Gutiérrez, Jesús; Pérez, Pablo; Orduna, Marta; Singla, Ashutosh; Cortés, Carlos; Mazumdar, Pramit; Viola, Irene; Brunnström, Kjell; Battisti, Federica; Cieplińska, Natalia; Juszka, Dawid; Janowski, Lucjan; Leszczuk, Mikołaj; Adeyemi-Ejeye, Anthony; Hu, Yaosi; Chen, Zhenzhong; Wallendael, Glenn Van; Lambert, Peter; Díaz, César; Hedlund, John; Hamsis, Omar; Fremerey, Stephan; Hofmeyer, Frank; Raake, Alexander; César, Pablo; Carli, Marco; García, Narciso

Subjective evaluation of visual quality and simulator sickness of short 360° videos: ITU-T Rec. P.919

Original published in:	IEEE transactions on multimedia / Institute of Electrical and Electronics EngineersNew York, NY : Institute of Electrical and Electronics Engineers. - 24 (2022), p. 3087-3100.
Original published:	2021-07-05
ISSN:	1941-0077 ; 1520-9210
DOI:	10.1109/TMM.2021.3093717
[*Visited:*	2023-05-04]

Subjective Evaluation of Visual Quality and Simulator Sickness of Short 360° Videos: ITU-T Rec. P.919

Jesús Gutiérrez, Pablo Pérez, Marta Orduna, Ashutosh Singla, Carlos Cortés, Pramit Mazumdar, Irene Viola, Kjell Brunnström, Federica Battisti, Natalia Cieplińska, Dawid Juszka, Lucjan Janowski, Mikołaj Leszczuk, Anthony Adeyemi-Ejeye, Yaosi Hu, Zhenzhong Chen, Glenn Van Wallendael, Peter Lambert, César Díaz, John Hedlund, Omar Hamsis, Stephan Fremerey, Frank Hofmeyer, Alexander Raake, Pablo César, Marco Carli, and Narciso García

***Abstract*—Recently an impressive development in immersive technologies, such as Augmented Reality (AR), Virtual Reality (VR) and 360° video, has been witnessed. However, methods for quality assessment have not been keeping up. This paper studies quality assessment of 360° video from the cross-lab tests (involving ten laboratories and more than 300 participants) carried out by the Immersive Media Group (IMG) of the Video Quality Experts Group (VQEG). These tests were addressed to assess and validate subjective evaluation methodologies for 360° video. Audiovisual quality, simulator sickness symptoms, and exploration behavior were evaluated with short (from 10 seconds to 30 seconds) 360° sequences. The following factors' influences were also analyzed: assessment methodology, sequence duration, Head-Mounted Display (HMD) device, uniform and non-uniform coding degradations, and simulator sickness assessment methods. The obtained results have demonstrated the validity of Absolute Category Rating (ACR) and Degradation Category Rating (DCR) for subjective tests with 360° videos, the possibility of using 10-second videos (with or without audio) when addressing quality evaluation of coding artifacts, as well as any commercial HMD (satisfying minimum requirements). Also, more efficient methods than the long Simulator Sickness Questionnaire (SSQ) have been proposed to evaluate related symptoms with 360° videos. These results have been instrumental for the development of the ITU-T Recommendation P.919. Finally, the annotated dataset from the tests is made publicly available for the research community.**

