EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems
Research Article EAI.EU

On the Consistency of 360° Video Quality Assessment in Repeated Subjective Tests: A Pilot Study

Majed Elwardy*, Hans-Jürgen Zepernick, Thi My Chinh Chu, and Yan Hu

Blekinge Institute of Technology, SE-37179 Karlskrona, Sweden
hans-jurgen.zepernick@bth.se (H.J.Z.); thi.my.chinh.chu@bth.se (T.M.C.C.); yan.hu@bth.se (Y.H.).

Abstract

Immersive media such as virtual reality, augmented reality, and 360° video have seen tremendous technological developments in recent years. Furthermore, the advances in head-mounted displays (HMDs) offer the users increased immersive experiences compared to conventional displays. To develop novel immersive media systems and services that satisfy the expectations of the users, it is essential to conduct subjective tests revealing users' perceived quality of immersive media. However, due to the new viewing dimensions provided by HMDs and the potential of interacting with the content, a wide range of subjective tests are required to understand the many aspects of user behavior in and quality perception of immersive media. The ground truth obtained by such subjective tests enable the development of optimized immersive media systems that fulfill the expectations of the users. This article focuses on the consistency of 360° video quality assessment to reveal whether users' subjective quality assessment of such immersive visual stimuli changes fundamentally over time or is kept consistent for each user. A pilot study was conducted under pandemic conditions with participants given the task of rating the quality of 360° video stimuli on an HMD in standing and seated viewing. The choice of conducting a pilot study is motivated by the fact that immersive media impose high cognitive load on the participants and the need to keep the number of participants under pandemic conditions as low as possible. To gain insight into the consistency of the participants' 360° video assessment over time, three sessions were held for each participant and each viewing condition with long and short breaks between sessions. In particular, the opinion scores and head movements were recorded for each participant and each session in standing and seated viewing. The statistical analysis of this data leads to the conjecture that the quality rating stays consistent throughout these sessions with each participant having their own quality assessment signature. The head movements, indicating the participants' scene exploration during the quality assessment task, also remain consistent for each participant according their individual narrower or wider scene exploration signature. These findings are more pronounced for standing viewing than for seated viewing. This work supports the role of pilot studies being a useful approach of conducting pre-tests on immersive media quality under opportunity-limited conditions and for the planning of subsequent full subjective tests with a large panel of participants. The annotated RQA360 dataset containing the data recorded in the repeated subjective tests is made publicly available to the research community.

Received on 07 November 2023; accepted on 29 December 2023; published on 08 January 2024

Keywords: 360° video, annotated dataset, opportunity-limited conditions, standing viewing, seated viewing, subjective tests, pilot study, quality of experience, quality assessment

doi:10.4108/eetinis.v11i1.4323

1. Introduction

In recent years, intelligent systems and advanced intelligent technologies together with extended realities

*Corresponding author: Majed Elwardy;
Tel.: +46-70-148-23-39.
Email: majed.elwardy@bth.se

(XRs) have seen tremendous advancements. In particular, XR serves as an umbrella term that captures virtual reality (VR), augmented reality (AR), and mixed reality (MR). It is anticipated that the metaverse will integrate several emerging technologies that support immersive cyber-virtual experiences in physical worlds such as digital twins, XR, artificial intelligence (AI), and 5G/6G mobile networks. In this context, digital twins provide a digital mirror of the physical world and may be integrated with advanced intelligent technologies to realize the metaverse [1, 2]. Interactive experiences in the different realizations of the real-virtual continuum are enabled by XR technologies allowing to be immersed in a virtual world. In view of networked immersive experiences, AI can enhance infrastructure reliability and performance while 5G/6G mobile networks offer seamless connectivity, high bandwidth, and significantly reduced latency.

Among the many immersive media, watching 360° videos on head-mounted displays (HMDs) has gained significant interest in recent years and has therefore been selected in this article as an application for studying the consistency of quality assessment. Viewing 360° videos on HMDs provides the users with a 360° × 180° viewing range and related 3+ degrees of freedom [3]. As such, the users may view 360° videos on HMDs while standing or seated exploring the full potential of unlimited rotational head movements (pitch, yaw, and roll) around the x, y, and z axes, and limited translational head movements along these axes. The fast developments of related technologies toward XR covering VR, AR, MR, and other immersive media modes are paving the way to novel applications including education, entertainment, healthcare, industry, marketing, and retail [4].

As humans are the final judges of the quality of experience (QoE) of immersive media applications, it is essential to base the design of stand-alone and networked immersive media systems and services on a suitable ground truth [5]. To obtain such ground truth, subjective tests are typically conducted in which participants assess the QoE of test stimuli that span over a wide range of quality levels. In relation to subjective assessment of 360° video quality, data collected during the tests may include quality ratings, rating times, head movements, eye tracking data, and galvanic skin responses. These psychophysical and psychophysiological data relate to explicit and implicit responses of the participants on the shown test stimuli which can be used to benchmark digital media processing chains with respect to QoE, and to develop objective perceptual quality models [6]. As the field of immersive media is developing fast [7], the question arises if the users' quality assessment to given immersive media applications fundamentally changes over certain periods of time or if it is kept rather similar.

Motivated by all of the above, this work focuses on the consistency of 360° video quality assessment through conducting subjective tests that were repeated after a long period of several months and a short period of a day or a few hours. The subjective test was conducted as a pilot study engaging experts on digital media processing and quality assessment. In this pilot study, the participants were presented a large number of 360° video stimuli for quality assessment in standing and seated viewing on an HTC Vive Pro HMD. It should be mentioned that the pilot study was chosen because of the more involved design of in-person experiments for immersive media which impose high cognitive load on the participants. In addition, the pilot study was conducted under the opportunity-limited conditions of the COVID-19 pandemic requiring to keep the number of participants as low as possible. Both of the above particulars of in-person experiments for immersive media, i.e., higher cognitive load imposed on participants and pandemic constraints, have increased the importance of pilot studies as a component of an overall subjective test framework. The statistical analysis of the opinion scores and head movements recorded in the repeated subjective tests not only allows conjectures on the consistency of 360° video quality assessment but also illustrates options for continuing related research under opportunity-limited conditions using pilot studies. Accordingly, the following objectives are pursued with this work:

O1: To conduct a pilot study that allows conjectures on the consistency of 360° video quality assessment.

O2: To perform a statistical analysis of the data recorded in the repeated subjective tests to evaluate the quality assessment behavior of the participants.

O3: To generate and publish an annotated dataset containing the psychophysical and psychophysiological data recorded in the pilot study allowing future research and use by the research community for meta-analysis.

In the rest of this section, related work is provided with respect to subjective quality assessment, subjective tests under opportunity-limited conditions, and repeated subjective test with few participants. Then, the contributions of this article are described.

1.1. Related Work

Subjective Quality Assessment. Experimental designs for subjective quality assessment of conventional videos have been well documented in literature and standardized by the International Telecommunication Union

(ITU). For example, in [8], a comprehensive introduction to psychophysical experiments and experimental design is given describing the design, execution, and analysis of perceptual studies. In [9], experimental designs and methodologies for the subjective assessment of television picture quality are provided. Similarly, in [10, 11], detailed recommendations on subjective quality assessment methods for multimedia applications are put forward. Regarding more immersive media, subjective assessment methods for 3D video quality are recommended in [12]. More recently, in [13], subjective test methodologies for 360° videos viewed on HMDs are described. In [14], focus is given to the QoE assessment of different types of XR telemeetings in VR, AR, or MR environments. The cross-lab subjective test campaign reported in [15] was carried out by the Immersive Media Group (IMG) of the Video Quality Experts Group (VQEG) and involved ten laboratories with a total of over 300 participants. This work evaluated the audiovisual quality, simulator sickness symptoms, and exploration behavior of short 360° videos on HMDs. An annotated dataset was generated containing the data recorded in these cross-lab subjective tests. It is noted that the results of these cross-lab subjective tests have fed into the development of Recommendation ITU-T P.919 [13]. As for the number of participants in subjective tests on media quality assessment, this is typically chosen to be above 20 participants. However, pilot studies with only a few participants are noted as an option to obtained general trends before conducting a time-consuming subjective test with a larger number of participants.

Subjective Tests Under Opportunity-Limited Conditions. Alternative experimental designs for conducting subjective tests under opportunity-limited conditions such as the COVID-19 pandemic, have been suggested. In-person subjective tests under pandemic conditions require stringent hygiene procedures and preferably a smaller number of participants to reduce the risk of infection. In [16], among other discussions, it is recommended developing better survey instruments and conducting meta-analysis that statistically combines results of multiple independent studies to derive overall conclusions about the posed research question. In [17], a status report is provided about the community discussions on in-person studies of immersive experiences under COVID-19 conditions and beyond. To deal with the additional issues on the availability and the risk of sharing specialized equipment such as HMDs, it is suggested to engage laboratory staff (experts) and infrastructure into in-person studies, and to recruit external participants (non-experts) possessing the required equipment. It is also suggested to consider participants' pooling through hardware distribution and to develop distributed experiments. In [18], a detailed experience report is given on how to conduct subjective tests under pandemic conditions. It is pointed out that the imposed hygiene measures such as the planning of experiments, securing sufficient ventilation, following stringent disinfection routines, and briefing the participants, consumed significant efforts prior to and after the experiments. An aim of this experimental design was to keep the time for the participants being required in the laboratory to a minimum. In [19], the impact of COVID-19 on conducting subjective tests for digital media quality assessment was discussed in the form of a position paper. Enablers suggested to facilitate QoE research under pandemic conditions include alternative experimental designs, adaptation of research methodologies, ethical vetting, standardization, and outsourcing work to a large-scale anonymous group of participants along with utilizing consumer-grade devices. To prevent ethical issues of QoE studies in advance, it is recommended to follow open science standards in terms of protocols, procedures, and tools, and making annotated datasets publicly available to support insightful meta-analyses. Further, it is suggested to replace test panels typically consisting of 15-28 participants that assess a relative small set of 25-30 test stimuli by a small panel that instead assesses a significantly larger set of test stimuli.

Repeated Subjective Tests with Few Participants. Given the need for subjective tests on immersive media quality assessment that engage only a few participants, related research on the number of participants has regained interest in recent years. Apart from keeping the number of participants low due to opportunity-limited situations, the recent developments on novel immersive media addressing a wide range of XR modes justify revisiting also the role of pilot studies with a few participants in an overall experimental design. In particular, in [20], the few observers with repetitions (FOWR) subjective test protocol has been proposed that engages four to six team member each rating the test stimuli several times. To reveal the suitability of the FOWR protocol for subjective quality assessment, a subjective test was conducted in which the participants were instructed to 10 times repeat the experiment of rating 110 processed video sequences. A total of 20 participants finished all 10 repetitions within 12 days and 8 months with a median time between consecutive repetitions of 2 weeks. The statistical analysis of the test results showed that FOWR-based experiments reach similar performance as conventional experiments in terms of association, agreement, perceptual similarity, and confusion analysis. The authors conjectured that this approach may be considered as a compromise between accurate but time-consuming subjective tests engaging a large number of users and less accurate but fast quality assessment using objective metrics. Further, pilot studies for pre-tests engaging a few participants with repeated subjective stimuli assessment are seen as a suitable methodology to reveal trends with reasonable accuracy.

1.2. Contributions

In continuation of our preliminary work and exploratory findings reported in [21, 22], this article provides a comprehensive study on participants' consistency in repeated subjective tests on 360° video quality assessment regarding opinion scores and head movements. In particular, in this article, the results of a comprehensive statistical analysis of different sets of opinion scores, average opinion scores, accumulated opinion scores, and head movements are presented using visualizations that offer more information such as violin plots and cumulative distribution functions (CDFs). Apart from measures of central tendency, average absolute deviations from these measures are used to examine the statistical deviation to avoid overweighing tail events. This statistical analysis is especially useful to make a comparison among the considered sets of data to reveal whether results are consistent or different.

The repeated subjective tests were conducted under the opportunity-limited conditions of the recent COVID-19 pandemic. The tests were executed as a pilot study engaging experts to reduce the risk of infection as suggested in [17, 19]. In contrast to large panels of participants viewing a relatively small set of test stimuli, this pilot study engaged two experts assessing a large number of test stimuli, i.e., 720 visual stimuli covering a wide range of quality levels.

