



<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *15th International Conference on Signal Processing and Communication Systems, ICSPCS 2021, Virtual, Online, 13 December 2021 through 15 December 2021*.

Citation for the original published paper:

Elwardy, M., Zepernick, H-J., Chu, T M., Hu, Y. (2021)
On the Opinion Score Consistency in Repeated 360° Video Quality Assessment for Standing and Seated Viewing on Head-Mounted Displays
In: Wysocki T.A., Wysocki B.J. (ed.), *2021 15th International Conference on Signal Processing and Communication Systems, ICSPCS 2021 - Proceedings* Institute of Electrical and Electronics Engineers Inc.
<https://doi.org/10.1109/ICSPCS53099.2021.9660331>

N.B. When citing this work, cite the original published paper.

©2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:bth-22674>

On the Opinion Score Consistency in Repeated 360° Video Quality Assessment for Standing and Seated Viewing on Head-Mounted Displays

Majed Elwardy, Hans-Jürgen Zepernick, Thi My Chinh Chu, and Yan Hu
Blekinge Institute of Technology, SE-371 79 Karlskrona, Sweden
E-mail: {mew, hjz, cch, yhx}@bth.se

Abstract—The development of immersive media systems and services relies on subjective tests that provide ground truth on the quality of experience as perceived by humans. In this paper, we investigate the opinion score consistency through a repeated subjective 360° video quality assessment experiment. The test stimuli are presented on a head-mounted display with participants given the task to rate the video quality in three repeated sessions for both standing and seated viewing. The statistical analysis of the data from the subjective tests aims at revealing if participants change their rating behavior over time or keep their opinion scores given to the test stimuli in standing and seated viewing consistent in a statistical sense. The experimental results are reported in terms of histograms of opinion scores, skewness and kurtosis of opinion scores, mean opinion scores, standard deviation of opinion scores (SOS), SOS fitting functions, and analysis of variance tests. The statistical analysis supports the conjecture that each participant has its own but consistent rating behavior throughout the respective sessions for standing and seated viewing. The finding that the quality rating behavior of the individual participant does not fundamentally change over time may, e.g., assist in scheduling subjective tests under pandemic conditions where experimental campaigns may need to be stalled for an unknown period of time.

Index Terms—360° videos, subjective tests, mean opinion score, standing viewing, seated viewing, head-mounted display.

I. INTRODUCTION

In recent years, an increasing demand on immersive media technology such as virtual reality (VR) and augmented reality (AR) devices and applications has been observed. Viewing 360° videos with the support of head-mounted displays (HMDs) offers immersive experiences which has contributed to the increased popularity of this type of immersive media. Further, 360° videos viewed on an HMD allows 3+ degrees of freedom (3DoF+) [1], i.e., three unlimited rotational movements around the x , y , and z axes, and limited translational head movements along these axes. The development of related stand-alone and networked immersive media systems and applications relies on conducting subjective tests which provide ground truth on the quality of experience as perceived by humans. In addition, given the 3DoF+ of movements supported by HMDs compared to viewing conventional videos on standard displays, immersive media may be viewed not only while seated on a chair but may also be freely explored while standing.

Regarding subjective tests on quality assessment of 360° videos on HMDs, these have often been conducted using methodologies that were recommended for conventional videos [2]–[4]. Subjective test methodologies focusing on 360° videos on HMD have recently been proposed in [5]. Apart from conventional data analysis of opinion scores such as the calculation of mean opinion scores (MOSs) [6], analysis of exploration behavior is also recommended including head movements and eye tracking data. However, participants are expected to sit on a chair during the subjective tests while the option of standing viewing is not covered in this recommendation.

Another research avenue pursued in the field of immersive media focuses on viewing conditions such as standing viewing, seated viewing, and locomotion interfaces. The impact on user’s experience in VR under different viewing conditions is typically studied in terms of immersion, presence, and efficient locomotion in VR. In [7], a motorized swivel chair was introduced to improve the immersive experience and to reduce motion sickness when watching 360° videos on HMDs. The work reported in [8] investigated the effect of fixed, half-swivel, and full-swivel chairs on users’ viewing experience of 360° videos on an HMD. The participants in the related subjective test reported better viewing experiences for half-swivel and full-swivel chair compared to fixed chair in terms of exploration, spatial awareness, and concerns of missing something for certain videos. In [9], a mixed-method experiment using four locomotion modes with different amounts of translational cues and control was conducted to assess task performance, task load, and simulator sickness. The research in [10] explored whether seated users in VR could be given the sensation of standing, walking, or even running. Among others, it was pointed out that hybrid positions where users can choose to sit or stand up may be considered in the design of VR applications if feasible. In [11], to stimulate research on different viewing conditions in VR, a systematic classification of the advantages and disadvantages between sitting and standing user interfaces in VR was provided.

In [12], the effect that standing and seated viewing of 360° videos on an HMD on subjective quality assessment was compared. The statistical analysis of the data gathered in this pilot study was conducted in terms of average rating times, MOS, standard deviation of opinion scores, head movements,

pupil diameter, galvanic skin response, and simulator sickness scores. A comparison of head movements in repeated 360° video quality assessment for standing and seated viewing on HMDs was given in [13]. The results deduced from the repeated subjective tests led to the conjecture that the participants have their own distinct exploration behavior for standing viewing which becomes less different for seated viewing. Further, head movements appear to be higher in standing viewing compared to seated viewing. However, a reduction in the rotational exploration of the 360° videos was observed in the repetitions of the quality assessment task for both viewing conditions.

Motivated by all of the above, this paper focuses on studying the opinion score consistency in repeated subjective tests in which 360° video quality assessment tasks were performed in standing and seated viewing on an HMD. In particular, given the tendency of reducing the rotational exploration in the repetitions of the quality assessment task reported in [13], the question arises whether changes would apply to the consistency of opinion scores given by the participants in the repeated quality assessment task. As such, the main research questions pursued in this paper are as follows:

- Do participants change their rating behavior over time or do they keep their opinion scores given to the test stimuli in standing and seated viewing consistent irrespective of when in time the subjective test is conducted?
- Do participants have the same generic rating behavior or does each participant has its distinct rating behavior?

Apart from supporting the concept of MOS in the quality assessment of 360° videos, i.e., the conventional averaging over opinion scores given by a panel of participants to a set of test stimuli in standing and seated viewing, answering the above research questions may also assist in conducting and scheduling subjective tests under pandemic conditions. For example, test campaigns may need to be stalled depending on the pandemic situation and be continued later which requires that quality rating behavior of participants would not fundamentally change over time. In view of the current pandemic situation, to keep hygiene related concerns to a minimum, two participants, referred to as P1 and P2, were engaged in the repeated subjective tests as in [13]. Following the suggestion of [14] to consider time averages instead of ensemble averages, each participant assessed a sequence of a large number of test stimuli in standing and seated viewing, i.e., a total of 720 different 360° videos covering a wide range of qualities, rather than many participants assessing a small number of test stimuli. This approach is also in line with the suggestion in [15] to consider in-person experiments with a low sample size during pandemic situations that reveal new research avenues and indicate the need for follow-up experiments with a larger sample size. The main contributions of this paper are summarized as follows:

- A statistical analysis of the opinion scores obtained from a repeated subjective test on 360° video quality assessment for standing and seated viewing on an HMD

is provided. A total of six sessions were conducted, i.e., three sessions for standing viewing and three sessions for seated viewing.