***Index Terms*—Quality of experience, 360° video, subjective test, methodology, simulator sickness, dataset.**

Manuscript received January 12, 2021; revised April 17, 2021; accepted June 11, 2021. Date of publication July 5, 2021; date of current version June 9, 2022. The work of Jesús Gutiérrez, Marta Orduna, Carlos Cortés, César Díaz, and Narciso García was supported in part by the Ministerio de Ciencia, Innovación y Universidades (AEI/FEDER) of the Spanish Government under Project TEC2016-75981 (IVME). The work of Jesús Gutiérrez was also supported in part by Juan de la Cierva fellowship (IJC2018-037816). The work of Pablo Pérez work was supported in part by the Spanish Administration Agency CDTI under Project IDI-20200861 (AMATISTA). The work of Ashutosh Singla, Stephan Fremerey, Frank Hofmeyer, and Alexander Raake was supported in part by the CYTEMEX project funded by the Free State of Thuringia, Germany (FKZ: 2018-FGI-0019). The work of Irene Viola and Pablo César work was supported in part by the European Commission as part of the H2020 program, under Grant 762111, "VRTogether" (http://vrtogether.eu/). The work of Kjell Brunnström, John Hedlund, and Omar Hamsis was supported in part by Vinnova (Sweden's Innovation Agency) in the Celtic-Next project 5G Perfecta (2018-00735). The work of Lucjan Janowski and Mikołaj Leszczuk was supported in part by the Norwegian Financial Mechanism 2014-2021 (Project number: 2019/34/H/ST6/00599). The work of Anthony Adeyemi-Ejeye was supported in part by the NVIDIA Corporation. The work of Yaosi Hu and Zhenzhong Chen was supported in part by the National Natural Science Foundation of China under Grant 61771348. The work of Glenn Van Wallendael and Peter Lambert was supported in part by the research project imec ICON ILLUMINATE HBC.2018.0201. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. L. Fang. *(Corresponding author: Jesús Gutiérrez.)*

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TMM.2021.3093717.

Digital Object Identifier 10.1109/TMM.2021.3093717

Jesús Gutiérrez, Marta Orduna, Carlos Cortés, César Díaz, and Narciso García are with the Grupo de Tratamiento de Imágenes, Information Processing and Telecommunications Center and Escuela Técnica Superior de Ingenieros de Telecomunicación, Universidad Politécnica de Madrid, 28040 Madrid, Spain (e-mail: jesus.gutierrez@upm.es; moc@gti.ssr.upm.es; ccs@gti.ssr.upm.es; cdm@gti.ssr.upm.es; narciso@gti.ssr.upm.es).

Pablo Pérez is with the Nokia Bell Labs, 28050 Madrid, Spain (e-mail: pablo.perez@nokia-bell-labs.com).

Ashutosh Singla, Stephan Fremerey, Frank Hofmeyer, and Alexander Raake are with the Audiovisual Technology Group, Technical University of Ilmenau, 98693 Ilmenau, Germany (e-mail: ashutosh.singla@tu-ilmenau.de; stephan.fremerey@tu-ilmenau.de; frank.hofmeyer@tu-ilmenau.de; alexander.raake@tu-ilmenau.de).

Pramit Mazumdar is with the Department of Computer Science and Engineering, Indian Institute of Information Technology Vadodara, 382028 Gandhinagar, India (e-mail: pramit.mazumdar@iiitvadodara.ac.in).

Irene Viola and Pablo César are with the Centrum voor Wiskunde en Informatica,, XG 1098 Amsterdam, The Netherlands (e-mail: irene.viola@cwi.nl; P.S.Cesar@cwi.nl).

Kjell Brunnström is with the RISE Research Institutes of Sweden AB, Kista 16440, Sweden, and also with the Mid Sweden University, 85170 Sundsvall, Sweden (e-mail: kjell.brunnstrom@ri.se).

John Hedlund and Omar Hamsis are with the RISE Research Institutes of Sweden AB, 16440 Kista, Sweden (e-mail: john.hedlund@ri.se; omar.hamsis@ri.se).

Federica Battisti is with the Department of Information Engineering, University of Padova, 35131 Padova, Italy (e-mail: federica.battisti@unipd.it).

Natalia Cieplińska, Dawid Juszka, Lucjan Janowski, and Mikołaj Leszczuk are with the AGH University of Science and Technology, 30059 Kraków, Poland (e-mail: natacalia@gmail.com; juszka@agh.edu.pl; janowski@kt.agh.edu.pl; leszczuk@agh.edu.pl).

Anthony Adeyemi-Ejeye is with the Innovative Media Laboratory, Department of Music and Media, University of Surrey, GU2 7XH Guildford, U.K. (e-mail: femi.ae@surrey.ac.uk).

Yaosi Hu and Zhenzhong Chen are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: ys_hu@whu.edu.cn; zzchen@whu.edu.cn).