The approach of performing repeated subjective tests is motivated by the following developments. First, immersive media applications such as watching 360° videos on HMDs are relatively new leveraging the recent advances in related software suites and hardware platforms. The perception of watching visual stimuli of immersive media on HMDs may therefore change once viewers' and even experts expectations have adapted to immersive media technologies. Second, subjective tests may need to be stalled in case of emerging opportunity-limited conditions such as a pandemic to prevent health risks and instead be continued at a later stage. However, a continuation of a subjective test would require that participants' quality perception on the immersive media under test does not fundamentally change over time but is kept consistent. As such, in relation to subjective quality assessment of 360° video stimuli, the main research questions pursued in this article are as follows:

RQ1: Does a participant's subjective quality assessment of 360° video stimuli viewed on an HMD change fundamentally over a certain period of time or does their quality assessment remain consistent? In this article, the characteristic quality assessment behavior associated with each participant is referred to as their "quality assessment signature".

RQ2: Do different participants have the same quality assessment signature or does each participant has a distinct quality assessment signature?

The pilot study and its results reported in this article not only shed light on these research questions but also may serve as an example for conducting in-person tests under opportunity-limited conditions where recruiting and engaging participants become a major challenge. The reported work is also sought to support the open science movement by making the wide range of data recorded during the pilot study publicly available to the research community allowing meta-analyses with other existing or future public annotated datasets. The main contributions of this article are summarized as follows:

C1: A repeated subjective test campaign is reported that was conducted under COVID-19 conditions which may serve as a guide for in-person immersive media quality assessment studies under opportunity-limited conditions.

C2: The RQA360 dataset established from this pilot study is made publicly available and a description of the dataset structure is provided. The RQA360 dataset contains opinion scores, head movements, eye tracking data, galvanic skin response (GSR) data, time stamps, rating durations, and demographic information about the participants.

C3: A statistical analysis of the recorded opinion scores and head movements is conducted with respect to four classifications of data for standing and seated viewing: (1) Original data for each session and participant; (2) Averaged or accumulated data over all video scenes for each session and participant; (3) Averaged or accumulated data over all video scenes and sessions for each participant; (4) Averaged or accumulated data over all video scenes, sessions, and participants.

C4: The statistical analysis of the opinion scores are presented as histograms, kernel fits to the histograms of opinion scores, and summary statistics, i.e., mean, median, mode, standard deviation, mean absolute deviation, median absolute deviation, skewness, and kurtosis.

C5: The statistical analysis of the different cases of averaged opinion scores are provided as violin plots with box plot inlets and the aforementioned summary statistics.

C6: CDFs for the head movements in terms of yaw, pitch, and roll angles are provided along with numerical values of these three rotational head movements for defined focus ranges.

C7: The obtained results lead to the conjecture that each participant possesses their individual but consistent quality assessment signature and head movement behavior throughout the three sessions for standing and seated viewing.

C8: The presented work supports the notion of pilot studies as a useful approach for conducting subjective tests on immersive media quality under opportunity-limited conditions and for the planning of subsequent full subjective tests with a large panel of participants.

It should be noted that the three objectives of this research are reached by the eight contributions as follows: Objective **O1** is achieved through contributions **C1, C7,** and **C8**; objective **O2** is fulfilled by contributions **C3** to **C6**; and objective **O3** is reached via contribution **C2**.

The remainder of this article is organized as follows. Section 2 summarizes the experimental design of the repeated subjective 360° video quality assessment tests. The structure of the RQA360 dateset is presented in Section 3. The measures used for the statistical analysis of the opinion scores and head movements gathered in the pilot study are presented in Section 4. The statistical analysis of the opinion scores and head movements are presented and discussed in Sections 5-6. Conclusions and future work are given in Section 7.

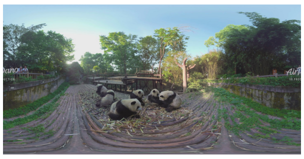
2. Experimental Design

The experimental design is based on our software suite and hardware platform that was developed for assessing subjective quality of immersive media when viewed on HMDs. Comprehensive details about the common components of the experimental setup used throughout our subjective test campaigns conducted over the years can be found in [6, 23]. In the following, the design of the repeated subjective quality assessment experiment for standing and seated viewing of 360° videos on an HMD is summarized to the extent needed for the understanding of the research reported in this article.

2.1. Visual Stimuli

Table 1 provides specifications of the 360° video stimuli that were used in this pilot study. The four natural scenes of 8K resolution were selected from the VQA-ODV dataset [24, 25] such that they span over a wide range of complexities and dynamics (see sample frames in Table 1). The bi-cubic scaling algorithm was used for downsampling the 8K reference videos which resulted in additional reference videos with lower resolutions. To significantly reduce the excessively high bitrates of these reference videos, the constant rate factor (CRF) option of the H.265 encoder with CRF=10 was used offering near perceptual lossless encoding. The libx265 encoder of the FFmpeg tool was then used to compress the perceptual lossless encoded reference videos with different quantization parameter (QP) to generate a set of test videos for each resolution. In total, a set of 120 video stimuli representing a wide range of quality levels was obtained from this processing, i.e., 4 scenes × 5 resolutions/scene × (1 reference + 5 QPs)/resolution.

Table 1. Summary of the 360° video scenes.

Sample frames of the chosen scenes [24, 25]	
Alcatraz	Blooming
Formation	Panda
	

Reference videos, test videos, tools	
Ref. videos	Resolution: 8K, 6K, 4K, optimal [26], 2K Frame rate: 29.97 fps Duration: 10 s Constant rate factor: CRF=10
Test videos	Resolution: 8K, 6K, 4K, optimal [26], 2K Frame rate: 29.97 fps Duration: 10 s Quantization: QP= 22, 27, 32, 37, 42
Tools	H.265 codec, libx265, FFmpeg [27, 28]

2.2. Software and Equipment

Table 2 provides specifications of the software suits, components of the human-machine interface (HMI), and hardware platform. The test platform for conducting subjective tests was developed using the Unity 3D game engine and Visual Studio 2017 which allow to create real-time interactive immersive applications for a wide range of devices. The iMotion Software Version 7.1 was used for recording the signals of the Shimmer biosensor.

Table 2. Software, human-machine interface, hardware.

Software suites	
Test platform	Unity 3D game engine V.2018.3.11f1 and Visual Studio 2017
Biosensor recordings	iMotion Software Version 7.1
Human-machine interface	
HMD	HTC Vive Pro with integrated eye-tracker Resolution: 1440×1600 pixels per eye Field of view: 110° Refresh rate: 90 Hz Gaze data output rate: 120 Hz
Interaction	HTC Vive controller
Sensors	Shimmer biosensor: (1) GSR (2) Heart rate
High-performance computing platform	
PC	Corsair One i160 Gaming PC with: (1) Intel I9-9900K processor of 3.6 GHz clock rate (2) NVIDIA GeForce RTX 2080 TI graphics card

The different components of the HMI comprise of the HTC Vive Pro HMD with integrated eye-tracker, HTC controller for interacting with the immersive virtual world, and the Shimmer biosensor for measuring the GSR and heart rate during HMD exposure.

Although the near perceptual lossless encoding of the original reference videos significantly reduced their excessively high bitrates, a high performance computing platform was needed to avoid stalling of the 360° video stimuli during streaming on the HMD. The requirement of smooth and uninterrupted streaming on the HMD is particular important as the conducted subjective tests aim at assessing quality of visual stimuli with spatial impairments rather then temporal impairments. The Corsair One i160 Gaming PC, offering the needed high-end processing performance, was hence selected for running the subjective tests.

2.3. Participants

As suggested in [19], hygiene measures under opportunity-limited conditions like a pandemic may be addressed by reducing the number of participants while the number of 360° video stimuli to be assessed by each participant is large instead. Two male participants, referred to as P1 (60 years of age) and P2 (31 years of age), took part in the three repeated subjective tests for standing and seated viewing. Informed consent was obtained from the participants before the subjective tests. Both participants were academic staff and familiar with multimedia signal processing and subjective test methodologies. It should be mentioned that the difference of 29 years between the ages of the two participants induce some level of demographic diversity. Each participant viewed the entire set of 120 different 360° video stimuli on the HMD in each of the three sessions for standing viewing and seated viewing on a fixed chair. As such, each participant viewed a total of 720 visual stimuli throughout the entire subjective test campaign.

2.4. Test Procedure

The quality rating of each 360° video of the sequence of stimuli shown on the HMD during a test session was performed using the five-level quality scale of the absolute category rating (ACR) method: (5) Excellent, (4) Good, (3) Fair, (2) Poor, (1) Bad [11, 13]. The 360° video stimuli shown on the HMD were presented in random order, one at a time, and rated independently on the five-level quality scale. Each session in standing viewing within a marked playing area and seated viewing on a fixed chair lasted around 30 minutes depending on the required time needed by each participant to rate the quality of the 120 different 360° video stimuli shown during a session.

Figure 1 shows the session schedule of the repeated 360° video quality assessment test accounting for a long break of several months between Session 1 (S1) and Session 2 (S2), and a short break of a day or some hours between S2 and Session 3 (S3). In this way, the participants' consistency of 360° video quality assessment over different periods of time can be captured.

The intervals between sessions in experimental designs for quality assessment tests of visual stimuli are conventionally kept short with slight variations in the timing among sessions due to the availability of participants and practical test schedule constraints. To study whether participants keep their quality assessment behavior consistent over longer periods of time, breaks of several months between consecutive sessions need to be considered. In this research, breaks of several months have therefore been induced between S1 and S2 while the conventionally used shorter breaks are placed between S2 and S3. Because the considered subjective tests reached into the COVID-19 pandemic, the breaks of six months and above relate to the time that passed to obtain organizational approval for the procedures of conducting in-person experiments under

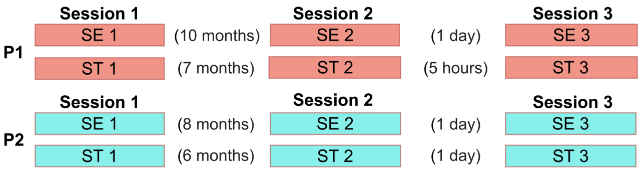


Figure 1. Session schedule (ST: standing, SE: seated) ©[2021] IEEE. Reprinted, with permission, from [21].

these opportunity-limited conditions and the related more time-consuming setup of the experiments.

3. RQA360 Dataset Structure

The data recorded in this pilot study is publicly available as RQA360 dataset under the GitHub link in [29]. The structure of the RQA360 dataset is shown in Figure 2 comprising of three main folders:

- *Standing_viewing_ACR_RQA360*
- *Seated_viewing_ACR_RQA360*
- *Test_Scenes*

The first and second main folder are associated with standing and seated viewing, respectively. Each of these main folders contains five sub-folders and one .csv file providing information about the participants as detailed in Table 3. The folder *Test_Scenes* contains specifications of the perceptual lossless encoded 360° reference videos, the generated 360° test videos with different resolutions and QPs, and provides instructions about the downloading of the 120 videos. The data files in these folders are in .csv format along with detailed information on the file structure given in *Readme* files.

Regarding the research reported in this article, the opinion scores and head movement data were analysed. The other data recorded during the pilot study may be used to address other research questions.

4. Statistical Measures

4.1. Averages of Opinion Scores

Let $u_{ijk}^{(n)}$ denote the opinion score, OS, that was given by participant $n \in \mathcal{N} = \{1, \ldots, N\}$ in session $i \in \mathcal{I} = \{1, \ldots, I\}$ of a certain viewing condition to test case $j \in \mathcal{J} = \{1, \ldots, J\}$ of 360° video scene $k \in \mathcal{K} = \{1, \ldots, K\}$. In this pilot study, $N = 2$ participants took part in $I = 3$ sessions for each viewing condition in which they assessed $J = 30$ test cases for $K = 4$ different 360° video scenes.