- The session repetition schedule of the subjective test for both viewing conditions considers long breaks of several months and short breaks of hours or a day between sessions. This allows providing preliminary insights on whether the rating behavior of the participants changes over time.
- Experimental results are provided as histograms of opinion scores, skewness and kurtosis of opinion scores, analysis of variance (ANOVA) tests among sets of opinion scores, standard deviations of opinion scores, and MOS of 360° videos covering a wide range of qualities in terms of content, resolution, and quantization.
- The statistical analysis of the opinion scores from the repeated quality assessment for both viewing conditions is conducted for each participant and each session, each participant and averaged over all sessions, and averaged over all participants and all sessions.

The remainder of this paper is organized as follows. Section II describes the experimental design of the subjective tests. The statistical measures used to analyse the opinion scores gathered in the subjective tests are introduced in Section III. A detailed statistical analysis of the data from the subjective tests is provided in Section IV. Conclusions are given in Section V.

II. EXPERIMENTAL DESIGN

The experimental setup uses our common platform developed for assessing the subjective quality of immersive media. Comprehensive details about the test stimuli, test methods, software suite, and technical equipment of the developed platform can be found in [16]. In the following, a summary of the experimental design is provided as needed for understanding the reported work.

Four natural scenes of 10s duration each with different resolutions (2K, optimal resolution (OR) [17], 4K, 6K, 8K) and six different quantization parameters (QPs) including the reference videos were used to establish a set of 120 different 360° videos covering a wide range of visual qualities. These test stimuli were presented on an HTC Vive Pro HMD for free exploration in standing viewing and seated viewing on a fixed chair. Each participant attended three sessions for both standing and seated viewing with the task to rate the 360° videos on a five-level quality scale according to the absolute category rating (ACR) method [5]. A session lasted around 30 minutes depending on the actual time required by the participant to cast a quality score for each test stimulus. To study the consistency of opinion scores given by the participants to the test stimuli at different times, the repeated 360° video quality assessment experiment accounted for a long break between Session 1 (S1) and Session 2 (S2), and a short break between S2 and Session 3 (S3) (see Fig. 1).

To keep hygiene related concerns to a minimum, we have engaged two participants, referred to as P1 and P2, in the repeated sessions. Each participant viewed a large number



Fig. 1. Session repetition schedule (ST: standing, SE: seated) [13].

of test stimuli, i.e., 360° videos over three sessions for both standing and seated viewing (a total of 720 360° videos) with the same set of 120 test stimuli shown in each session but in random order. The two male participants were 60 (P1) and 31 (P2) years of age, academic staff, and familiar with the ACR method. It should be mentioned that the approach of engaging a small number of participants assessing a larger number of test stimuli follows the alternative experimental design proposed in [14] as an option to conduct subjective tests under pandemic conditions.

III. STATISTICAL MEASURES

A. Mean Opinion Score

In the considered context, the MOS over the different 360° video scenes shall be defined as

$$\mu_j^{(n)} = \frac{1}{K} \sum_{k=1}^K u_{ijk}^{(n)} \quad (1)$$

where $u_{ijk}^{(n)}$ denotes the opinion score given by participant $n \in \{1, \dots, N\}$ in session $i \in \{1, \dots, I\}$ to test case $j \in \{1, \dots, J\}$ of 360° video scene $k \in \{1, \dots, K\}$. Here, the term ‘test case’ refers to a distinct resolution-QP pair of a given 360° video.

Similarly, the average MOS over both the different sessions for a given viewing condition and the different 360° video scenes can be formulated with (1) as

$$\mu_j = \frac{1}{I} \sum_{i=1}^I \mu_j^{(n)} \quad (2)$$

An additional averaging of opinion scores may be performed over the number of participants which results with (2) in the following average MOS:

$$\mu_j = \frac{1}{N} \sum_{n=1}^N \mu_j^{(n)} \quad (3)$$

In the repeated subjective tests described in Section II, $N = 2$ participants took part, $K = 4$ different 360° video scenes were considered, and $I = 3$ sessions were conducted for both standing and seated viewing. In each of the six sessions, the participants watched $J = 30$ test cases for each 360° video scene. Accordingly, $K \times J = 120$ test stimuli were shown in each session and $K \times J \times I = 360$ test stimuli were shown throughout the respective three sessions for standing and seated viewing. This results in a total of 720 test stimuli that were shown to each participant altogether.

B. Standard Deviation of Opinion Scores

In this work, the standard deviation (SD) is used to measure the amount of dispersion of a set of opinion scores with respect

to the MOS in (1), (2) and (3). As such, a lower SD indicates a more confident rating while a higher SD relates to a less confident rating. The SD with respect to the MOS in (1) is calculated as

$$\sigma_{ij}^{(n)} = \sqrt{\frac{1}{K-1} \sum_{k=1}^K [u_{ijk}^{(n)} - \mu_{ij}^{(n)}]^2} \quad (4)$$

The SD with respect to the average MOS in (2) and (3), respectively, can be formulated as

$$\sigma_j^{(n)} = \sqrt{\frac{1}{IK-1} \sum_{k=1}^K \sum_{i=1}^I [u_{ijk}^{(n)} - \mu_j^{(n)}]^2} \quad (5)$$

$$\sigma_j = \sqrt{\frac{1}{IKN-1} \sum_{n=1}^N \sum_{k=1}^K \sum_{i=1}^I [u_{ijk}^{(n)} - \mu_j]^2} \quad (6)$$

C. Standard Deviation of Opinion Score Hypothesis

To assess and model the diversity of subjective ratings on a five-level quality scale, the SD of opinion scores (SOS) hypothesis has been proposed in [18], [19]. These works suggest the following SOS fitting function for modeling the relationship between MOS and SOS:

$$SOS^2(x) = -ax^2 + 6ax - 5a = a(-x^2 + 6x - 5) \quad (7)$$

where a denotes the SOS parameter and variable x represents the MOS. As the ACR method selected to rate the quality of the 360° videos uses a five-level quality scale, the SOS fitting function is also applied in the next section for the statistical analysis of the experimental results.

IV. EXPERIMENTAL RESULTS

This section provides a statistical analysis of the opinion scores gathered in the repeated subjective tests on 360° video quality for standing and seated viewing. In particular, histograms of the opinion scores given by the two participants in the different sessions are presented along with the skewness and the kurtosis of the related opinion score distributions. The dispersion of the opinion scores around the MOS is assessed in terms of the SD and the SOS fitting function to the respective data points. Further, MOS progressions versus QPs are presented for different resolutions.