Glenn Van Wallendael and Peter Lambert are with the Department of Electronics and Information Systems, Ghent University - imec, 9000 Ghent, Belgium (e-mail: glenn.vanwallendael@ugent.be; peter.lambert@ugent.be).

Marco Carli is with the Department of Engineering, Universitá degli Studi Roma Tre, 00154 Rome, Italy (e-mail: marco.carli@uniroma3.it).

I. Introduction

RECENT years have witnessed many impressive technological and scientific advances in fields such as Augmented Reality (AR), Virtual Reality (VR), and immersive communication systems, such as 360° video, multiview video, immersive audio-haptic systems, etc. The availability of these technologies is paving the way to some extremely appealing new applications and services in different domains, such as entertainment, communications, social relations, healthcare, and industry. The interest in such services and their potential impact on society and economy are enormous. The technology revolution led to a significant growth of the telepresence/AR/VR market, which is predicted to become a multi-billion business [1], [2].

The users of these new technologies can explore and experience the contents in a more interactive and personalized way than previous technologies [3], intensifying their sensation of "being there". These new perceptual dimensions and interaction behaviors provided by immersive technologies, together with the new challenges concerning the whole processing chain (i.e., from acquisition to rendering), require an exhaustive study and understanding in order to satisfy the demands and expectations of the users [4]. In this sense, the research on evaluation of Quality of Experience (QoE) allows, on one side, the extraction of useful outcomes to optimize the audiovisual systems, and on the other side, to identify possible inconveniences that deteriorate the user experience and hinder the success of emerging technologies, such as 360° video [5].

To evaluate the end-users' QoE, subjective experiments are usually performed following standard methodologies, such as the International Telecommunication Union (ITU) recommendations, designed explicitly for particular technologies or services. Although several subjective experiments with immersive media technologies have been already published in the literature, the majority apply assessment methodologies initially designed for 2D video, and there is still the lack of an international recommendation for 360° video, like the ones existing for images [6], video [7], and 3D video [8]. Thus, a revision of the QoE evaluation methods designed for previous technologies is required to develop robust and reliable methodologies for immersive media technologies.

Also, to foster the research and development of immersive media, it is essential to access databases with appropriate contents and users' data from subjective experiments. This allows the reproducibility of the research, the comparison of results from different tests, and the development of models to estimate the QoE of the users of immersive technologies.

Taking this into account, a cross-lab test was carried out within the Immersive Media Group (IMG)[1] of the Video Quality Experts Group (VQEG) with the following objectives:

- To validate and recommend test methodologies to evaluate the audiovisual quality of 360° videos, taking into account:
  - The duration of the test sequences, considering short ones (10-30 seconds). Longer sequences, which may entail the evaluation of other aspects such as presence, immersion, etc. are left for future work.
  - Influence factors such as the Head-Mounted Display (HMD), the source content characteristics, and the impact of uniform and non-uniform artifacts.
- To recommend methods to assess simulator sickness, considering:
  - One multi-item questionnaire (SSQ or derivation from it), or one single-question item.
  - When/how to assess simulator sickness and how to process and analyze the results.
- To generate and publish a dataset of subjectively assessed 360° content for future research, which is available in the databases section of the VQEG website.

The fulfilment of these objectives has supported the development of the recent ITU-T Recommendation P.919 [9]. This recommendation provides guidelines for subjective test methodologies for 360° video on HMDs, in line with the recommendations ITU-R BT.500 [6], ITU-T P.910 [10], and ITU-T P.913 [7] for 2D video, and ITU-T P.915 [8] for 3D video. This paper presents the details of the subjective experiment and the results that supported the majority of the guidelines included in the new recommendation.

The rest of the paper is structured as follows. Section II provides an overview of related works in the state-of-the-art. Section III provides a detailed description of the test setup, whose main results are provided in Section IV. Finally, Section V exposes the main conclusions of the work.