To support content independent quality assessment, let us define $\overline{u}_{ij}^{(n)}$ as average opinion score, $\overline{\text{OS}}_1$, over the K different 360° video scenes associated with participant n, session i of a certain viewing condition, and test case j as

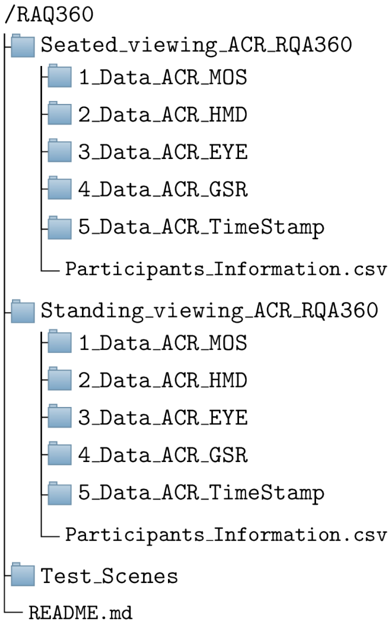

```
/RAQ360
├─ Seated_viewing_ACR_RQA360
│   ├─ 1_Data_ACR_MOS
│   ├─ 2_Data_ACR_HMD
│   ├─ 3_Data_ACR_EYE
│   ├─ 4_Data_ACR_GSR
│   ├─ 5_Data_ACR_TimeStamp
│   └─ Participants_Information.csv
├─ Standing_viewing_ACR_RQA360
│   ├─ 1_Data_ACR_MOS
│   ├─ 2_Data_ACR_HMD
│   ├─ 3_Data_ACR_EYE
│   ├─ 4_Data_ACR_GSR
│   ├─ 5_Data_ACR_TimeStamp
│   └─ Participants_Information.csv
├─ Test_Scenes
└─ README.md
```

Figure 2. RQA360 dataset directory structure.

$$\overline{u}_{ij}^{(n)} = \frac{1}{K} \sum_{k=1}^{K} u_{ijk}^{(n)} \tag{1}$$

The mean of the set comprising of the average opinion scores of the J different test cases associated with participant n and session i of a certain viewing condition can therefore be defined as

$$\mu_i^{(n)} = \frac{1}{J} \sum_{j=1}^{J} \overline{u}_{ij}^{(n)} \tag{2}$$

Similarly, to support content and session independent quality assessment, let us define $\overline{u}_j^{(n)}$ as average opinion score, $\overline{\text{OS}}_2$, over both K different 360° video scenes and I sessions of a certain viewing condition associated with participant n and test case j as

$$\overline{u}_j^{(n)} = \frac{1}{I} \sum_{i=1}^{I} \overline{u}_{ij}^{(n)} \tag{3}$$

where $\overline{u}_{ij}^{(n)}$ is defined in (1). The mean of the set of the average opinion scores of the J different test cases irrespective of content k and session i of a certain viewing condition associated with participant n can be defined as

$$\mu^{(n)} = \frac{1}{J} \sum_{j=1}^{J} \overline{u}_j^{(n)} \tag{4}$$

Table 3. Content of the RQA360 dataset.

Name	Content of folders, sub-folders, and files
Data_ACR_MOS	Opinion scores of the 360° videos per scene and corresponding rating durations. Four .csv files are provided for each participant, i.e., one dedicated .csv file for each of the four 360° video scenes. Resolution, QP, duration, frame rate (fps), and bitrate (Mbps) are also provided for the 360° video stimuli of each scene.
Data_ACR_HMD	Head movements during HMD exposure are provided as head positions in Cartesian coordinates x, y, and z, and head rotations as yaw, pitch, and roll angles.
Data_ACR_EYE	Eye tracking data is provided in three files as gaze left, gaze right, and gaze combined. Gaze left and gaze right data is given as pupil dilation and pupil dilation validity (true/false). Gaze combined data considers both eyes specifying origin and direction of the eye tracking and its validity.
Data_ACR_GSR	GSR files containing the GSR signal in micro-Siemens (μS), GSR quality (valid/not-valid), and heart rate (beats/min.).
Data_ACR_TimeStamp	Timestamps that mark start and end of each event during HMD exposure: Experiment, video, and rating.
Participants_Information.csv	Age, gender, occupation, and self-reported level of experience with immersive media.

Finally, to support content, session, and participant independent quality assessment, let us define $\overline{u}_j$ as mean opinion score, MOS, over K different 360° video scenes, I sessions of a certain viewing condition, and N participants as

$$\overline{u}_j = \frac{1}{N}\sum_{n=1}^{N}\overline{u}_j^{(n)} \tag{5}$$

where $\overline{u}_j^{(n)}$ is defined in (3). The mean of the set of the average opinion scores of the J different test cases irrespective of content k, session i, and participant n can be defined as

$$\mu = \frac{1}{J}\sum_{j=1}^{J}\overline{u}_j \tag{6}$$

4.2. Standard Deviation of Opinion Scores

The standard deviation (SD) measures the dispersion of data in relation to the average of a given dataset. Given the different levels of average opinion scores and their means, the corresponding SDs can be formulated with (1) to (6) as

$$\sigma_i^{(n)} = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}\left[\overline{u}_{ij}^{(n)} - \mu_i^{(n)}\right]^2} \tag{7}$$

$$\sigma^{(n)} = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}\left[\overline{u}_j^{(n)} - \mu^{(n)}\right]^2} \tag{8}$$

$$\sigma = \sqrt{\frac{1}{J-1}\sum_{j=1}^{J}\left[\overline{u}_j - \mu\right]^2} \tag{9}$$

4.3. Average Absolute Deviation

The average absolute deviation (AAD) of a dataset quantifies the absolute statistical dispersion from the dataset's measure of central tendency [30], e.g., mean or median. While the SD gives large weight to large observations due to the squaring of the deviation of a sample value from the selected measure of central tendency, the AAD avoids overweighing such tail events by processing the absolute value of the deviations [31]. In other words, the AAD is considered as being more resilient to outliers compared to the SD.

In the considered context, given a dataset $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ of opinion scores u_n, the mean absolute deviation, MAD$_1$, and the median absolute deviation,

MAD_2, respectively, can be defined as

$$\mathrm{MAD}_1 = \mathrm{mean}(|u_n - \mathrm{mean}(\mathcal{U})|) \quad (10)$$
$$\mathrm{MAD}_2 = \mathrm{median}(|u_n - \mathrm{median}(\mathcal{U})|) \quad (11)$$

where $|\cdot|$, mean$(\cdot)$, and median$(\cdot)$ denote absolute value, mean operator, and median operator, respectively.

4.4. Higher-Order Statistics

Additional insights into the participants' consistency of 360° video quality assessment can be obtained by higher-order statistics for estimating shape parameters such as skewness and kurtosis. While lower-order statistics such as mean and SD use constant, linear, and quadratic terms of the samples, higher-order statistics use third and higher moments. In this article, we use the unbiased versions of skewness and kurtosis of the different sets of opinion scores with different levels of averaging to evaluate the asymmetry and tailedness of the various histograms of average opinion scores. This allows us to not only study the participants' quality assessment consistency throughout the three sessions conducted for each viewing condition in terms of measures of central tendencies and dispersion of the sets of average opinion scores but also the shape of their histograms.

The skewness of a dataset $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ of opinion scores u_n can be defined as [32]

$$\mathrm{Skewness} = \frac{\sqrt{n(n-1)}}{n-2} s_1 \quad (12)$$

where

$$s_1 = \frac{\frac{1}{N}\sum_{n=1}^{N}(u_n-\mu)^3}{\left[\sqrt{\frac{1}{N}\sum_{n=1}^{N}(u_n-\mu)^2}\right]^3} \quad (13)$$

and μ denotes the mean of the dataset $\mathcal{U}$. For unimodal histograms, a negative (positive) skewness indicates that the left (right) tail of the histogram is longer compared to the right (left) tail.

The kurtosis of a dataset $\mathcal{U}$ of opinion scores u_n can be defined as [32]

$$\mathrm{Kurtosis} = \frac{N-1}{(N-2)(N-3)}\left[(N+1)k_1 - 3(N-1) + 3\right] \quad (14)$$

where

$$k_1 = \frac{\frac{1}{N}\sum_{n=1}^{N}(u_n-\mu)^4}{\left[\frac{1}{N}\sum_{n=1}^{N}(u_n-\mu)^2\right]^2} \quad (15)$$

In the considered context, the kurtosis measures the tailedness [33] of the various histograms or sets of opinion scores and sets of average opinion scores.

5. Statistical Analysis of Sets of Opinion Scores and Average Opinion Scores

This section presents the statistical analysis of the opinion scores recorded for each participant during the three sessions for standing and seated viewing and associated average opinion scores. In particular, the sets of opinion scores and average opinion scores that have undergone statistical analysis are composed with respect to the following data categories:

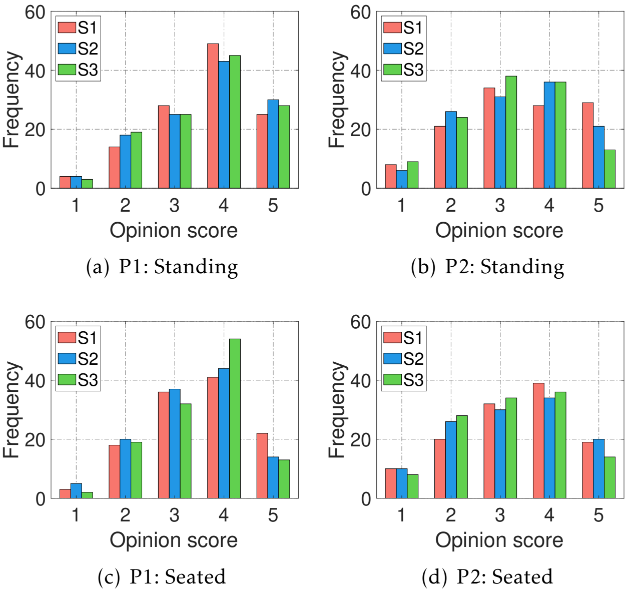
- **OS**: Opinion scores for each session, each participant, and each viewing condition.
- $\overline{\mathbf{OS}}_1$: Average opinion scores over the 360° video scenes for each session, participant, and viewing condition. The results represent the quality rating irrespective of the 360° video scene.
- $\overline{\mathbf{OS}}_2$: Average opinion scores over the 360° video scenes and sessions for each participant and viewing condition. The results represent the quality rating irrespective of the 360° video scene and session.
- **MOS**: Average opinion scores over the 360° video scenes, sessions, and participants for each viewing condition. The results represent the quality rating irrespective of the 360° video scene, session, and participant.

To support the findings conjectured from the statistical analysis of these sets, Appendix B provides histograms, kernel fits, and summary statistics for accumulated opinion scores that are grouped according to the above categories but are not averaged. In this way, it is verified that the averaging of opinion scores considered in this section does not remove important information on the quality assessment consistency.

5.1. Statistical Analysis of Opinion Scores

Figure 3 shows the histograms of opinion scores given by Participant 1 (P1) and Participant 2 (P2) to the $K \times J = 120$ test stimuli shown in each of the three sessions, S1, S2, and S3, for standing viewing (ST) and seated viewing (SE).

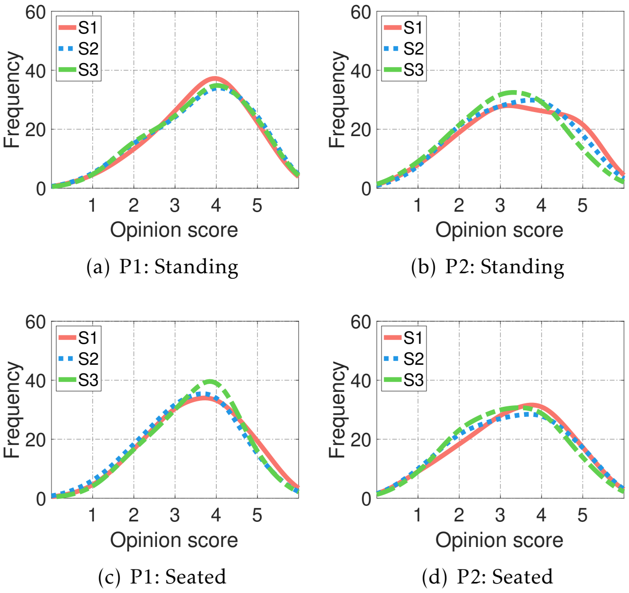
Figures 3(a)-(b) depict the histograms of OSs for P1 and P2, respectively, for each session in ST. Clearly, the frequencies of opinion scores for a given quality score are kept at similar levels for most of the cases throughout the three sessions for each participant. However, the shapes of the histograms for P1 and P2 differ but are consistent for a given participant throughout their sessions. While P1 gives more ratings toward good quality (OS = 4) for all sessions, the ratings given by P2 are concentrated in the mid-range of the quality scale for all sessions making the histograms more symmetric.



(a) P1: Standing (b) P2: Standing

(c) P1: Seated (d) P2: Seated

Figure 3. Histograms of opinion scores for P1 and P2 in each session and viewing condition.