A. Histograms of Opinion Scores

Figs. 2(a)-(f) show the histograms of opinion scores given by the two participants to the wide range of test stimuli in each of the three sessions in standing and seated viewing. In addition, Kernel distributions are provided in these figures as fit to the data conveyed by the histograms. In these fits, normal distributions are used as nonparametric kernel smoothing functions along with a bandwidth value of 0.6028. Regarding the results for standing viewing in Figs. 2(a)-(c), it is observed that each participant follows its own but consistent quality rating of the test stimuli throughout the three sessions. This finding is supported by the respective fits to the histograms which more clearly reveal the consistent quality rating of each participant.

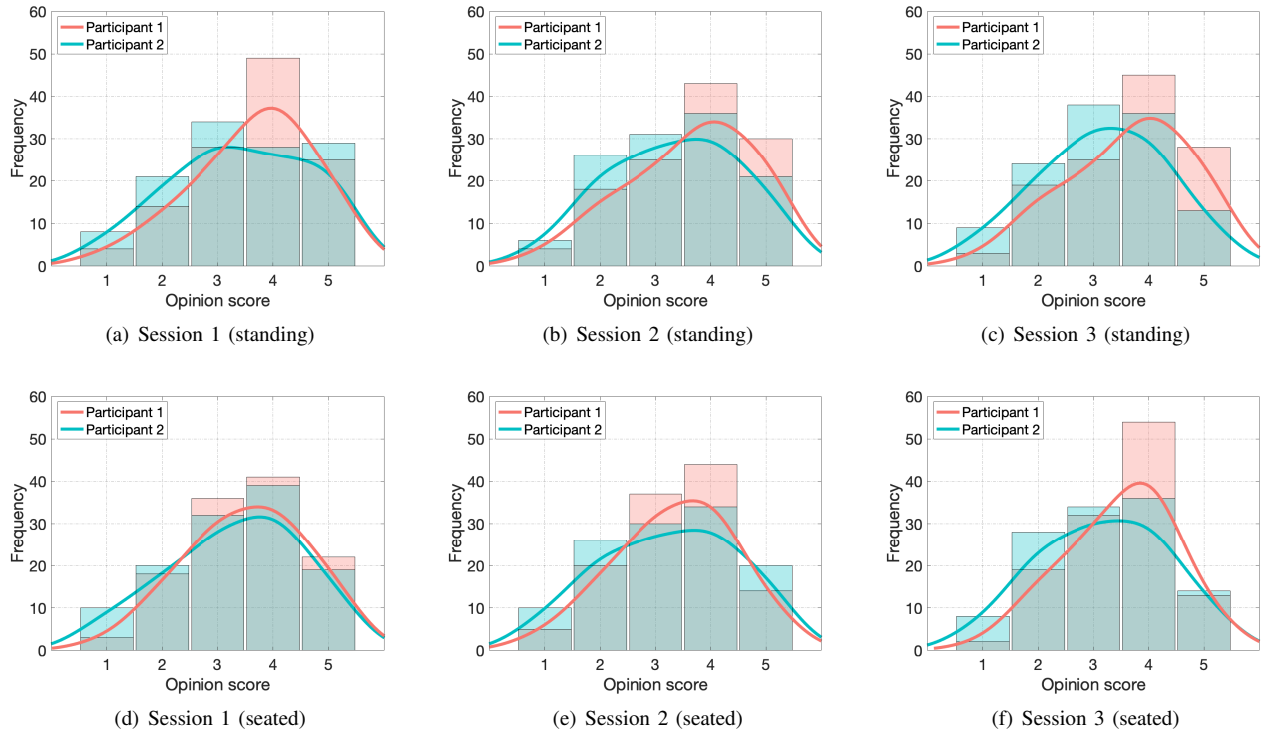


Fig. 2. Histograms of opinion scores given by the two participants to the test stimuli in standing and seated viewing, and their Kernel fit distributions.

Similar results are obtained for seated viewing in Figs. 2(d)-(f) along with the respective fits, i.e., each participant follows its own quality rating pattern of the test stimuli irrespective of the session.

Additional insights on the sets of opinion scores can be obtained from the skewness and kurtosis that are provided in Table I. First, let us consider the skewness assessing the asymmetry of the subjective data around the mean value of the distributions of opinion scores. Note that a negative or positive skewness, respectively, indicates that the tail is on the left or right side of a distribution. In the considered context, left-skewed distributions with the left tail being drawn out are obtained for both participants and all sessions. However, as the skewness magnitudes are small, the asymmetries are rather mild (see also fits in Figs.2(a)-(f)). Overall, more left-skewed opinion score distributions are obtained for P1 compared to P2. Second, the kurtosis measures how outlier-prone a distribution is with higher kurtosis indicating a greater extremity of outliers. The kurtosis of the normal distribution is 3, which is commonly used for comparison. Distributions that have a kurtosis greater than 3 are more outlier-prone than the normal distribution while having a kurtosis less than 3 applies to a less outlier-prone distribution. As can be seen from Table I, the kurtoses obtained for the sets of opinion scores of both participants are less than 3 for all sessions.

In addition, ANOVA tests [20]–[22] were conducted to determine if variations in the sets of opinion scores arise among the three sessions in standing and seated viewing. Table II and Table III show the summary statistics of the ANOVA tests

TABLE I
SKEWNESS AND KURTOSIS OF OPINION SCORES

| | P1 | | P2 | |
|-------|----------|----------|----------|----------|
| | Skewness | Kurtosis | Skewness | Kurtosis |
| S1 | -0.58 | 2.77 | -0.23 | 2.08 |
| ST S2 | -0.51 | 2.38 | -0.17 | 2.10 |
| S3 | -0.47 | 2.34 | -0.18 | 2.33 |
| S1 | -0.27 | 2.39 | -0.33 | 2.27 |
| SE S2 | -0.32 | 2.53 | -0.17 | 2.06 |
| S3 | -0.41 | 2.58 | -0.11 | 2.19 |

for P1 and P2 in terms of the sum of squares (SS), degree of freedom (DF), mean square (MS) value, F-value, and p-value. The sources of the tests are the respective three sets of opinion scores obtained for each participant in standing and seated viewing. Here, the p-values are of main interest giving the probability of the difference in the samples due to sampling errors. If the p-value is higher (lower) than a selected significance level α , the hypothesis of equal means is accepted (rejected). Regarding the p-values shown in the tables, the significance level of $\alpha = 0.05$ was chosen to assess if there exist statistically significant differences in the MOS among the sets of opinion scores obtained for the three sessions in standing and seated viewing. Clearly, as all p-values are significantly above the significance level of $\alpha = 0.05$, there exists strong evidence for statistically significant similarity

TABLE II
ANOVA TESTS AMONG SETS OF OPINION SCORES FOR P1

| Standing Viewing | | | | | |
|------------------|---------|-----|--------|--------|---------|
| Source | SS | DF | MS | F | p-value |
| Groups | 0.006 | 2 | 0.0028 | 0.0024 | 0.9976 |
| Error | 417.050 | 357 | 1.1682 | | |
| Total | 417.056 | 359 | | | |
| Seated Viewing | | | | | |
| Source | SS | DF | MS | F | p-value |
| Groups | 1.672 | 2 | 0.8361 | 0.8300 | 0.4365 |
| Error | 359.217 | 357 | 1.0062 | | |
| Total | 360.889 | 359 | | | |