II. Related Work

A. QoE and Immersive Media Technologies

QoE [11] is "the degree of delight or annoyance of the user with an application or service. It results from the fulfilment of his or her expectations concerning the utility and enjoyment of the application or service in the light of the user's personality and current state," as defined by EU Cost Action 1003 Qualinet [12], which has now also been standardised by the ITU in the recommendation P.10/G.100 [13]. This definition goes beyond the traditional QoE research (carried out by the telecommunication community) and overlaps with the User Experience (UX)[2] research tradition from the Human-Computer Interaction (HCI) community. In fact, the QoE community is in the process of embracing some of the more user-centric and UX-like methods, especially due the emergence of immersive media technologies, which are offering the end users more interactive and personalized experiences. Thus, the term Quality of User eXperience (QUX) has been introduced [14]. In this sense, the role of QUX is twofold. On the one hand, the technologies developed for the next-generation of immersive communications need to be user-centric. On the other hand, new standardised evaluation methodologies are needed to assess developed technologies' QUX.

[1] [Online]. Available: www.its.bldrdoc.gov/vqeg/projects/immersive-media-group

[2] QoE is strongly related to but different from the field of UX, which also focuses on users' experiences with services.

When services become truly immersive, the impact of the influence factors, as described in Reiter *et al.* [15], becomes crucial to consider. In this regard, in addition to human and perceptual factors, several system factors influence the users' new immersive experiences [16]. For instance, hardware solutions for immersive environments are also associated with UX issues, such as large and bulky HMDs leading to user discomfort (e.g., eye-strain, dizziness, fatigue and nausea) [17]. The acquisition, compression, and transmission of immersive data are also emerging problems that must be solved efficiently to correctly deal with the target applications [5]. In this sense, the research community is active in studying the QoE of the users of immersive technologies, including international groups and organisations, such as Qualinet [4], the Moving Picture Expert Group (MPEG) [18], the Institute of Electrical and Electronics Engineers (IEEE) [19]–[21], the ITU [16], and the VQEG.

The VQEG [22], [23] is an international and independent organisation of technical experts in perceptual video quality assessment from industry, academia and government organisations. The main goals of the VQEG are to advance the field of perceptual video quality assessment, establish best practices for subjective experiments, conduct large scale subjective experiments, and evaluate new assessment metrics and models. The VQEG pursues several objectives, and one of them is to support standardisation work. Based on results produced by the VQEG, more than twenty-five international recommendations on video quality have been approved by the ITU [24], [25]. The VQEG has initiated related research projects: IMG, Quality Assessment for Computer Vision Applications (QaCoViA), and Key Performance Indicators for 5 G (5GKPI). This research could impact QoE and QUX standardisation in global bodies such as ITU-T [16] or IEEE [19]–[21].

In particular, the IMG researches on quality assessment of immersive media, with the main goals of generating immersive media content datasets, validating subjective test methods, and providing guidelines for QoE evaluation of immersive systems, including 360° content, virtual/augmented/mixed reality, 3D content, free-viewpoint video, multiview technologies, light field content, etc. Apart from welcoming contributions related to any of these topics, during the last years several members of the IMG have been jointly working on a test plan for the cross-lab tests presented in this paper. The work presented here has been instrumental in the development of the recent recommendation ITU-T P.919 [9].

B. Subjective and Objective Methods for Quality Assessment of 360° Videos

Different subjective quality assessment methodologies have been standardised and widely used to evaluate the quality of videos on computer screens or TVs. These include the Double-Stimulus Continuous Quality-Scale (DSCQS) method, the Double-Stimulus Impairment Scale (DSIS) method (also known as Degradation Category Rating (DCR) [7]), the Absolute Category Rating methods (ACR), and the ACR with Hidden Reference (ACR-HR) [6], [10]. However, the equivalent standards for 360° videos are non-existent, and only some works in the literature have proposed guidelines based on experience [26]. Mainly, the existing works in the state-of-the-art have been using those methods (e.g., ACR and ACR-HR), originally developed for 2D content [27], to subjectively evaluate omnidirectional video and image quality [28]–[32]. However, there have been few proposals of methodologies modified for 360° video. For example, Fremerey *et al.* proposed the Modified Pair-Comparison (M-PC) method to evaluate slight perceptual differences [33], [34], while Singla *et al.* proposed the Modified-ACR (M-ACR) method [35], and compared it with the DSIS and ACR methodologies, finding that DSIS is statistically more reliable [36].