(a) P1: Standing (b) P2: Standing

(c) P1: Seated (d) P2: Seated

Figure 4. Kernel fits to the histograms obtained for P1 and P2 in each session and viewing condition.

Figures 3(c)-(d) show the histograms of OSs for P1 and P2, respectively, for each session in SE. The results indicate that both participants become more critical about the quality of the visual stimuli when they are seated. In particular, the frequency of giving the quality score of OS = 5 is generally lower for SE compared to ST. This behavior is thought to be due to the participants' being more focused on the quality assessment task for SE with less distraction caused by other tasks such as keeping the body in balance during HMD exposure in ST. Despite this difference between the quality assessment in ST and SE, both participants possess their own quality assessment signature in terms of distinct histograms of OSs throughout the sessions. While the quality scores given by P1 are negatively skewed around good quality, quality ratings are more symmetrically distributed for P2.

To estimate the OS frequencies for the different scenarios, kernel fits to the data of the different sets of OSs were performed using non-parametric kernel-smoothing. The inherent smoothing processing allows another means of comparing the consistency of the quality assessment among sessions and participants. Figures 4(a)-(d) show the kernel fits to the histograms obtained for P1 and P2 in each session and viewing condition. The results clearly show that both participants have their own but consistent quality assessment signature for all three sessions of a given viewing condition. The kernel fits obtained for P1 show pronounced tails to the left while more symmetric progressions are observed for P2.

Tables 4-5 show the summary statistics of the OSs given by P1 and P2 to the 120 test stimuli shown in each session and viewing condition. Mean, median, and mode of the different sets of OSs assess the central tendencies of the data shown in the histograms. SD and MAD_1 represent the dispersion of the data with respect to the mean while MAD_2 quantifies the dispersion in relation to the median. The shape of the histograms in terms of asymmetry and tailedness are measured by the skewness and kurtosis, respectively.

Table 4 shows the summary statistics of the sets of OSs obtained for P1. The minor variations of the means and medians of the OSs for a given viewing condition indicate consistent quality assessment throughout the three sessions irrespective of the duration of the breaks between the sessions (see Figure 1). The mode of 4 is located toward the upper end of the quality scale which indicates the asymmetry of the related histograms with a tail toward the lower end of the quality scale. The values obtained for SD, MAD_1, and MAD_2 indicate that the dispersion of the OSs around the mean and median are similar for all sessions and given viewing condition. This finding suggests that not only the central tendencies contained in the sets of OSs but also the dispersion associated with the quality ratings given by P1 to the test stimuli during the sessions remain consistent irrespective of the duration of the breaks between the sessions. The noted asymmetry is confirmed by the negative skewness obtained for the histograms in Figures 3(a),(c) and the kernel fits in Figures 4(a),(c). In other words, the asymmetry of the unimodal histograms and kernel fits consistently

Table 4. Summary statistics of the sets of OSs obtained for P1 in each session and viewing condition.

	Mean	Med.	Mod.	SD	MAD_1	MAD_2	Skew.	Kurt.
S1	3.64	4.00	4.00	1.04	0.86	1.00	−0.59	2.81
ST S2	3.64	4.00	4.00	1.11	0.94	1.00	−0.51	2.41
S3	3.63	4.00	4.00	1.08	0.91	1.00	−0.47	2.37
S1	3.51	4.00	4.00	1.04	0.88	1.00	−0.28	2.41
SE S2	3.35	3.00	4.00	1.03	0.86	1.00	−0.32	2.56
S3	3.48	4.00	4.00	0.94	0.80	1.00	−0.42	2.61

Table 5. Summary statistics of the sets of OSs obtained for P2 in each session and viewing condition.

	Mean	Med.	Mod.	SD	MAD_1	MAD_2	Skew.	Kurt.
S1	3.41	3.00	3.00	1.22	1.05	1.00	−0.24	2.09
ST S2	3.33	3.00	4.00	1.15	0.98	1.00	−0.18	2.11
S3	3.17	3.00	3.00	1.10	0.90	1.00	−0.18	2.36
S1	3.31	3.00	4.00	1.17	0.99	1.00	−0.34	2.30
SE S2	3.23	3.00	4.00	1.21	1.02	1.00	−0.17	2.07
S3	3.17	3.00	4.00	1.12	0.93	1.00	−0.12	2.21

suggest longer tails to the left with the mass of the histograms and kernel fits being concentrated to the right for both viewing conditions. However, the asymmetry becomes less pronounced for SE compared to ST as the lower sknewness values indicate. The kurtosis in the range between 2.37 to 2.81 indicates a broadening of the peaks and thickening of the tails.

Table 5 shows the summary statistics of the sets of OSs obtained for P2. As with P1, the results obtained for P2 support the conjecture of a consistent quality assessment throughout each session of a given viewing condition. In particular, the means and medians tend more toward the mid-quality score of OS = 3 for both viewing conditions. On the other hand, the mode tends to a value of 3 for ST while the mode of 4 is obtained for SE. Similarly, the SD, MAD_1, and MAD_2 reveal consistent dispersion of the OSs around the mean and median throughout each session of a given viewing condition. However, the SD and MAD_1 values obtained for P2 are larger than those for P1 which indicates a slightly higher dispersion around the central tendencies of the quality ratings of P2. Given the more symmetric histograms and kernel functions of the different sets of OSs for P2, the smaller negative skewness values suggest a less developed tail to the lower quality level compared to P1. Similarly, the kurtosis values for P2 in the range between 2.07 and 2.36 indicate a lower tailedness compared to P1.

In addition, analysis of variance (ANOVA) tests [34, 35] were performed among the different sets of opinion scores to determine if statistical significant variations exist that would point against the above conjectured consistent quality assessment of P1 and P2 in their sessions in standing and seated viewing. The p-values produced by the ANOVA tests represent the probability of the differences in the samples due to sampling errors. Here, we have used a significance level of $\alpha = 0.05$. The ANOVA tests among the three sets of opinion scores for P1 in standing viewing (P1–ST), P1 in seated viewing (P1–SE), P2 in standing viewing (P2–ST), and P2 in seated viewing (P2–SE), returned the following p-values:

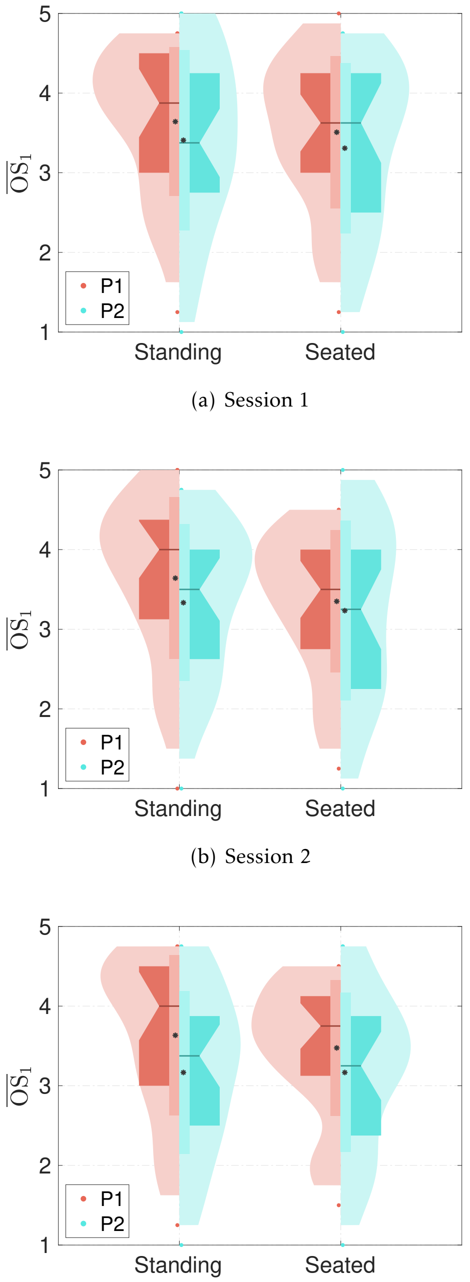
- P1–ST: p = 0.9976
- P1–SE: p = 0.4365
- P2–ST: p = 0.2553
- P2–SE: p = 0.6424

Because the p-values are well above the significance level α, it can be conjectured that the variations contained in the sets of OSs are statistically insignificant for these cases. In other words, the ANOVA tests confirm that the quality assessment of P1 and P2 is consistent throughout the three sessions conducted for both viewing conditions.

5.2. Statistical Analysis of Average Opinion Scores Over Video Scenes

In this section, a first averaging of the OSs is performed with respect to the four different 360° video scenes leading to average opinion scores $\overline{\text{OS}}_1$ as defined in (1). The $\overline{\text{OS}}_1$ values are real numbers which express the center of a set of four OSs that were given by a participant to the four different 360° video scenes of the same resolution-QP pair. Accordingly, $\overline{\text{OS}}_1$ values capture the participants' quality assessment irrespective of the content of the video scenes. The results obtained by the statistical analysis of the different sets of 30 $\overline{\text{OS}}_1$ values are visualized by violin plots which comprise of a box plot and a kernel density plot of the respective data.

Figures 5(a)-(c) show the violin plots of the sets of $\overline{\text{OS}}_1$ for each participant and session in ST and SE. The inner shape of each violin plot represents the box plots conveying summary statistics about $\overline{\text{OS}}_1$ and the outer shape shows the kernel densities of the $\overline{\text{OS}}_1$ values. It is observed that higher means (black asterisk marker) of the sets of $\overline{\text{OS}}_1$ values are obtained for P1 for each session in ST and SE compared to those obtained for P2. The medians (central mark on the notch of the box plots) of the sets of $\overline{\text{OS}}_1$ values are the same for both participants in seated viewing of S1. However, as with the means, the medians are higher for P1 in standing viewing of S1 and both viewing conditions of S2 and S3. Generally, the means and medians are higher for ST compared to SE for both participants indicating that P1 and P2 become



(a) Session 1

(b) Session 2

(c) Session 3

Figure 5. Violin plots of $\overline{\text{OS}}_1$ for each participant, session, and viewing condition.

more reluctant to give higher ratings when they are seated allowing more focus on the quality assessment task. Furthermore, the kernel densities indicate that both participants keep their distinct quality assessment signatures consistent throughout their sessions also for these content independent average opinion scores.

The numerical values of the summary statistics obtained from the sets of $\overline{\text{OS}}_1$ values for each session and viewing condition are provided in Table 6 for

Table 6. Summary statistics of the sets of $\overline{\text{OS}}_1$ values obtained for P1 in each session and viewing condition.

		Mean	Med.	Mod.	SD	MAD_1	MAD_2	Skew.	Kurt.
	S1	3.64	3.88	4.50	0.93	0.75	0.63	−0.85	2.96
ST	S2	3.64	4.00	4.00	1.01	0.81	0.50	−0.62	2.72
	S3	3.63	4.00	4.50	1.01	0.85	0.50	−0.87	3.18
	S1	3.51	3.63	4.00	0.96	0.78	0.63	−0.81	2.77
SE	S2	3.35	3.50	4.00	0.89	0.72	0.63	−0.84	2.59
	S3	3.48	3.75	3.75	0.85	0.67	0.50	−0.86	2.91

Table 7. Summary statistics of the sets of $\overline{\text{OS}}_1$ values obtained for P2 in each session and viewing condition.