TABLE III
ANOVA TESTS AMONG SETS OF OPINION SCORES FOR P2

| Standing Viewing | | | | | |
|------------------|---------|-----|--------|--------|---------|
| Source | SS | DF | MS | F | p-value |
| Groups | 3.672 | 2 | 1.8361 | 1.3700 | 0.2553 |
| Error | 478.325 | 357 | 1.3398 | | |
| Total | 481.997 | 359 | | | |
| Seated Viewing | | | | | |
| Source | SS | DF | MS | F | p-value |
| Groups | 1.206 | 2 | 0.6028 | 0.4400 | 0.6424 |
| Error | 485.725 | 357 | 1.3606 | | |
| Total | 486.931 | 359 | | | |

between the MOS of the considered groups. In other words, it may be conjectured that the respective sets of opinion scores obtained for each participant in the three sessions for each of the two viewing conditions are statistically significant similar.

B. Standard Deviation of Opinion Scores and SOS fitting Function

Figs. 3(a)-(f) show the SD of the MOS obtained for P1 and P2 in the three sessions for standing and seated viewing. The SOS fitting function given in (7) to the data points are also shown in these figures with the respective SOS parameter a and mean square error (MSE) provided in Table IV.

Similar to the results for conventional images and videos in [19], [23], [24], it is observed from the SOS fitting function that the SD becomes lower toward the lower and upper end of the five-level quality scale compared to the mid-quality range for both participants and all sessions. This progression means that it was more difficult for both participants to rate the mid-quality 360° videos. In the majority of the sessions in standing and seated viewing, the SD is lower for P1 compared to P2 indicating that P1 was more confident about the given opinion scores.

TABLE IV
SOS PARAMETER a AND MSE OF THE SOS FITTING FUNCTIONS

| | | S1 | | S2 | | S3 | |
|----|----|--------|--------|--------|--------|--------|--------|
| | | a | MSE | a | MSE | a | MSE |
| ST | P1 | 0.0804 | 0.6381 | 0.0843 | 0.6281 | 0.0638 | 0.5856 |
| | P2 | 0.0928 | 0.5692 | 0.1405 | 0.5736 | 0.0704 | 0.4668 |
| SE | P1 | 0.0507 | 0.6334 | 0.0907 | 0.5742 | 0.0535 | 0.5659 |
| | P2 | 0.0997 | 0.5301 | 0.0834 | 0.5072 | 0.0961 | 0.5478 |

C. Mean Opinion Scores

Fig. 4 and Fig. 5 show the MOS according to (1) over all four scenes versus QP for P1 and P2, respectively. In contrast to the histograms and SOS fitting functions, the progressions of MOS give fewer insights on the opinion score consistency across the three sessions. However, as expected, the MOS tends to decrease with increasing QP for each resolution. The spread of MOS for the 360° reference videos (Ref.) gives indication that P1 has less resolving power of the different resolutions (narrow cluster of MOS) compared to P2 (wider cluster of MOS). An indication for giving consistent opinion scores throughout the three sessions can be observed for the lowest resolution of 2K. In particular, the MOS progression for P1 for the 360° videos with 2K resolution over different QPs is very similar for all sessions in standing and seated viewing. Further, a plateau of almost constant MOS for QP = 32 and QP = 37 and subsequent drop of MOS for QP = 42 is observed for P1 for all sessions.

D. Statistical Analysis of Opinion Scores Over All Sessions

Additional insights on the quality rating behavior of each participant can be obtained through a statistical analysis of sets that contain their opinion scores from all three sessions for standing and seated viewing. In this way, the variation of results among the sessions is averaged revealing the typical quality rating behavior of each participant.

Figs. 6(a)-(c) show the histograms of opinion scores given by each participant accumulated over the respective three sessions for standing and seated viewing as well as histograms of opinion scores accumulated over both viewing conditions. The Kernel distributions that were fitted to the histograms are also provided in these figures. As can be seen from the histograms and Kernel distributions, each participant has its own quality rating behavior which varies little between standing and seated viewing. The skewness and kurtosis obtained for these sets of opinion scores confirm this finding (see Table V). In particular, the asymmetry of the histograms of opinion scores is slightly more skewed to the left for P1 while the kurtosis indicates a less outlier-prone distribution for P2. The quality rating behavior of each participant becomes even more apparent for the sets containing the opinion scores from both standing and seated viewing.

Figs. 7(a)-(c) show the SD defined in (5) versus average MOS over the three sessions for standing viewing, seated

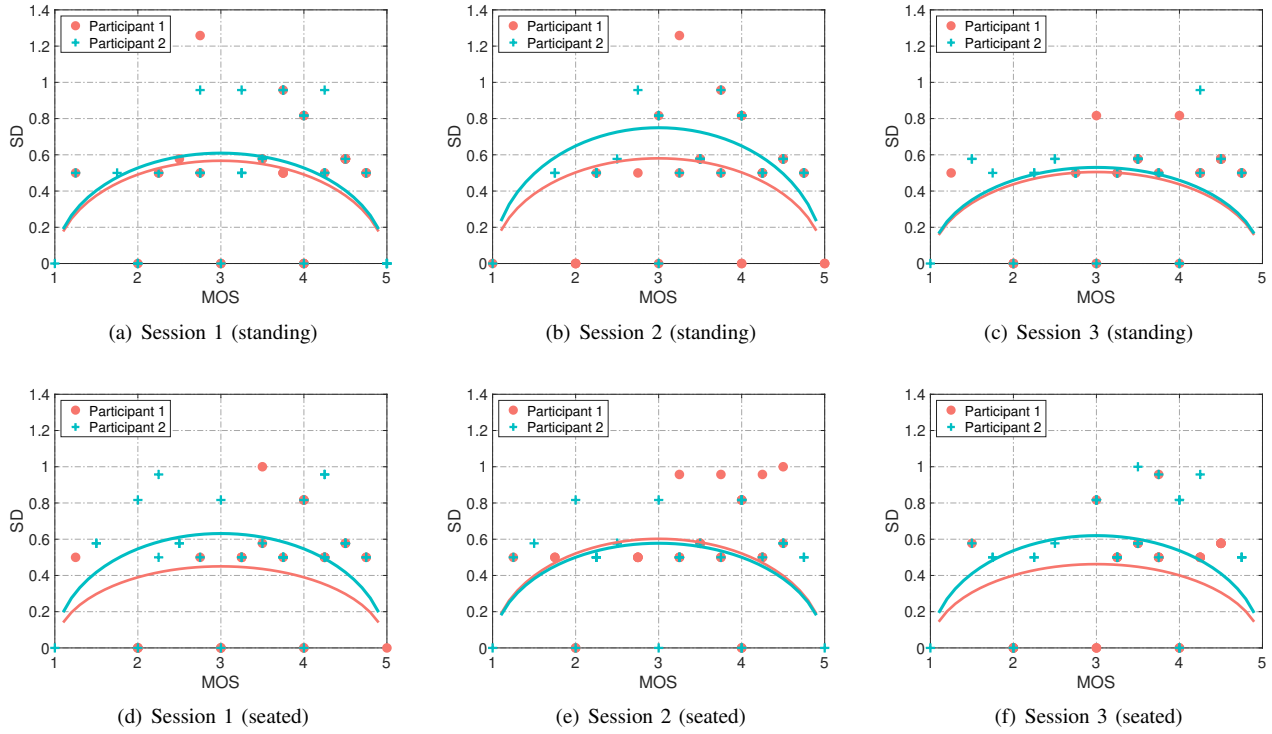


Fig. 3. Standard deviation versus MOS along with the SOS fitting function to the data points for standing and seated viewing.