As for the subjective test methodologies, the objective metrics for the instrumental evaluation of 360° video quality already reported in the state-of-the-art are based on the proposals originally developed for 2D content [37]. Most of the solutions are adaptations of either Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Multi Scale-SSIM (MS-SSIM), such as Spherical PSNR (S-PSNR) [38], Weighted to Spherically PSNR (WS-PSNR) [39], Area Weighted spherical PSNR (AW-PSNR) [40], Craster Parabolic Projection PSNR (CPP-PSNR) [41], Omnidirectional Video PSNR (OV-PSNR) [42], Weighted-SSIM (W-SSIM), and Weighted MS-SSIM (WMS-SSIM) [28]. Recently, Video Multimethod Assessment Fusion (VMAF)[3] has been considered one of the most robust metrics for traditional contents, and its application on 360° videos has been validated and compared to other metric adaptations [43], and new specific models for 360° content are being developed based on machine-learning approaches [44]. Despite the usefulness of these metrics in providing quality estimations, further research is required, given that QoE factors are not always considered within their frameworks, and therefore do not necessarily correspond to what users perceive.

C. Assessment of Audiovisual Quality With 360° videos

Several studies have been already carried out addressing factors influencing the audiovisual quality of 360° videos. For instance, a great effort in the literature focuses on designing coding and transmission schemes that improve QoE with bandwidth limitations, proposing to provide higher quality to the Field of View (FoV) that corresponds to the area in the HMD where the user is looking at. Thus, the content is partitioned spatially in tiles and temporally in segments, providing multi-quality scenes [45], [46]. In this sense, Muñoz *et al.* [47] proposed a methodology to monitor the quality perceived by users based on the FoV. Other studies focused on the impact of frame rate on 360° video quality [33], [34], [48], concluding that a higher frame rate could have a positive impact on the overall perceived 360° video quality. However, further research is required on how this type of coding and transmission schemes affects observers [5].

Another essential aspect when visualising 360° content is the influence of the HMD on the audiovisual quality. For instance, Zhang *et al.* [49] investigated the resolution of the source 360° video needed to guarantee a per-pixel presentation according to the FoV and resolution of the HMD screen. Singla *et*

[3][Online]. Available: https://github.com/Netflix/vmaf

al. [30] found that HTC Vive provides slightly better audiovisual quality than Oculus Rift. Orduna *et al.* [31] also compared two HMDs (Samsung Gear VR and Lenovo Mirage Solo), without finding significant differences between the perceived quality. However, they obtained significant differences between the usability of the evaluation mechanism, touchpad and handheld controller, respectively, showing that to use the controller is more natural for the participants. In this sense, another aspect being analysed is the way of collecting the responses from the participants during the subjective test. Participants can rate the sequences using a controller or touchpad of the HMD [50], avoiding removing the goggles. Additionally, it can be recorded on a paper [30], verbally [35], [36], [51], [52], or online using a web application [53].

Moreover, the users' freedom to explore 360° content adds another variable to the influence of a crucial factor in subjective video quality assessment: the duration of the test sequences. On one side, the designers of subjective QoE tests have to balance time constraints and limits of human attention and fatigue to measure the phenomenon under investigation. Typically, video quality subjective tests use short sequences (e.g., 8-10 seconds) [7], [54], especially when the performance of some systems or algorithms (e.g., encoders) are under evaluation. On the other side, using short sequences is far from real video consumption scenarios, since most people do not watch television or 360° video in 10-second intervals [55]. Thus, real-life scenarios and QoE definition [12], [56] provide arguments supporting the use of longer sequences that can attract the interest of the participants in the tests and immerse them in the narrative [57]–[60]. Other research questions motivate the use of stimuli longer in duration than is typical in standard tests allowing to study a more realistic experience and deeper user engagement [61] or higher-order cognitive processes [62]. Additionally, users of 360° videos may spend some time at the beginning of the sequence to explore the whole scene before focusing on the assessments [63], [64].