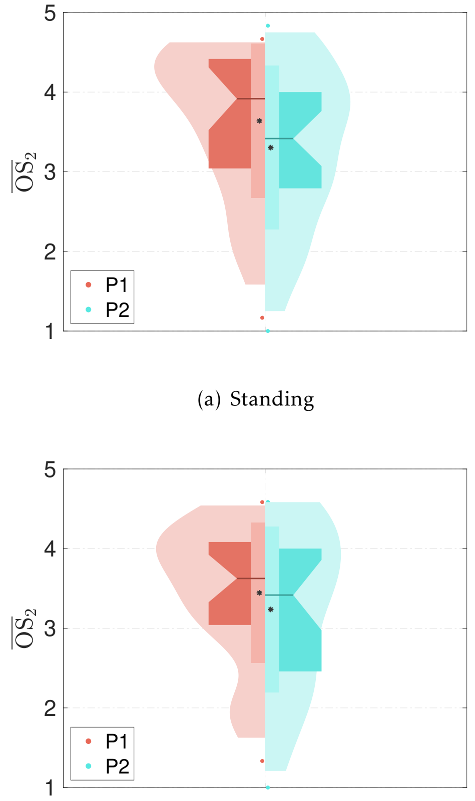
		Mean	Med.	Mod.	SD	MAD_1	MAD_2	Skew.	Kurt.
	S1	3.41	3.38	2.75	1.13	0.92	0.75	−0.37	2.41
ST	S2	3.33	3.50	4.00	0.98	0.82	0.63	−0.51	2.19
	S3	3.17	3.38	3.50	1.02	0.84	0.75	−0.55	2.51
	S1	3.31	3.63	3.75	1.07	0.90	0.75	−0.27	2.01
SE	S2	3.23	3.25	4.00	1.13	0.95	1.00	−0.32	2.31
	S3	3.17	3.25	3.25	1.00	0.80	0.75	−0.35	2.44

P1 and Table 7 for P2. Because averaging of the OS values was performed over the four different scenes for each resolution-QP pair, the mode and MAD_2 can also assume real values. It is observed that MAD_1 and MAD_2 become smaller than the SD indicating their ability of avoiding overweighing the tails of the distributions. Overall, the numerical values support the conjecture of consistent but distinct quality assessment signatures for each participant for the considered sets of content independent averaged opinion scores.

5.3. Statistical Analysis of Average Opinion Scores Over Video Scenes and Sessions

In addition to averaging the OSs over the four different 360° video scenes for a given resolution-QP pair, in this section, an additional level of averaging over all sessions is performed leading to average opinion scores $\overline{\text{OS}}_2$ as defined in (3). The sets of $\overline{\text{OS}}_2$ values represent the participants' quality assessment irrespective of the content of the 360° video scenes and the session conducted for a given viewing condition.

Figures 6(a)-(b) show the violin plots of the sets of $\overline{\text{OS}}_2$ values for each participant and viewing condition. Although the consistency of the quality assessment cannot be revealed any longer because of the averaging of opinion scores over the sessions in each viewing condition, the kernel fits still support the conjecture that each participant has their own distinct quality assessment signature. The violin plots also show that the means and medians of the sets of $\overline{\text{OS}}_2$ values for each viewing condition is higher for P1 than for P2. As



(a) Standing

(b) Seated

Figure 6. Violin plots of $\overline{\text{OS}}_2$ for each participant and viewing condition.

such, P1 tends to give higher quality ratings compared to P2 irrespective of the viewing condition.

Tables 8-9 provide the summary statistics of the sets of $\overline{\text{OS}}_2$ values for P1 and P2, respectively, for each viewing condition. The comparisons of the summary statistics also support the conjecture of each participant having their own quality assessment signature.

Table 8. Summary statistics of the sets of $\overline{\text{OS}}_2$ values for P1 in each viewing condition.

	Mean	Med.	Mod.	SD	MAD_1	MAD_2	Skew.	Kurt.
ST	3.64	3.92	4.58	0.97	0.80	0.62	−0.94	2.95
SE	3.44	3.63	1.92	0.88	0.70	0.54	−0.87	2.92

Table 9. Summary statistics of the sets of $\overline{\text{OS}}_2$ values for P2 in each viewing condition.

	Mean	Med.	Mod.	SD	MAD_1	MAD_2	Skew.	Kurt.
ST	3.30	3.42	2.83	1.03	0.85	0.62	−0.44	2.40
SE	3.24	3.42	3.83	1.04	0.88	0.92	−0.46	2.16

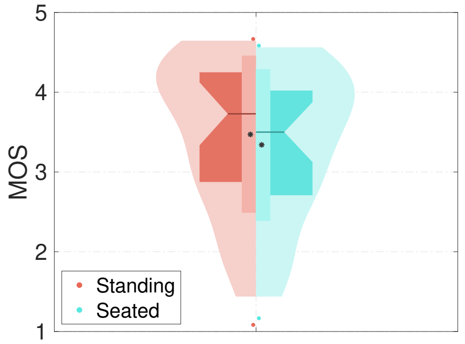


Figure 7. Violin plots of the sets of MOS values for each viewing condition irrespective of video content, session, and participant.

5.4. Statistical Analysis of Average Opinion Scores Over Video Scenes, Sessions, and Participants

An averaging of opinion scores is performed over all video scenes, sessions, and participants for each viewing condition. The results represent conventional MOS values which are commonly used with subjective tests on digital media quality assessment. Due to the three-fold averaging, only comparisons of the quality assessment between viewing conditions remain possible. Conjectures about quality assessment consistency throughout sessions and differences among participants are no longer supported.

Figure 7 shows the violin plot of the sets of MOS values as defined in (5) for standing and seated viewing while summary statistics are provided in Table 10. As can be seen from the figure, the mean and median of the MOS values are higher for ST compared to SE while the kernel fits for both viewing conditions become similar. In particular, the obtained skewness values suggest that the distribution of the MOS values for ST is slightly more left-tailed than for SE. These observations suggest that the participants generally tend to give ratings that are more leaned toward the higher end of the quality scale in ST compared to SE.

Table 10. Summary statistics of the sets of MOS values for each viewing condition irrespective of video content, session, and participant.

	Mean	Med.	Mod.	SD	MAD_1	MAD_2	Skew.	Kurt.
ST	3.47	3.73	2.08	0.98	0.82	0.67	−0.72	2.63
SE	3.34	3.50	2.71	0.95	0.79	0.77	−0.64	2.47

Finally, the statistics of the MOS values obtained in the full subjective test reported in [36] are compared with the results of the pilot study reported in this article. In particular, 30 participants (7 females and 23 males) of ages between 20 to 36 years took part in

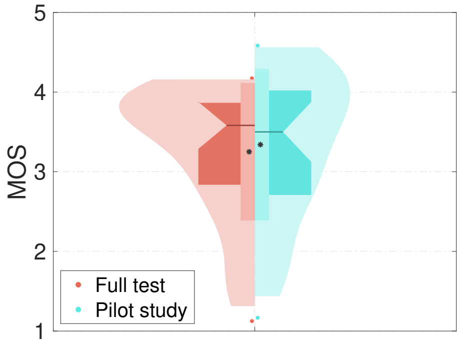


Figure 8. Violin plots of the sets of MOS values for the full subjective test and pilot study for seated viewing irrespective of video content, session, and participant.

the full subjective test with an average age of 25.37 years. In the full subjective test and the pilot study, the participants rated the same set of 360° video stimuli on an HTC Vive Pro HMD in seated viewing using the ACR method. However, while the visual stimuli in the pilot study were repeated over three sessions, they were shown only once in the full subjective test. The results for the full subjective test and pilot study shed light on the relevance of a subjective test with a few participants but repeated assessment compared to a conventional subjective test with a larger panel of participants.

Figure 8 and Table 11 show the violin plots and summary statistics for the sets of OSs obtained for the full subjective test and the pilot study. The figure and table indicate that the measures of central tendencies are similar in both test scenarios. As expected, the dispersion of the OSs in terms of SD, MAD_1, and MAD_2, is lower for the full subjective test with more participants compared to the pilot study. Also, the skewness and kurtosis become stronger for the full subjective test which results in enhancing the trends already revealed in the pilot study. These findings support the view of pilot studies being an integral component of an overall experimental design and a suitable means for conducting subjective tests under opportunity-limited conditions.

Table 11. Summary statistics of the sets of MOS values for the full subjective test (FT) and the pilot study (PS) for seated viewing irrespective of video content, session, and participant.

	Mean	Med.	Mod.	SD	MAD_1	MAD_2	Skew.	Kurt.
FT	3.25	3.58	3.05	0.86	0.70	0.49	−1.04	3.15
PS	3.34	3.50	2.71	0.95	0.79	0.77	−0.64	2.47

6. Statistical Analysis of Head Movements

Additional insights about the participants' consistency in 360° video quality assessment can be gained by analysing the head movements that were recorded for each participant, session, and viewing condition. In this context, the rotational head movements, i.e., yaw, pitch, and roll angles, represent the participants' exploration behavior of the 360° videos viewed on the HMD. In particular, yaw represents motion along the equator, pitch describes up and down movements between south and north pole, and roll captures circular movements with respect to the front view:

$$\begin{aligned} \text{Yaw} &\in [-180°, 180°] \\ \text{Pitch} &\in [-90°, 90°] \\ \text{Roll} &\in [-90°, 90°] \end{aligned}$$

Here, the rotational head movements are accumulated over the four 360° video scenes and then analysed with respect to the following data categories:

- Yaw, pitch, and roll angles for each participant, session, and viewing condition.
- Yaw, pitch, and roll angles accumulated over all sessions for each participant and viewing condition.
- Yaw, pitch, and roll angles accumulated over all sessions and participants for each viewing condition.

The statistical analysis of these data categories is performed in terms of CDFs and percentages of the rotational head movements falling into certain focus ranges. These focus ranges are defined here through visual inspection of the CDFs of the three rotational head movements and are denoted as follows:

$$\begin{aligned} \text{Yaw}[-60°, 60°] &:= \text{Yaw} \in [-60°, 60°] \\ \text{Pitch}[-30°, 30°] &:= \text{Pitch} \in [-30°, 30°] \\ \text{Roll}[-15°, 15°] &:= \text{Roll} \in [-15°, 15°] \end{aligned}$$

6.1. Head Movements for Each Participant, Session, and Viewing Condition

Figures 9-11 show the CDFs of the yaw, pitch, and roll rotations, respectively, obtained for each participant, session and viewing condition. It shall be mentioned that the yaw angle of 0° relates to the front view shown to the participants at the start of each session and the pitch angle of 0° relates to the equator. Although the roll angle can range from −90° to 90°, the results shown are limited to the range from −45° to 45° since the recorded data falls in this narrower range. The reason for not obtaining larger roll angles may be due to the difficulty to physically tilt the head in roll direction.

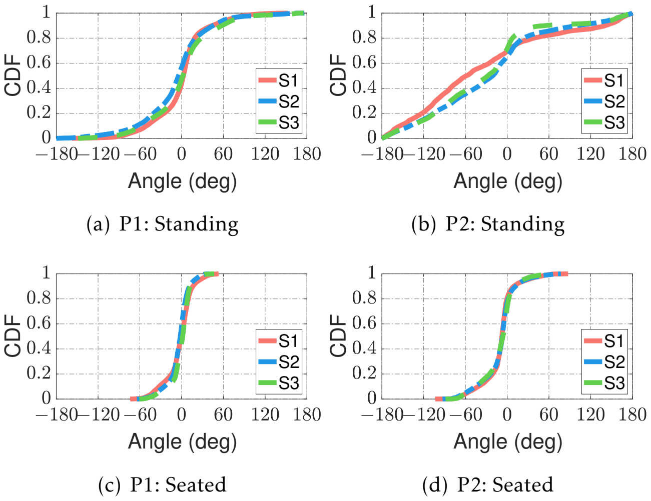
First, Figures 9(a)-(d) depict the CDFs of the yaw angles recorded for each participant, session, and viewing condition. It can be seen from these figures that the participants' exploration of the 360° videos regarding yaw rotations remains consistent throughout the sessions for a given viewing condition. However, for standing viewing, P2 explores the visual stimuli much wider compared to P1, i.e., making use of the entire yaw range [see Figures 9(a),(b)]. On the other hand, for seated viewing, both participants explore the scenes in a rather small yaw range with very similar CDFs of the yaw angle [see Figures 9(c),(d)].

Second, Figures 10(a)-(d) show the CDFs of the pitch angles recorded for each participant, session, and viewing condition. It can be observed that the pitch rotations also remain consistent throughout the sessions for each participant and given viewing condition. However, the scene exploration behavior regarding pitch rotations from the equator toward the north pole and south pole is much narrower compared to the yaw rotations along the equator between left and right direction from the front view. In addition, the participants tend to explore the shown visual stimuli more toward the north pole as indicated by the positive offsets of the CDFs of the pitch angles with respect to the equator. Similar as for the yaw angle, each participant has their distinct pitch rotation behavior, referred to as their exploration signature, throughout the sessions with P2 exploring a wider range of pitch angles.

Third, Figures 11(a)-(d) present the CDFs of the roll angles recorded for each participant, session, and viewing condition. Again, the exploration of the visual stimuli with roll rotations remains consistent for each participant throughout the sessions in standing viewing and slightly differs in seated viewing. Furthermore, P1 appears to be inclined tilting the head to the left (negative offset of CDFs) while P2 tends tilting the head to the right (positive offset of CDFs).