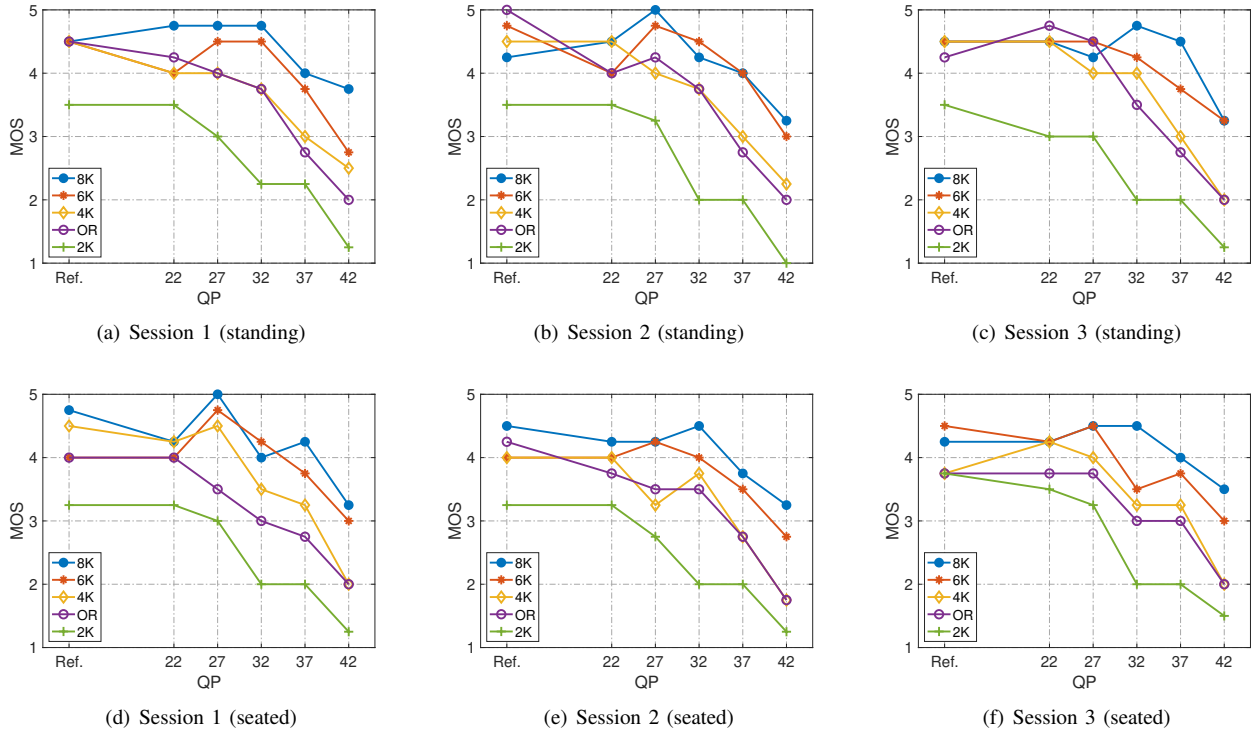


Fig. 4. MOS over the four scenes versus QP obtained for P1 in the three sessions for standing and seated viewing.

viewing, and both viewing conditions. The SOS fitting functions are also provided with SOS parameter a and MSE given in Table VI. The SOS fitting functions reveal that

the uncertainty in the given opinion scores is lower for P1 compared to P2 for the entire five-level quality scale. In addition, for each participant, the progression of the SOS

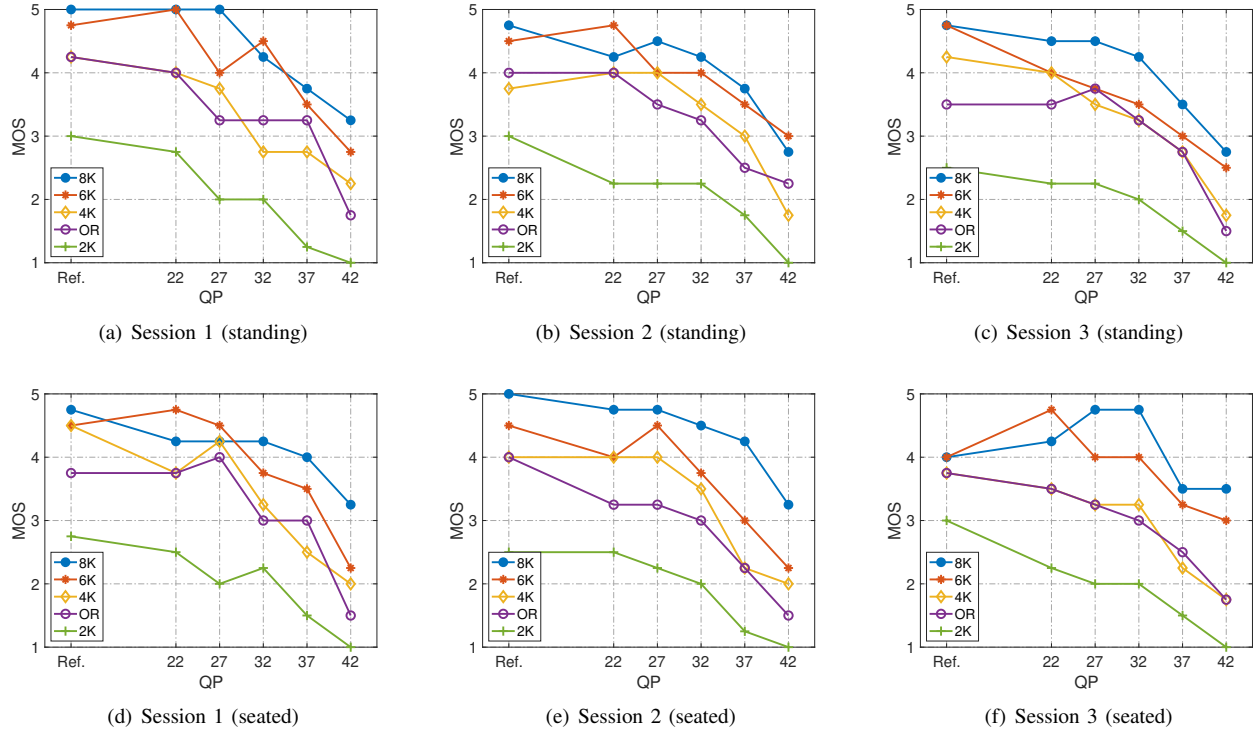


Fig. 5. MOS over the four scenes versus QP obtained for P2 in the three sessions for standing and seated viewing.