Furthermore, although in many cases, the test sequences in subjective quality experiments are watched by the participants without audio, subjective studies have shown that it can influence the perceived quality, and that the degradations affecting (jointly or separately) the audio and video signals impact the user QoE [60], [65], [66]. Thus, objective metrics have been developed trying to model the overall audiovisual quality [65]–[67]. This influence can be emphasized when visualising 360° videos, given that audio can impact the exploration and attention of the observers [68]–[71], thus, possibly influencing the quality assessments. Also, in immersive environments, to wear headphones enhances the involvement of the user [72] and can reduce simulator sickness [73].

D. Assessment of Simulator Sickness With 360° videos

When watching VR stimuli with HMDs, the users may suffer from cybersickness or simulator sickness [74], which can be worse with moving or dynamic stimuli [75]. In order to evaluate the related symptoms, the most popular method is the Simulator Sickness Questionnaire (SSQ) developed by Kennedy *et al.* [76]. It consists of 16 symptoms grouped by three factors (oculomotor, nausea, and disorientation). It has been successfully applied to video quality tests. For instance, it was used to study the symptoms caused by 3D content, showing a significant increase in symptoms after viewing long stereoscopic 3D videos [77], and after watching 3D content on a 3DTV and with immersive 3D glasses (a kind of HMD), although in a different way for the two 3D viewing technologies [78]. The SSQ has been also used lately to study the possible symptoms caused by viewing 360° videos. For example, Singla *et al.* [30] investigated the impact of resolution, HMDs, and content on simulator sickness symptoms, and found that resolution and contents have a statistically significant impact on the scores of the SSQ.

Given that the SSQ was originally developed to deal with military flight simulators, more appropriate alternatives have been investigated to be used with immersive technologies, such as VR. For example, Kim *et al.* [79] proposed the Virtual Reality Sickness Questionnaire (VRSQ). It was derived from the SSQ by reducing the number of symptoms from 16 to 9, grouped by two factors (oculomotor and disorientation). Also, Cybersickness Questionnaire (CSQ) was proposed by Stone III *et al.* [80], derived from the SSQ by retaining only nine symptoms, also grouped by two factors (dizziness and difficulty focusing). In addition, there are quite a few single-scale questions such as Fast Motion Sickness Scale (FMS) [81], Misery Scale Index (MISC) [82], Vertigo scale [83], and Short SSQ [84] that can be used to assess sickness. This Short SSQ was compared to the long SSQ [76] in a subjective quality test with 360° videos, carried out by Singla *et al.* [85]. Their results showed that to investigate the impact of individual technical factors (e.g., bitrate, resolution, etc.) the SSQ should be used, while the Short SSQ can replace it to differentiate videos causing low or high simulator sickness.

E. Datasets of 360° videos

In recent years, several studies have been published proposing different datasets for 360° content [86]. For instance, the Joint Video Exploration Team (JVET) created datasets with sequences (in uncompressed format) of 10 seconds for research on 360° video encoding [87], [88]. Also, several databases can be found in the literature with associated ratings of visual quality [31], [32], [89], [90], presence [31], [64], valence and arousal [91], and simulator sickness [31], [64], [90], [92]. In addition, many of these datasets also include data related to head movements [90], [92]–[97] and eye movements [63] of the participants.

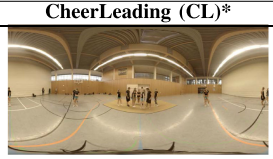
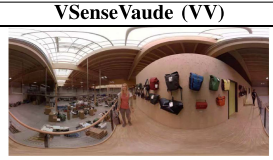
Furthermore, 360° sequences can also be downloaded from online platforms such as YouTube, Arte, and from project repositories, such as ImmersiaTV[4]. However, they are not available in uncompressed or very high-quality versions, and they may have copyright restrictions. Taking this into account, and considering that for our tests we required free, high-quality videos, with at least 30 seconds of duration, and covering a wide range of spatial, temporal and exploration properties, a new dataset was

