewska investigated the influence of a number of participants on the stability of subjective quality evaluation. They compared the stability of subjective evaluation using the DCR method between 52 participants and 10 participants and then showed that there is a difference around the middle–quality range. Here, we estimated the number of participants who will achieve the same stability in evaluating the grid view as the 15 participants for the single view. The minimum number of participants was defined as 15 in ITU-T Recommendation P.910 [3]. We derived this value by formulating the relationship between the number of participants (from 15 to 28) and the mean 95% CI of MOS ranging from 2 to 4. As shown in [23], we developed this relationship by reducing the number of participants by *N* (*N* = 1, 2, ..., 14) from 29. For each *N*, we randomly selected them in 15 ways. From the results, we estimated that 23 participants are needed to evaluate the grid view as stably as the single view with 15 participants, as shown in Fig. 7.

## Conclusion

In this paper, we conducted experiments to clarify the stability of subjective evaluations for single and grid views, which are layouts unique to web conferencing videos, to acquire stable evaluation data. For grid views, we prepared evaluation videos with different display patterns (the face image size, the number of faces, and the placement of the face images) and quality distributions in the display. As a result, we showed the 95% CIs of the grid view are wider than those of the single view. This means that the grid view is less stable than the single view. In particular, for display patterns, the evaluation stability tends to decrease when the face images are small. In addition, differences in placement do not affect the evaluation. For quality distributions, the evaluation stability tends to decrease when quality degrades in one

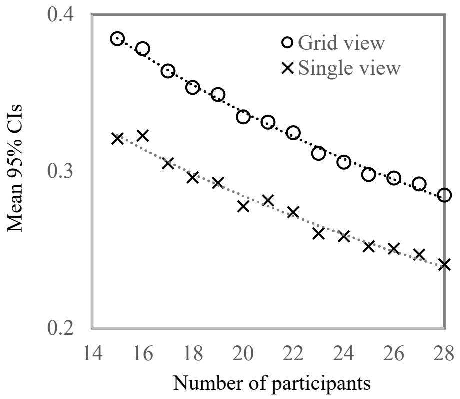


***Figure 7.*** *Estimation of number of participants.*

part of the display. Finally, we estimated the minimum number of participants required to stably evaluate the grid views.

Future work will further investigate how the number of face images on a display affects quality even when the face images are the same size. In addition, the impact of quality distributions when face images are small needs further examination.

## References

[1] A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, C. Dietzel, et al., “Implications of the covid-19 pandemic on the internet traffic,” Broadband Coverage in Germany; 15th ITG-Symposium (IEEE, New York, NY, 2021) pp. 1–5.

[2] S. Liu, P. Schmitt, F. Bronzino, and N. Feamster, “Characterizing service provider response to the covid-19 pandemic in the united states,” Passive and Active Measurement (Springer, Cham, Switzerland, 2021) pp. 20–38.

[3] Rec. ITU-T P.910: Subjective video quality assessment methods for multimedia applications (ITU, Geneva), www.itu.int.

[4] Rec. ITU-R BT.500: Methodology for the subjective assessment of the quality of television pictures (ITU, Geneva), www.itu.int.

[5] Rec. ITU-T G.1070: Opinion model for videotelephony applications (ITU, Geneva), www.itu.int.

[6] Rec. ITU-T P.1203.1: Parametric bitstream-based quality assessment of progressive download and adaptive audio-visual streaming services over reliable transport - video quality estimation module (ITU, Geneva), www.itu.int.

[7] Rec. ITU-T P.1201.1: Parametric non-intrusive assessment of audiovisual media streaming quality - lower resolution application area (ITU, Geneva), www.itu.int.

[8] Rec. ITU-T P.1201.2: Parametric non-intrusive assessment of audiovisual media streaming quality - higher resolution application area (ITU, Geneva), www.itu.int.

[9] zoom Support, “Adjusting your video layout during a virtual meeting,” https://support.zoom.us/hc/en-us/articles/201362323-Adjusting-your-video-layout-during-a-virtual-meeting, accessed Aug. 2023.