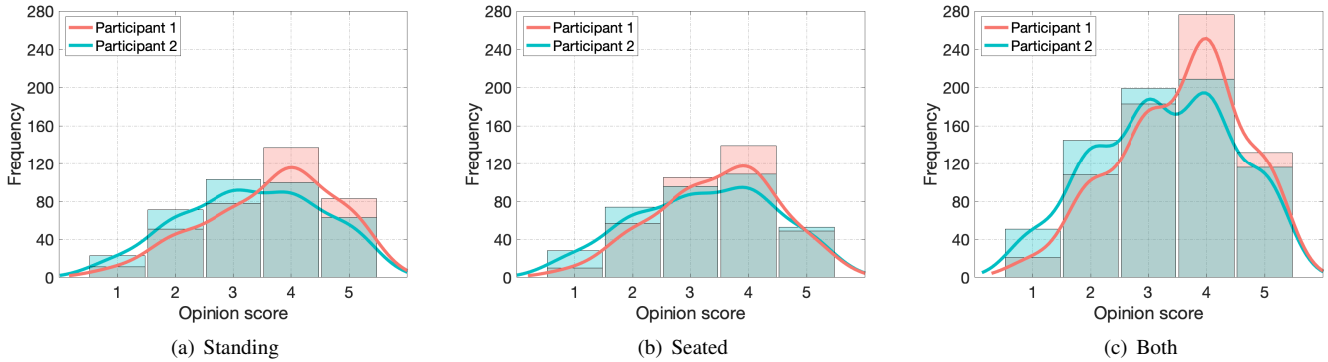


Fig. 6. Histograms of opinion scores accumulated over the three sessions for standing viewing, seated viewing, and both viewing conditions.

TABLE V
SKEWNESS AND KURTOSIS OF OPINION SCORES

| | P1 | | P2 | |
|------|----------|----------|----------|----------|
| | Skewness | Kurtosis | Skewness | Kurtosis |
| ST | -0.52 | 2.49 | -0.18 | 2.17 |
| SE | -0.33 | 2.51 | -0.20 | 2.16 |
| Both | -0.41 | 2.47 | -0.19 | 2.17 |

fitting functions differs very little among standing viewing, seated viewing, and both viewing conditions accumulated.

Figs. 8(a)-(f) show the average MOS defined in (2) over the respective three sessions for standing viewing, seated viewing,

and both viewing conditions versus QP. The averaging of opinion scores over all three sessions for standing and seated viewing shows that P1 has less resolution resolving power in standing viewing compared to seated viewing. This can be clearly seen for the reference videos with QP='Ref.' which receive similarly high average MOS for P1 in standing viewing for the resolutions OR, 4K, 6K, and 8K while the average MOS for these resolutions are not as clustered in seated viewing. On the other hand, the resolution resolving power of P2 is similar for standing and seated viewing. This observation indicates that a participant may get distracted from the quality assessment task in standing viewing compared to seated viewing.

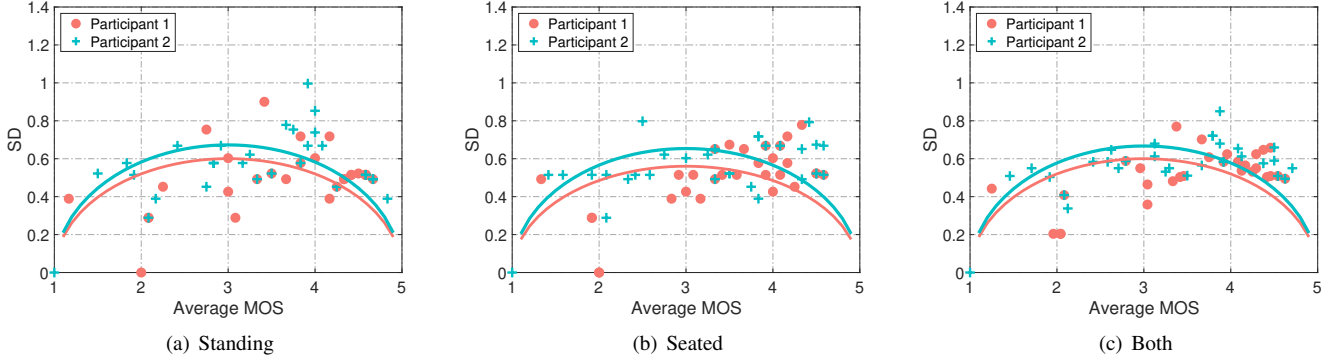


Fig. 7. Standard deviation versus average MOS over the three sessions for standing viewing, seated viewing, and both viewing conditions.