[4][Online]. Available: http://www.immersiatv.eu/project-outcomes/datasets/

TABLE I
DISTRIBUTION OF THE NINE TEST CONDITIONS AND PARTICIPANT LABORATORIES

ID	Test Condition	Methodology	Lab	HMDs	Num. of PVSs	PVSs' Length
A	Video duration	ACR	Wuhan	HTC Vive	64	10s & 20s
B	Video duration	ACR	AGH	Oculus Rift	40	20s & 30s
C	Video duration	DCR	Roma3	HTC Vive	40	10s & 20s
D	Video duration	DCR	CWI	Oculus Rift	30	20s & 30s
E	Video duration	ACR	Surrey	HTC Vive	48	10s & 30s
F	Influence of HMD (desktop/mobile, High/low resolution)	ACR	UPM & Nokia	GearVR vs. HTC Vive vs. HTC Vive Pro	48	20s
G	Influence of HMD (Tethered vs. untethered)	ACR	Ghent	HTC Vive Pro	48	20s
H	Influence of audio (Videos with vs. without audio)	ACR	RISE	HTC Vive	48	20s
I	Influence of scoring method (App. vs. voice)	ACR	TU Ilmenau	HTC Vive Pro	48	20s

TABLE II
PROPERTIES OF THE SOURCE SEQUENCES

Name (ID)	**NokiaDojo (ND)***	**NokiaFlamenco (NF)**	**CheerLeading (CL)***	**BrazilMusic (BM)**
Screenshot				
Resolution	3840x2160, 30fps	3840x2160, 30fps	4096x2048, 25fps	4096x2048, 25fps
Provider	Nokia	Nokia	TU Ilmenau	TU Ilmenau
Description	Video of an indoor sport course, with ambient audio. Contains stitching artifacts.	Indoor dance course, with ambient audio. Contains stitching artifacts.	Cheerleading session indoors, with ambient audio.	Indoor scene of a band playing Brazilian music. With audio.
Name (ID)	**VSenseLuther (VL)**	**VSenseVaude (VV)**	**OculusMotion (OM)**	**OculusBeach (OB)***
Screenshot				
Resolution	4096x2048, 30fps	4096x2048, 30fps	3840x1920, 30fps	3840x1920, 30fps
Provider	VSense	VSense	Oculus	Oculus
Description	Video with animation content and a main character. Contains various shots (indoors and outdoors) and audio.	Video where a girl speaks to the camera. Contains audio and various indoor and outdoor shots.	Camera moving in a city. Contains music and two shots: one in daylight and one at night.	Scene with music of a beach at sunset with people dancing and moving.

collected, which is made publicly available with quality and simulator sickness annotations and head-rotation data collected form the cross-lab experiment.

III. SUBJECTIVE EXPERIMENT

A. Test Conditions

According to the objectives reported in Section I, the nine test conditions shown in Table I were established to be evaluated in the cross-lab tests, including: two test methodologies (ACR and DCR), test videos of 10, 20 and 30 seconds, and different HMDs (desktop, mobile, tethered, untethered, etc.), methods to collect observers' ratings, and using sequences with and without audio. The selected test conditions cover factors influencing the assessment of audiovisual quality, including the impact of spatial degradations (e.g., coding artifacts), which is commonly done with short sequences [7]. Several other factors influence the overall QoE of the users when watching 360° videos [16], such as immersion [64] or temporal degradations (e.g, transmission degradations [45], latency [98], etc.), which may require longer sequences to be properly evaluated [58]–[60], and were out of the scope of the test campaign presented in this paper. In addition, given that even with short sequences the users may experience simulator sickness, different questionnaires were considered to analyze how and when to assess it during the test session. The following subsections provide details on these test conditions and the experimental setups used in the tests.