Tables 12-13 provide numerical values of the percentages of the yaw, pitch, and roll angles falling in the above defined focus ranges. As such, the higher the percentage in the range of a given rotational angle, the lower is the exploration toward angles outside this range. Clearly, confirming the consistency deduced from the CDFs, P2 explores the visual stimuli in standing viewing with much wider yaw rotations and slightly wider pitch directions compared to P1. The results also show the consistency of the scene exploration behavior throughout the sessions for each participant and viewing condition.

The distinct differences in scene exploration between participants regarding yaw rotations in standing viewing observed in Figures 9(a),(b) corresponds well with their quality assessment signatures show in Figures 4(a),(b). The narrower scene exploration of P1 corresponds to more skewed kernel fits to the histograms of OSs for all sessions [see Figure 9(a) and Figure 4(a)]. The wider scene exploration of P2 corresponds to more symmetric kernel fits to the histograms of OSs for all sessions [see Figure 9(b) and Figure 4(b)]. As such, it may be conjectured that participants' scene exploration behavior and quality assessment signatures for more relaxed viewing conditions and increased locomotion options are correlated.



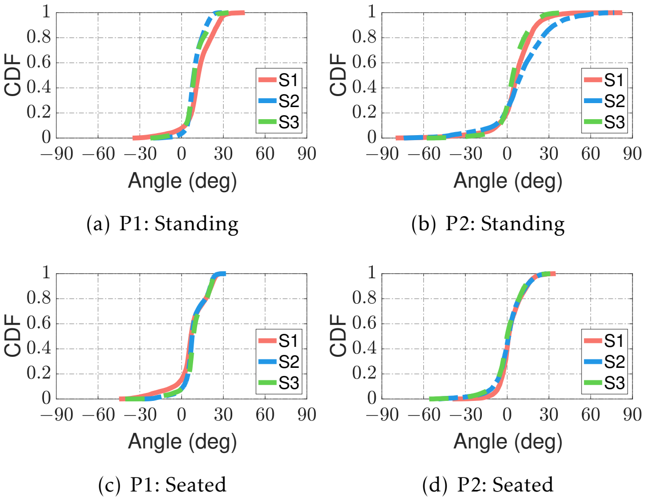
(a) P1: Standing

(b) P2: Standing

(c) P1: Seated

(d) P2: Seated

Figure 9. CDFs of the yaw angles for each participant, session, and viewing condition.



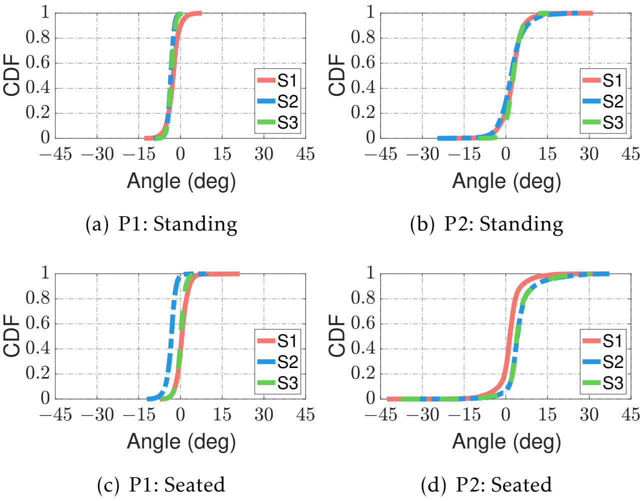
(a) P1: Standing

(b) P2: Standing

(c) P1: Seated

(d) P2: Seated

Figure 10. CDFs of the pitch angles for each participant, session, and viewing condition.



(a) P1: Standing

(b) P2: Standing

(c) P1: Seated

(d) P2: Seated

Figure 11. CDFs of the roll angles for each participant, session, and viewing condition.

Table 12. Percentages of yaw, pitch, and roll rotations falling in the focus ranges for P1, each session, and viewing condition.

		S1	S2	S3
Yaw $[-60°, 60°]$	ST	86.69%	82.90%	80.82%
	SE	99.55%	99.95%	99.97%
Pitch $[-30°, 30°]$	ST	97.48%	100.00%	99.58%
	SE	98.32%	99.95%	99.74%
Roll $[-15°, 15°]$	ST	100.00%	100.00%	100.00%
	SE	99.88%	100.00%	100.00%

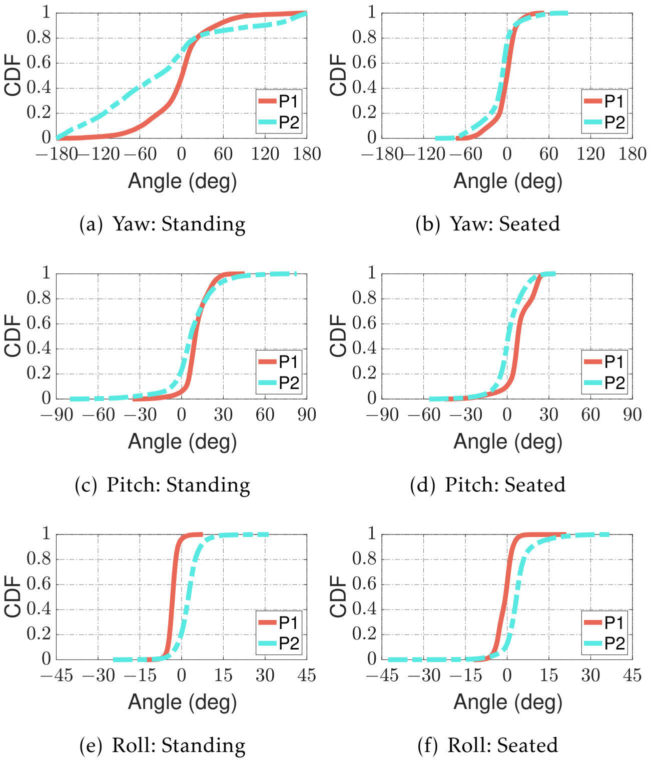
Table 13. Percentages of yaw, pitch, and roll angles falling in the focus ranges for P2, each session, and viewing condition.

		S1	S2	S3
Yaw $[-60°, 60°]$	ST	32.85%	50.41%	52.48%
	SE	95.31%	94.29%	95.44%
Pitch $[-30°, 30°]$	ST	93.88%	82.57%	97.84%
	SE	99.68%	98.41%	98.73%
Roll $[-15°, 15°]$	ST	99.35%	98.49%	99.99%
	SE	98.30%	94.94%	94.59%

6.2. Statistical Analysis of Head Movements Accumulated Over All Sessions

In this section, the rotational head movements are accumulated over all sessions for each participant and viewing condition. In this way, the 360° video exploration signature associated with each participant can be revealed irrespective of the session in a given viewing condition.

Figures 12(a)-(f) show the CDFs of the yaw, pitch, and roll angles accumulated over all sessions for each participant and viewing condition. Most notably, the CDFs in Figure 12(a) indicate that the exploration of the shown 360° videos with yaw rotations may significantly differ among participants in standing viewing compared to the accumulated pitch and roll rotations [see Figures 12(b)-(f)]. In particular, P1 explores the shown visual stimuli in a relatively small yaw range of around $[-60°, 60°]$ while P2 explores a much wider yaw range of around $[-180°, 180°]$. On the other hand, the accumulated yaw angles for seated viewing become very similar for both participants and are almost entirely kept within a range of $[-60°, 60°]$. As for the pitch rotations accumulated over all sessions, both participants kept their scene exploration within a range of about $[-30°, 30°]$ for both viewing conditions but with a slight inclination toward the north pole. The less steep transition from 0 to 1 observed in the CDF of the pitch angles for P2 indicates a slightly wider exploration of the visual stimuli between south pole and north pole compared to P1. Regarding roll rotations, the exploration of the visual stimuli becomes even more constrained for both participants residing within a range of about $[-15°, 15°]$.



(a) Yaw: Standing

(b) Yaw: Seated

(c) Pitch: Standing

(d) Pitch: Seated

(e) Roll: Standing

(f) Roll: Seated

Figure 12. CDFs of yaw, pitch, and roll angles accumulated over all sessions for each participant and viewing condition.

Table 14 supports the above conjecture through the percentages that the accumulated yaw, pitch, and roll angles over all sessions reside in the respective focus ranges for each participant and viewing condition. The numerical results clearly show that P1 limits their yaw rotations to 83.47% within the focus range of $[-60°, 60°]$

in standing viewing while P2 attends this range to only 45.07%. As for the remaining combinations of yaw, pitch, and roll rotations in standing and seated viewing, exploration within the focus ranges is kept above 90% for both participants.

In view of the above findings deduced from the head movement over all sessions, it is conjectured that participants have their distinct exploration signature when viewing 360° videos on HMDs. These exploration signatures can be rather different among participants for yaw rotations while differences are less striking for pitch and roll rotations. Similar as P2 explores a wider range of yaw angles in standing viewing compared to P1, the average opinion scores over all sessions for P2 spread a wider range of the quality scale compared to P1 [see Figure 6(a) and Tables 8-9]. As such, the higher degree of rotational freedom allowed in standing viewing compared to being constrained to a fixed chair in seated viewing appears not only to have an impact on participants' exploration signatures of visual stimuli on HMDs but can also influence the quality assessment of visual stimuli.

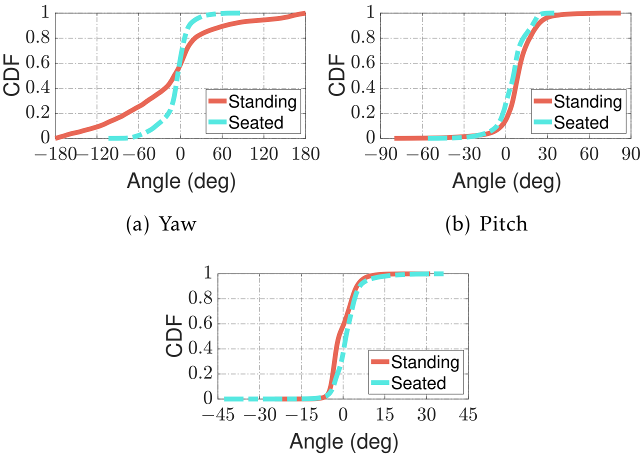
Table 14. Percentages of yaw, pitch, and roll angles falling in the focus ranges accumulated over all sessions for each participant and viewing condition.

	ST		SE	
	P1	P2	P1	P2
Yaw [−60°, 60°]	83.47%	45.07%	99.83%	95.01%
Pitch [−30°, 30°]	99.02%	91.43%	99.33%	98.94%
Roll [−15°, 15°]	100.00%	99.28%	99.96%	95.94%

6.3. Statistical Analysis of Head Movements Accumulated Over All Sessions and Participants

As a further means of condensing the recorded rotational head movements, yaw, pitch, and roll angles are accumulated over all sessions and participants. The obtained results allow a comparison of the exploration of the visual stimuli in standing and seated viewing irrespective of the session and participant.

Figures 13(a)-(c) show the CDFs of the yaw, pitch, and roll angles for these accumulated rotational head movements for both viewing conditions. Clearly, as can be seen from Figure 13(a), a much wider exploration of the visual stimuli is observed for the yaw angle for standing viewing compared to seated viewing. Further, the CDFs of the yaw angles of both viewing conditions progress almost symmetrically with respect to the yaw angle of 0° (front view at session start). Regarding the pitch angles, similar narrow exploration is observed for both viewing conditions with a positive offset of the CDFs toward the north pole. Similarly, the CDFs of the roll angles become even more narrow with little difference between standing and seated viewing.



(a) Yaw

(b) Pitch

(c) Roll

Figure 13. CDFs of yaw, pitch, and roll angles for both viewing conditions over all sessions and participants.

Table 15 shows the percentages that the accumulated yaw, pitch, and roll angles for standing and seated viewing fall into the focus ranges. These numerical results reinforce the findings deduced from the CDFs, i.e., accumulated yaw rotations are much wider for standing viewing with only 64.10% in the focus range compared to seated viewing with 97.41% in the focus range. The percentages of all other scene explorations being within the focus ranges are well above 90%.

Table 15. Percentages of yaw, pitch, and roll angles falling within the focus ranges for both viewing conditions accumulated over all sessions and participants.