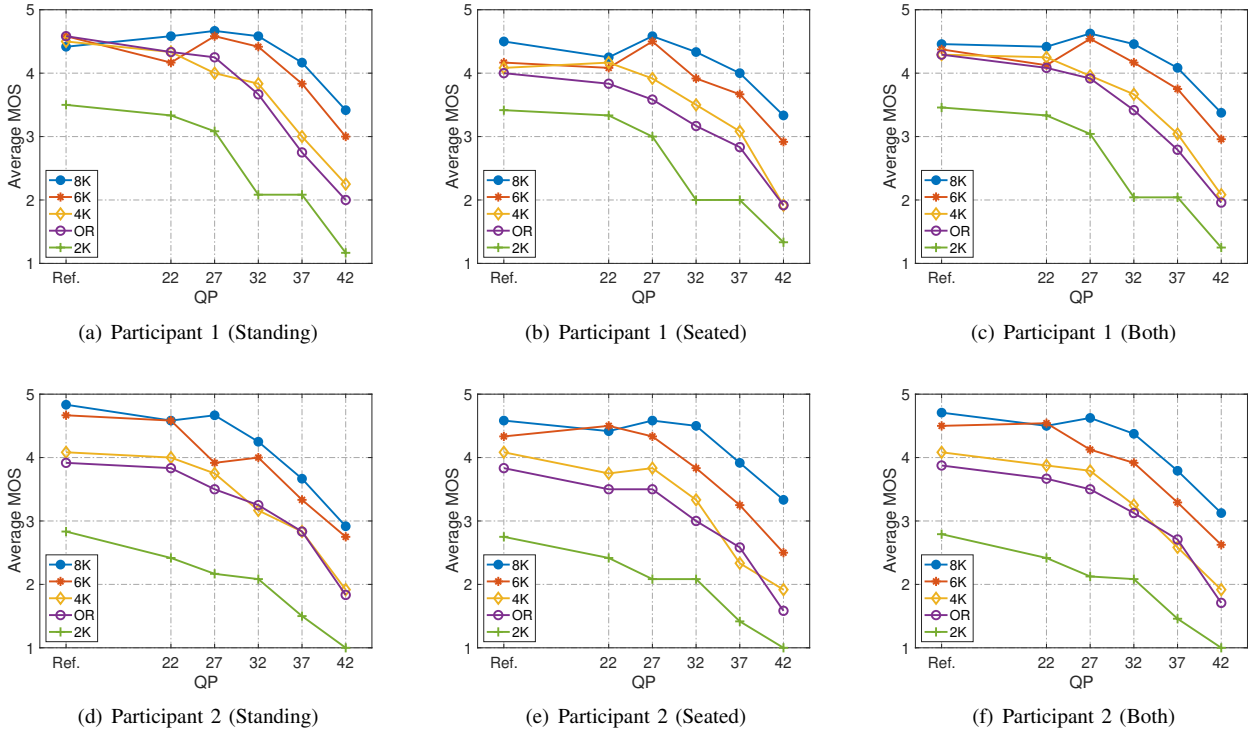


Fig. 8. Average MOS over the three sessions in standing viewing, seated viewing, and both viewing conditions: (a)-(c) Participant 1, (d)-(f) Participant 2.

TABLE VI
SOS PARAMETER a AND MSE OF THE SOS FITTING FUNCTIONS

| | ST | | SE | | All | |
|----|--------|--------|--------|--------|--------|--------|
| | a | MSE | a | MSE | a | MSE |
| P1 | 0.0902 | 0.3648 | 0.0785 | 0.3859 | 0.0901 | 0.3174 |
| P2 | 0.1132 | 0.3029 | 0.1069 | 0.2963 | 0.1113 | 0.2456 |

E. Statistical Analysis of Opinion Scores for All Sessions and All Participants

Finally, a statistical analysis is conducted for the sets containing the opinion scores given by both participants in the respective three sessions for standing and seated viewing.

As such, this additional accumulation of opinion scores aligns with the conventional statistical analysis of subjective tests aiming at obtaining MOS over the number of participants for each test case.

Figs. 9(a)-(b) show the histograms of opinion scores and SD of average MOS using (6) along with the SOS fitting functions. The histograms are similar for standing and seated viewing which is also supported by the skewness and kurtosis provided in Table VII. Given that the magnitude of the skewness of the set of opinion scores for seated viewing is slightly lower compared to standing viewing, the distribution of opinion scores conveyed by the Kernel fit is slightly less asymmetric for seated viewing. Similarly, the kurtosis for seated viewing is slightly closer to that of a normal distribution compared to standing viewing. The SOS fitting functions (see Table VIII)

to the SD versus average MOS progression indicate that the participants on the average are slightly more uncertain about their opinion scores given in standing viewing.

TABLE VII
SKEWNESS AND KURTOSIS OF OPINION SCORES

| | Skewness | Kurtosis |
|----|----------|----------|
| ST | -0.3500 | 2.2700 |
| SE | -0.2900 | 2.3500 |

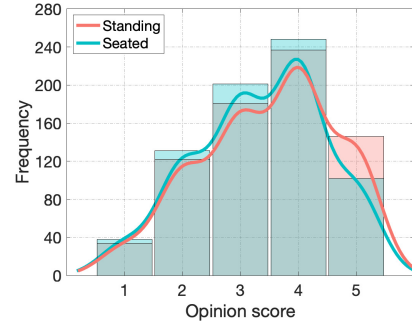
TABLE VIII
SOS PARAMETER a AND MSE OF THE SOS FITTING FUNCTIONS

| | a | MSE |
|----|--------|--------|
| ST | 0.1199 | 0.2587 |
| SE | 0.1048 | 0.2277 |

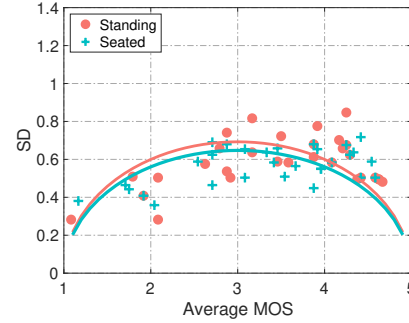
Figs. 10(a)-(c) present the average MOS versus QP using (3). Accordingly, for a given resolution-QP pair, the average of opinion scores is taken over the four 360° video scenes, over the three sessions for standing and seated viewing, and over the two participants. However, the consistency of opinion scores is not assessable anymore due to the three levels of averaging while differences among standing and seated viewing become more pronounced. In particular, the resolution resolving power is lower in standing viewing with the average MOS for the 360° reference videos of resolutions OR, 4K, 6K, and 8K being narrower clustered compared to seated viewing. Further, while the 360° videos with resolutions of OR and 4K received almost the same average MOS for the different resolution-QP pairs in standing viewing, they can be clearly differentiated on the average in seated viewing. This finding is consistent with the results of the pilot study reported in [12].

V. CONCLUSIONS

In this paper, we have studied the consistency of opinion scores given by the participants in repeated subjective tests on 360° video quality for standing and seated viewing on an HMD. In particular, three sessions were conducted for both standing and seated viewing with long breaks of several months and short breaks of hours or a day between sessions. A comprehensive statistical analysis of the data gathered in these subjective tests has been provided, including histograms of opinion scores, skewness and kurtosis of opinion scores, ANOVA tests, SD of opinion scores, SOS fitting functions, and MOS. The main conjecture supported by the statistical analysis is that each participant has its own but consistent quality rating behavior throughout the three sessions for standing and seated viewing. In other words, given that even long and short breaks between sessions were considered, the quality rating behavior of an individual participant does not fundamentally change over time. This important finding may assist in scheduling subjective tests under pandemic conditions where experimental campaigns may need to be stalled for an unknown period



(a)



(b)

Fig. 9. Statistical analysis of the sets containing the opinion scores from both participants given in the respective three sessions for standing and seated viewing: (a) Histogram, (b) Standard deviation.

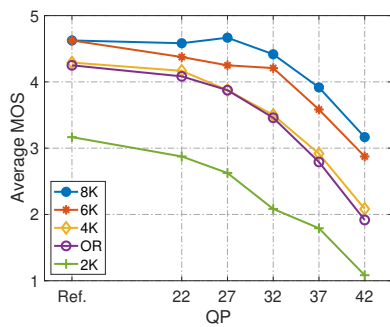
of time. As such, the opinion scores obtained from sessions that are scheduled several months apart could still be used to conduct a statistically meaningful analysis in terms of MOS. The statistical analysis also gives evidence that the resolving power among different 360° video resolutions may be lower for standing viewing compared to seated viewing. Given the important indicative results and conjectures of this study, future work may consider conducting large-scale subjective tests on quality assessment, opinion score consistency, and viewing behavior of 360° videos on HMDs. This future work may engage a larger panel of participants and a larger set of viewing conditions, e.g., fixed chair, half-swivel chair, full-swivel chair, couch, options of larger rotational and translational movements, and free walking.