[10] webex Help Center, “Switch your view in Webex Meetings, Webex Webinars, and Webex Events (classic),” https://help.webex.com/en-us/article/dy3xzq/Switch-your-view-in-Webex-Meetings,-Webex-Webinars,-and-Webex-Events-(classic), accessed Aug. 2023.

[11] Google Meet Help, “Learn about the Meet layout for your computer,” https://support.google.com/meet/answer/10550593?hl=en-GB, accessed Aug. 2023.

[12] D. Vucic and L. Skorin-Kapov, “The impact of packet loss and google congestion control on qoe for webrtc-based mobile multiparty audiovisual telemeetings,” MultiMedia Modeling (Springer, Cham, Switzerland, 2019) pp. 459–470.

[13] M. Schmitt, J. Redi, D. Bulterman, and P. S. Cesar, “Towards individual qoe for multiparty videoconferencing,” IEEE Trans. Multimedia, 20: 1781 (2018).

[14] Z. Wang, Y. Liu, Y. Li, H. Yang, and D. Yang, “An estimated qoe model for video telephone service,” IC-NIDC (IEEE, New York, NY, 2016) pp. 273–278.

[15] S. Jana, A. Chan, A. Pande, and P. Mohapatra, “Qoe prediction model for mobile video telephony,” Multimed Tools Appl., 75: 7957 (2016).

[16] L. Janowski and M. Pinson, “The accuracy of subjects in a quality experiment: A theoretical subject model,” IEEE Trans. Multimedia, 17: 2210 (2015).

[17] J.-S. Lee, “On designing paired comparison experiments for subjective multimedia quality assessment,” IEEE Trans. Multimedia, 16: 564 (2014).

[18] Z. Li, C. Bampis, L. Janowski, and I. Katsavounidis, “A simple model for subject behavior in subjective experiments,” EI (IS&T, Springfield, VA, 2020), pp. 131-1—131-13.

[19] meetecho, ”janus-gateway,” https://github.com/meetecho/janus-gateway, accessed Aug. 2023.

[20] Rec. ITU-T P.913: Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment (ITU, Geneva), www.itu.int.

[21] K. Yamagishi, N. Egi, N. Yoshimura, and P. Lebreton, “Derivation procedure of coefficients of metadata-based model for adaptive bitrate streaming services,” IEICE Trans. Commun., 104: 725 (2021).

[22] R. Kłoda and A. Ostaszewska, “Subjective video quality evaluation: an influence of a number of subjects on the measurement stability,” Recent Advances in Mechatronics (Springer, Berlin, Heidelberg, 2007) pp. 611–615.

[23] K. Kawashima, J. Okamoto, and T. Hayashi, “Verification on stability and reproducibility of dscqs method for assessing 4k ultra-hd video quality,” QoMEX (IEEE, New York, NY, 2014) pp. 214–219.

## Author Biography

*Kimiko Kawashima received her B.E. and M.E. degrees in engineering from Keio University in 2008 and 2010. She joined NTT Laboratories, Tokyo, Japan in 2010, she is researching video quality assessment.*

*Yuichiro Urata received his B.E. and M.E. degrees in Engineering from University of Electro-Communications, Tokyo, Japan in 2009 and 2011. Since then he has worked in NTT Laboratories, Tokyo. His work has focused on the Quality of Experience for videos.*

*Noritsugu Egi received his B.E. and M.E. degrees in Electrical Communication Engineering from Tohoku University, Japan in 2003 and 2005. He joined NTT Laboratories, Tokyo, Japan, in 2005. Currently, he is researching speech and audio quality assessment.*

*Kazuhisa Yamagishi received his B.E. degree in Electrical Engineering from the Tokyo University of Science, Japan, in 2001 and his M.E. and Ph.D. degrees in Electronics, information, and Communication Engineering from Waseda University, Japan, in 2003 and 2013. Since joining NTT Laboratories in 2003, he has been engaged in the development of objective quality-estimation models for multi-media telecommunications.*