	ST	SE
Yaw [−60°, 60°]	64.10%	97.41%
Pitch [−30°, 30°]	95.19%	99.14%
Roll [−15°, 15°]	99.63%	97.94%

7. Conclusions and Future Work

In this article, a pilot study on the consistency of 360° video quality assessment that was conducted under an opportunity-limited condition has been reported. The test procedure that was approved by the Corona group of the Blekinge Institute of Technology (BTH) and used in this pilot study has been provided (see Appendix A). Three repeated subjective tests for both standing and seated viewing were performed in which the participants rated 360° videos on an HMD using the ACR method. To reveal whether participants' quality assessment stays consistent or significantly changes over time, long breaks of a few months and short breaks of a day or a few hours were placed between sessions of each viewing condition. In each repeated session for a given viewing condition, 120 different 360° videos covering a wide range of quality levels were rated. Accordingly, each participant rated 360 visual stimuli in each viewing condition resulting in 720 visual stimuli over both viewing conditions. The opinion scores, head movements, eye tracking data, GSR data, time stamps, and demographic information about the participants recorded in this pilot study have been made available in the RQA360 dataset under the GitHub link in [29]. The RQA360 dataset allows the research community conducting meta-analyses with other existing or future public annotated datasets. In relation to the objectives pursued and the research questions posed in this article, the opinion scores given by each participant and their head movements during HMD exposure in each viewing condition have been analysed. The exploratory findings of this pilot study can be summarized as follows:

- **OS:** The histograms of the sets of OSs, their kernel fits, and summary statistics have shown that the quality assessment of the visual stimuli is kept consistent for each participant throughout the three sessions of each viewing condition. However, each participant has their distinct quality assessment signature.

- **$\overline{\textbf{OS}}_1$:** The above findings extend to the sets of average OSs over the four 360° video scenes confirming consistent quality assessment signatures for each participant. Furthermore, the violin plots and summary statistics indicate that the participants are more reluctant to give ratings toward the higher end of the quality scale in seated viewing. This behavior is thought to be due to participants not being distracted by other tasks, e.g., keeping the body in balance in standing viewing, but are primarily focused on the quality assessment task in seated viewing.

- **$\overline{\textbf{OS}}_2$:** The statistical analysis of the sets of average OSs over video scenes and sessions reveal that P1 tends to give higher ratings compared to P2 irrespective of the viewing condition. This finding also supports that the participants have their distinct quality assessment signature.

- **MOS:** The violin plots and summary statistics of the sets of average OSs over video scenes, sessions, and participants indicate that the participants lean more toward the higher end of the quality scale in standing viewing compared to seated viewing.

- **Pilot study:** A comparison of the quality assessment results for seated viewing obtained in this pilot study with those of a full subjective test on the same set of 360° videos supports the view of pilot studies being an important component of an overall experimental design. In particular, the measures of central tendencies of the pilot study and the full subjective test are similar while the measures of dispersion become lower for the full subjective test.

- **Accumulated OS:** The statistical analysis of the sets of accumulated OSs presented in Appendix B support the above exploratory finding that participants' follow consistently their distinct quality assessment signature throughout the sessions for each viewing condition. These results verify that the averaging of OSs does not has removed important information about the consistency of 360° video quality assessment.

- **Head movements:** The CDFs of the head movements for different levels of accumulated yaw, pitch, and roll angles have shown that the participants possess their own consistent but distinct exploration behavior of the 360° videos throughout the sessions for each viewing condition. This exploratory finding applies in particular to yaw rotations in standing viewing which is clearly different between the participants. In relation to the quality assessment, the higher degree of rotational freedom allowed in standing viewing compared to seated viewing on a fixed chair appears to also have an impact on participants' quality assessment signatures.

The large amount of data obtained and related exploratory findings have clearly shown the importance of a pilot study as part of an overall experimental design. The presented work has also shown the potential of using pilot studies as an efficient means of continuing in-person experiments under opportunity-limited conditions.

In view of pilot studies leading to exploratory findings, considerations toward larger subjective tests and areas of future work may be suggested as follows:

- Conduct subjective tests on the consistency of 360° video quality assessment engaging larger panels of participants and larger sets of 360° video scenes.

- Conduct larger subjective tests to investigate whether eye tracking data, GSR data, and rating durations provide further insights on the consistency of 360° video quality assessment.

- Perform meta-analysis exploring other existing or future annotated datasets on 360° video quality assessment in order to increase the precision of findings.

- Study the impact of viewing conditions such as fixed chair, half-swivel chair, full-swivel chair, couch, options of larger rotational and translational movements, and free walking on 360° video quality assessment.

Acknowledgments

This work was supported in part by the Knowledge Foundation, Sweden, through the ViaTecH and HINTS projects under contract numbers 20170056 and 20220068, respectively. This study was granted ethical approval (Dnr. 2018/863). We thank the volunteers who generously shared their time to participate in the pilot study. We thank Kjell Brunnström for the fruitful discussions and advise on conducting subjective tests under opportunity-limited conditions. Finally, we thank our colleagues in the ViaTecH and HINTS projects for feedback.

References

[1] Aloqaily, M., Bouachir, O., Karray, F., Al Ridhawi, I. and Saddik, A.E. (2023) Integrating Digital Twin and Advanced Intelligent Technologies to Realize the Metaverse. *IEEE Consumer Electronics Magazine* **12**(6): 47–55.

[2] Han, Y., Niyato, D., Leung, C., Miao, C. and Kim, D.I. (2022) A Dynamic Resource Allocation Framework for Synchronizing Metaverse with IoT Service and Data. In *Proc. IEEE Int. Conf. on Communications* (Seoul, Republic of Korea): 298–301.

[3] 3GPP TR 26.918 V16.0.0 (2018) *Virtual Reality (VR) Media Services Over 3GPP (Release 16)*, 3rd Generation Partnership Project, Technical Specification Group Services and System Aspects.

[4] Yrjölä, S., Ahokangas, P. and Matinmikko-Blue, M. (Jun. 2020) *White Paper on Business of 6G. (6G Research Visions, No. 3)*, University of Oulu, Finland.

[5] Duanmu, Z., Zeng, K., Ma, K., Rehman, A. and Wang, Z. (2017) A Quality-of-Experience Index for Streaming Video. *IEEE Journal of Selected Topics in Signal Processing* **11**(1): 154–166.

[6] Hu, Y., Elwardy, M. and Zepernick H.-J. (2021) On the Effect of Standing and Seated Viewing of 360° Videos on Subjective Quality Assessment: A Pilot Study. *Computers* **10**(6): 1–28.

[7] Liu, R., Peng, C., Zhang, Y., Husarek, H. and Yu, Q. (2021) A Survey of Immersive Technologies and Applications for Industrial Product Development. *Computers & Graphics* **100**: 137–151.

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

[9] Recommendation ITU-R BT.500-13 (2012) *Methodology for the Subjective Assessment of the Quality of Television Pictures*, International Telecommunication Union, Geneva, Switzerland.

[10] Recommendation ITU-T BT.1788 (2007) *Methodology for the Subjective Assessment of Video Quality in Multimedia Applications*, International Telecommunication Union, Geneva, Switzerland.

[11] Recommendation ITU-T P.910 (2008) *Subjective Video Quality Assessment Methods for Multimedia Applications*, International Telecommunication Union, Geneva, Switzerland.

[12] Recommendation ITU-T P.915 (2016) *Subjective Assessment Methods for 3D Video Quality*, International Telecommunication Union, Geneva, Switzerland.

[13] Recommendation ITU-T P.919 (2020) *Subjective Test Methodologies for 360° Video on Head-Mounted Displays*, International Telecommunication Union, Geneva, Switzerland.

[14] Recommendation ITU-T P.1320 (2022) *Quality of Experience Assessment of Extended Reality Meetings*, International Telecommunication Union, Geneva, Switzerland.

[15] J. Gutierrez, et al. (2022) Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360 Videos: ITU-T Rec. P.919. *IEEE Trans. Multimedia* **24**: 3087–3100.

[16] Feil-Seifer, D., Haring, K.S., Rossi, S., Wagner, A.R. and Williams, T. (2020) Where to Next? The Impact of COVID-19 on Human-Robot Interaction Research. *ACM Trans. Human-Robot Interaction* **10**(1): 1–7.

[17] A. Steed, et al. (2020) Evaluating Immersive Experiences During Covid-19 and Beyond. *Interaction* **27**(4): 62–67.

[18] Spang, R.P. and Pieper, K. (2021) Durchführung von psychophysiologischen und subjektiven Experimenten während einer Pandemie. *ITG News* (2): 13–14.

[19] Zepernick H.-J., Pieper, K., Spang, R.P., Engelke, U., Hirth, M. and Naderi, B. (2021) On the Impact of COVID-19 on Subjective Digital Media Quality Assessment. In *Proc. IEEE Int. Workshop on Multimedia Signal Processing* (Tampere, Finland): 1–6.

[20] Perez, P., Janowski, L., Garcia, N. and Pinson, M. (2022) Subjective Assessment Experiments That Recruit Few Observers With Repetitions (FOWR). *IEEE Trans. Multimedia* **24**: 3442–3454.

[21] Elwardy, M., Zepernick H.-J. and Hu, Y. (2021) 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): 71–74.

[22] Elwardy, M., Zepernick H.-J., Chu, T.M.C. and Hu, Y. (2021) On the Opinion Score Consistency in Repeated 360° Video Quality Assessment for Standing and Seated Viewing on Head-Mounted Displays. In *Proc. IEEE Int. Conf. on Signal Processing and Commun. Systems* (Sydney, Australia): 1–10.

[23] Elwardy, M., Zepernick H.-J., Sundstedt, V. and Hu, Y. (2019) 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): 40–49.

[24] Beihang University, School of Electronic and Information Engineering, Beijing, China *VQA-ODV. 2017, Accessed on: 12. Sep. 2020*. URL https://github.com/Archer-Tatsu/VQA-ODV.

[25] Li, C., Xu, M. and Wang, Z. (2018) Bridge the Gap Between VQA and Human Behavior on Omnidirectional Video: A Large-Scale Dataset and a Deep Learning Model. In *Proc. ACM Int. Conf. on Multimedia* (Seoul, Republic of Korea): 932–940.

[26] Zhang, Y., Wang, Y., Liu, F., Liu, Z., Li, Y., Yang, D. and Chen, Z. (2018) Subjective Panoramic Video Quality Assessment Database for Coding Applications. *IEEE Trans. Broadcast.* **64**(2): 42–51.

[27] FFmpeg *H.264 Video Encoding Guide. 2018, Accessed on: 12. Sep. 2020.* URL https://trac.ffmpeg.org/wiki/Encode/H.264#crf.

[28] FFmpeg *FFmpeg and H.265 Encoding Guide. 2018, Accessed on: 12 Sep. 2020.* URL https://trac.ffmpeg.org/wiki/Encode/H.265.

[29] Blekinge Institute of Technology, Karlskrona, Sweden *RQA360. 2023, Accessed on: 28. Aug. 2023.* URL https://github.com/MajedElwardy/RQA360.

[30] Dodge, Y. (2008) *The Concise Encyclopedia of Statistics* (New York, NY, USA: Springer).

[31] Telab, N.N. *What Scientific Idea is Ready for Retirement?, Accessed on: 17. Jul. 2023.* URL https://web.archive.org/web/20140116031136/http://www.edge.org/response-detail/25401.

[32] The MathWorks Inc. (2023) *Statistics and Machine Learning Toolbox (R2021b)*, Natick, MA, USA. URL https://www.mathworks.com.

[33] Westfall, P.H. (2014) Kurtosis as Peakedness, 1905-2014. R.I.P. *The American Statistical Association* **68**(3): 191–195.

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

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

[36] Elwardy, M., Hu, Y., Zepernick H.-J., Chu, T.M.C. and Sundstedt, V. (2020) Comparison of ACR Methods for 360° Video Quality Assessment Subject to Participants' Experience with Immersive Media. In *Proc. Int. Conf. on Signal Process. and Commun. Systems* (Adelaide, Australia): 1–10.

[37] Blekinge Institute of Technology *Visual and Interactive Computing Laboratory (ViaLab)*. URL https://a.bth.se/viatech-synergy/vialabs/.

[38] Brunnström, K., Andrén, B., Schenkman, B., Djupsjöbacka, A. and Hamsis, O. (2020) *Recommended Precautions Because of Covid-19 for Perceptual, Behavioural, Quality and User Experiments with Test Persons in Indoor Labs*, RISE Report:2020:84, Stockholm, Sweden.