ACKNOWLEDGMENT

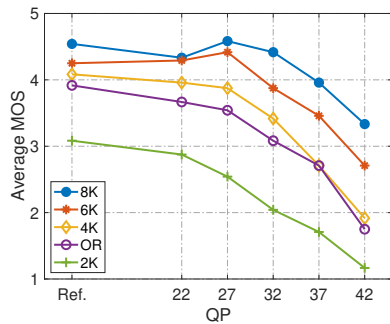
This work has been supported in part by the Knowledge Foundation, Sweden, through the ViaTech project under contract 20170056. We thank the volunteers for their time participating in this study.

REFERENCES

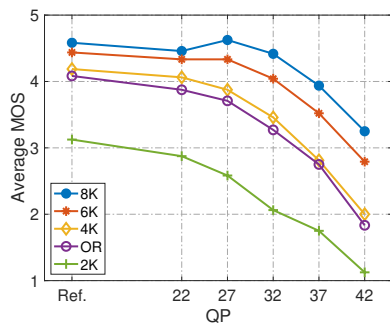
- [1] 3GPP TR 26.918 V16.0.0, *Virtual Reality (VR) Media Services Over 3GPP (Release 16)*, 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects, Dec. 2018.
- [2] Recommendation ITU-T BT.1788, *Methodology for the Subjective Assessment of Video Quality in Multimedia Applications*, International Telecommunication Union, Feb. 2007.



(a) Standing



(b) Seated



(c) Both

Fig. 10. Average MOS versus QP for the sets containing the opinion scores from both participants given in the respective three sessions for standing viewing, seated viewing, and both viewing conditions.

[3] Recommendation ITU-T P.910, *Subjective Video Quality Assessment Methods for Multimedia Applications*, International Telecommunication Union, Apr. 2008.

[4] Recommendation ITU-R BT.500-13, *Methodology for the Subjective Assessment of the Quality of Television Pictures*, International Telecommunication Union, Jan. 2012.

[5] Recommendation ITU-T P.919, *Subjective Test Methodologies for 360 Degree Video on HMD*, International Telecommunication Union, Oct. 2020.

[6] Recommendation ITU-T P.800.2, *Mean Opinion Score Interpretation and Reporting*, International Telecommunication Union, May 2013.

[7] J. Gugenheimer, D. Wolf, G. Haas, S. Krebs, and E. Rukzio, "SwiVR-Chair: A Motorized Swivel Chair to Nudge Users' Orientation for 360 Degree Storytelling in Virtual Reality," in *Proc. CHI Conf. on Human Factors in Comp. Systems*, New York, NY, USA, May 2016, pp. 1996–2000.

[8] Y. Hong, A. MacQuarrie, and A. Steed, "The Effect of Chair Type on Users' Viewing Experience for 360-degree Video," in *Proc. ACM*

Symposium on Virtual Reality Software and Technology, Tokyo, Japan, Dec. 2018, pp. 1–11.

[9] T. Nguyen-Vo, B. E. Riecke, W. Stuerzlinger, D. M. Pham, and E. Kruijff, "NaviBoard and NaviChair: Limited Translation Combined with Full Rotation for Efficient Virtual Locomotion," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 1, pp. 65–177, Jan. 2019.

[10] D. Zielasko and B. E. Riecke, "Can We Give Seated Users in Virtual Reality the Sensation of Standing or Even Walking? Do We Want To?" in *Proc. IEEE Conf. on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, Atlanta, GA, USA, Mar. 2020, pp. 281–282.

[11] D. Zielasko and B. Riecke, "Sitting vs. Standing in VR: Towards a Systematic Classification of Challenges and (Dis)Advantages," in *Proc. IEEE Conf. on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, Atlanta, GA, USA, Mar. 2020, pp. 297–298.

[12] Y. Hu, M. Elwardy, and H.-J. Zepernick, "On the Effect of Standing and Seated Viewing of 360° Videos on Subjective Quality Assessment: A Pilot Study," *Computers*, vol. 10, no. 6, pp. 1–28, Jun. 2021.

[13] M. Elwardy, H.-J. Zepernick, and Y. Hu, "On Head Movements in Repeated 360° Video Quality Assessment for Standing and Seated Viewing on Head Mounted Displays," in *Proc. IEEE Conf. on Virtual Reality and 3D User Interfaces Abstracts and Workshops*, Lisbon, Portugal, Mar./Apr. 2021, pp. 71–74.

[14] H.-J. Zepernick, K. Pieper, R. P. Spang, U. Engelke, M. Hirth, and B. Naderi, "On the Impact of COVID-19 on Subjective Digital Media Quality Assessment," in *Proc. IEEE Int. Workshop on Multimedia Signal Processing*, Tampere, Finland, Oct. 2021, accepted.

[15] D. Feil-Seifer, K. S. Haring, S. Rossi, A. R. Wagner, and T. Williams, "Where to Next? The Impact of COVID-19 on Human-Robot Interaction Research," *ACM Trans. Human-Robot Int.*, vol. 10, no. 1, pp. 1–7, Jun. 2020.

[16] M. Elwardy, H.-J. Zepernick, V. Sundstedt, and Y. Hu, "Impact of Participants' Experiences with Immersive Multimedia on 360° Video Quality Assessment," in *Proc. IEEE Int. Conf. on Signal Processing and Commun. Systems*, Gold Coast, Australia, Dec. 2019, pp. 40–49.

[17] Y. Zhang, Y. Wang, F. Liu, Z. Liu, Y. Li, D. Yang, and Z. Chen, "Subjective Panoramic Video Quality Assessment Database for Coding Applications," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 42–51, Jun. 2018.

[18] T. Hößfeld, R. Schatz, and S. Egger, "SOS: The MOS is Not Enough!" in *Proc. Int. Workshop on Quality of Multimedia Experience*, Mechelen, Belgium, Sep. 2011, pp. 131–136.

[19] T. Hossfeld, P. E. Heegaard, M. Varela, and S. Möller, "QoE Beyond the MOS: An In-depth Look at QoE via Better Metrics and Their Relation to MOS," *Quality and User Experience*, vol. 1, no. 2, pp. 1–23, Sep. 2016.

[20] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman, *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill, 1996.

[21] C. F. J. Wu and M. Hamada, *Experiments: Planning, Analysis, and Parameter Design Optimization*. New York: John Wiley & Sons, 2000.

[22] D. W. Cunningham and C. Wallraven, *Experimental Design: From User Studies to Psychophysics*. Boca Raton: CRC Press, 2012.

[23] F. De Simone, M. Naccari, M. Tagliasachi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective Assessment of H.264/AVC Video Sequences Transmitted Over a Noisy Channel," in *Proc. Int. Workshop on Quality of Multimedia Experience*, San Diego, CA, USA, Jul. 2009, pp. 204–209.

[24] U. Engelke, *Modelling Perceptual Quality and Visual Saliency for Image and Video Communications*. Doctoral Dissertation, Blekinge Institute of Technology, Karlskrona, Sweden, 2010.