Appendix A. Test Procedure Under COVID-19 Conditions

Recommendations were established for conducting pilot studies under COVID-19 conditions within the Visual and Interactive Computing Laboratory (ViaLAB) [37] at the Department of Computer Science of BTH, Karlskrona, Sweden. A subset of the guidelines released by RI.SE Digital Systems Networks in [38] were adapted, amended as needed, and then approved by the BTH Corona group as follows (Italic font with quotation marks: Guideline quoted from [38]; Italic font: Guideline adapted from [38] with rewording; Times roman font: Amendment):

- *The test leader should prepare the experiment in advance to reduce the interaction time with the test person.*
- *The test leader should wash and disinfect hands thoroughly (for 30 s or more).*
- *The test leader must wear a visor and gloves.*
- *The test leader should be ready to receive the test person wearing a visor and gloves.*
- *"Physical distancing is to be maintained; the test leader should not shake hands with the test persons when they arrive and leave."*
- *"The test person will only meet the test leader. Communication should be done at a safe distance, and preferably through electronic means."*
- *"Verbal agreement shall be obtained from both the test person and the test leader that he or she is healthy (asymptomatic)."*
- *The test person should wash and disinfect their hands thoroughly (for 30 s or more).*
- *The test person is to be instructed to use gloves, as is most suitable for the experiment.*
- *"The test person is to be instructed to use a face mask or a visor, as is most suitable for the experiment."*
- *"A safe physical distance should be kept as much as is possible when giving the test instructions. Some test equipment may require the test leader to be closer than 1 meter but then preferably in as short a time as possible."*
- Different test persons should not be tested on the same day.
- *"Products and other equipment as well as the testing room with its furniture should be disinfected and cleaned using appropriate methods between sessions."*
- *The testing room should be ventilated between test sessions or air purifiers with preferably HEPA-13 or 14 filters could be run in the lab between test sessions.*
- *"Ideally, the test leader should stay in an adjacent lab/room during the tests. If the test leader is in the same room, a plexiglass shield should be used as a partition."*

- *"The equipment that is used shall be adapted as much as is possible to the test conditions and Covid-19 safety., i.e., keyboards, computer mice, etc."*
- *"When using VR/AR glasses, headsets, etc. for respective tests, they are to be cleaned between test sessions with disinfectants or UV-C light and ozone."*
- No breaks are scheduled during a test session.

Appendix B. Statistical Analysis of Sets of Opinion Scores and Accumulated Opinion Scores

This appendix provides an additional statistical analysis of sets of opinion scores and accumulated opinion scores in terms of histograms, kernel fits, and summary statistics. Recall that different levels of averaging of the recorded OSs have been performed in Section 5 resulting in $\overline{\text{OS}}_1$, $\overline{\text{OS}}_2$, and MOS values. Similarly, the following different levels of accumulated opinion scores are considered and analyzed in this appendix:

- Opinion scores for each session, participant, and viewing condition (same data as in Section 5-A but in different groupings).
- Opinion scores accumulated over all sessions for each participant and viewing condition.
- Opinion scores accumulated over all sessions and participants for each viewing condition.

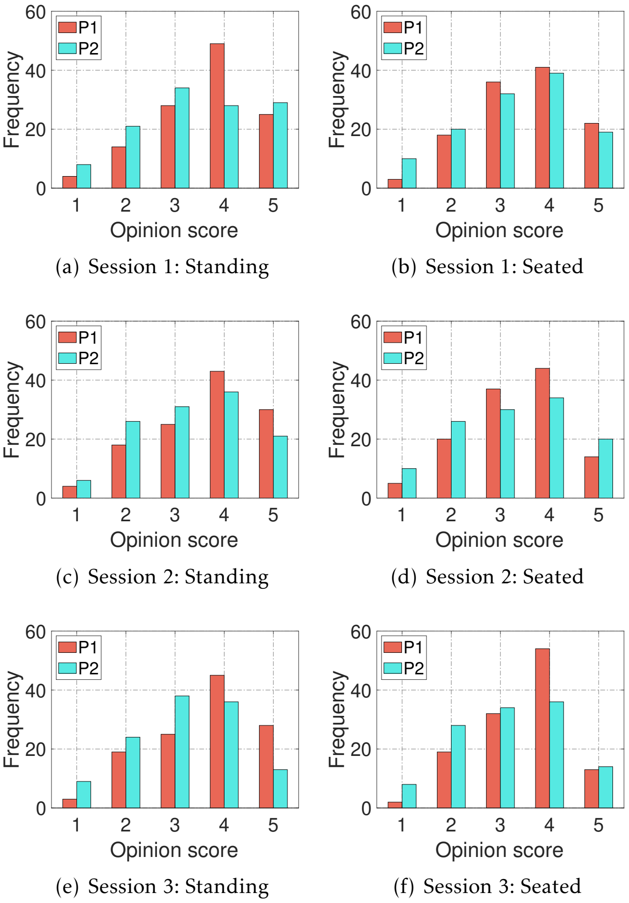
The statistical analysis reported in this appendix supports the conjectures given in Section 5 and verifies that the averaging does not has removed important information about the consistency of 360° video quality assessment.

B.1. Statistical Analysis of Opinion Scores

Figures B.1(a)-(f) and Figures B.2(a)-(f) show the histograms of opinion scores and related kernel fits, respectively, for each session, participant, and viewing condition. Especially, the kernel fits to the histograms show that each participant has their own quality assessment signature throughout the sessions in both viewing condition. Because Figures B.1-B.2 are based on the same data as that shown in Figures 3-4 but only differently grouped, the summary statistics presented in Tables 4-5 apply here as well.

B.2. Statistical Analysis of Opinion Scores Accumulated Over All Sessions for Each Participant and Viewing Condition

An accumulation over all sessions is achieved by adding up the respective frequencies for each of the five possible opinion scores reported in Figures B.1(a),(c),(e) and Figures B.1(b),(d),(f). As such, the respective frequencies of the now accumulated opinion scores are larger as shown in Figures B.3(a),(c). The histograms of these sets of accumulated OSs, their kernel fits, and summary statistics verify that both participants have their distinct quality assessment signature in both viewing condition [see Figures B.3(a)-(d) and Tables B.1-B.2]. In particular, the skewness and kurtosis in Tables B.1-B.2 indicate different shapes of the histograms and their kernel fits for P1 and P2.

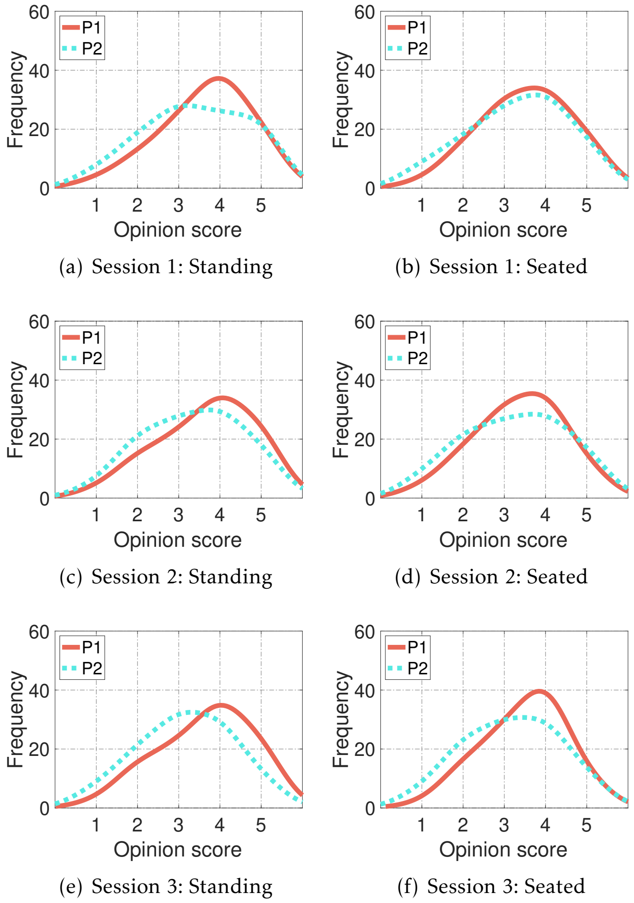


(a) Session 1: Standing
(b) Session 1: Seated
(c) Session 2: Standing
(d) Session 2: Seated
(e) Session 3: Standing
(f) Session 3: Seated

Figure B.1. Histograms of sets of OSs for each participant, session, and viewing condition.

B.3. Statistical Analysis of Opinion Scores Accumulated Over All Sessions and Participants for Each Viewing Condition

The accumulation of opinion scores over all sessions and participants is achieved by adding up the respective frequencies for each of the five possible opinion scores reported in Figures B.3(a),(c). Figures B.4(a)-(b) show the histograms of the obtained sets of accumulated OSs and their kernel fits while the summary statistics are given in Table B.3. The results indicate only small



(a) Session 1: Standing
(b) Session 1: Seated
(c) Session 2: Standing
(d) Session 2: Seated
(e) Session 3: Standing
(f) Session 3: Seated

Figure B.2. Kernel fits to the histograms of OSs for each participant, session, and viewing condition.

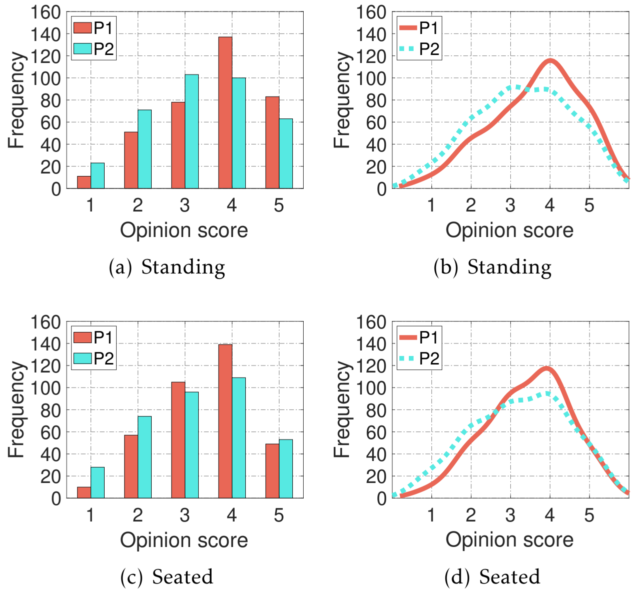
Table B.1. Summary statistics of sets of OSs accumulated over all sessions for P1 and each viewing condition.

	Mean	Med.	Mod.	SD	MAD$_1$	MAD$_2$	Skew.	Kurt.
ST	3.64	4.00	4.00	1.08	0.90	1.00	−0.52	2.50
SE	3.44	4.00	4.00	1.00	0.85	1.00	−0.33	2.52

Table B.2. Summary statistics of sets of OSs accumulated over all sessions for P2 and each viewing condition.

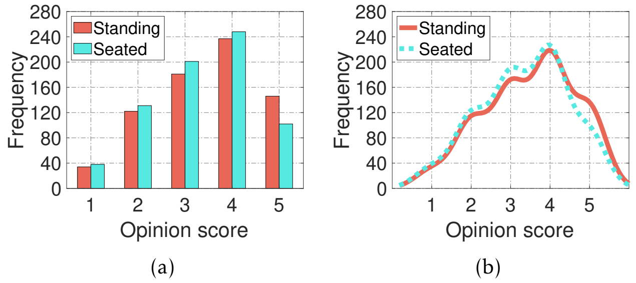
	Mean	Med.	Mod.	SD	MAD$_1$	MAD$_2$	Skew.	Kurt.
ST	3.30	3.00	3.00	1.16	0.98	1.00	−0.18	2.17
SE	3.24	3.00	4.00	1.16	0.98	1.00	−0.20	2.17

differences between the viewing conditions with the accumulated OSs for standing viewing slightly tending toward the higher end of the quality scale.



(a) Standing
(b) Standing
(c) Seated
(d) Seated

Figure B.3. Histograms of sets of OSs accumulated over all sessions and their kernel fits for each participant and viewing condition.



(a)
(b)

Figure B.4. Histograms of sets of OSs accumulated over all sessions and participants for both viewing conditions and their kernel fits.

Table B.3. Summary statistics of sets of OSs accumulated over all sessions and participants for both viewing conditions.

	Mean	Med.	Mod.	SD	MAD$_1$	MAD$_2$	Skew.	Kurt.
ST	3.47	4.00	4.00	1.13	0.97	1.00	−0.35	2.27
SE	3.34	3.00	4.00	1.09	0.92	1.00	−0.29	2.35