B. Test Stimuli

Eight 360° videos of 30 seconds were used as source sequences (SRCs) in the tests. They were all in equirectangular projection, monoscopic, and had at least a resolution of 3840x1920 pixels and 25 fps. Screenshots of these sequences and their main characteristics are shown in Table II. The original videos were provided by Nokia, TU Ilmenau, VSense [93], and

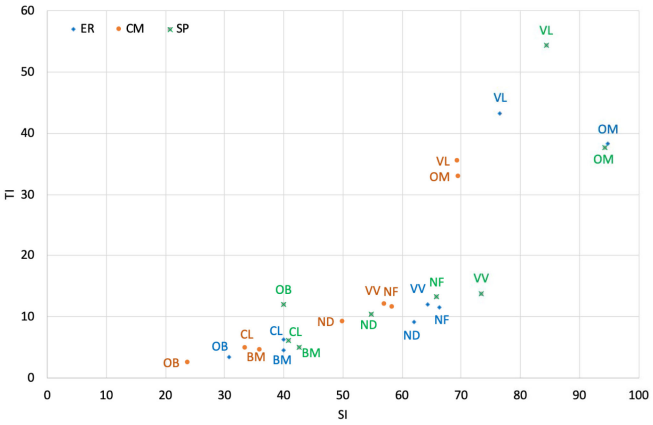


Fig. 1. Scatter plot of SI and TI of the source sequences in ER, CM and SP projections.

#Tiles	Transition	ROI	QPs							
8x5**	Smooth	90°	42	37	32	22	22	32	37	42
6x3	Smooth	120°	42		32	22	22		32	42
8x5**	Abrupt	180°	42	42	22	22	22	22	42	42
6x3	Abrupt	120°	37		37	22	22		37	37

Fig. 2. Settings for the non-uniform coding configurations.

Oculus. The selected sequences present a wide range of content characteristics, including one video with camera motion (OM), one with animation scenes (VL), and different spatial and temporal complexities, as shown by the Spatial Information (SI) and Temporal Information (TI) indices [10] represented in Fig. 1. As it can be seen, SI and TI have been computed in three different projections, i.e., equirectangular (ER), cube-map (CM) and spherical (SP), to account for possible inaccuracies due to projection distortions [9], [99], [100]. Although small differences can be observed, the three domains' computations are highly correlated and show a wide distribution of spatial and temporal properties of the dataset.

Eight different HEVC coding configurations were applied to generate the test videos, including four uniform encodings (using homogeneous QPs) and four non-uniform encodings (using different configurations of tiles). For the uniform configurations, the following QPs were used: 15, 22, 32, 42, while Fig. 2 shows the settings for the non-uniform ones. As it can be seen, two different structures of tiles were used, and smooth and abrupt transitions between adjacent tiles were considered. The encoding of all the test sequences was done using the Kvazaar encoder, applying $period = 2s$, $gop = 0$ (structure disabled), and $ref = 1$ (forcing reference frames). Also, for each encoded video, three sequences were created with different duration using the first 10 seconds, the first 20 seconds, and the whole 30 seconds video[5].

[5]This dataset is publicly available. Access details at: [Online]. Available: https://www.its.bldrdoc.gov/vqeg/video-datasets-and-organizations.aspx

C. Evaluation Methodologies

The participants in the tests were asked to freely watch and explore the test contents and rate them in terms of audiovisual quality and simulator sickness according to the following methodologies.

1) Audiovisual Quality: In order to validate test methodologies for subjective quality assessment of 360° videos, two methodologies were implemented in different laboratories [7]:

- ACR: Single-stimulus method where the test videos are presented to the observers in random order, and they rate the stimuli independently on a five-grade category scale, from 5 (excellent) to 1 (bad).
- DCR (or DSIS): Double-stimulus method where, for each test video, the observers first watch the corresponding reference video, and they rate the degradations on a five-point scale, from 5 (imperceptible) to 1 (very annoying).

The impact of two different ways to collect observers' ratings was investigated. On one side, the Unity-based application Miro360[6] [101] was used, which allows the presentation and rating of the videos in the HMDs and the recording of the ratings and head-rotation data. On the other side, one lab also collected the ratings that the observers provided verbally [36]. In this case, the participant had to say the number of the rating aloud, and the experimenter noted it down. In both cases, the rating scales were displayed in the HMD after each test video, and the observers were able to evaluate all the test videos without removing the HMD to rate.

2) Simulator Sickness: In order to study appropriate methods to evaluate simulator sickness with 360° video, three different questionnaires were used in the cross-lab tests:

- Simulator Sickness Questionnaire (SSQ) [76]: The widely-used method by Kennedy, which evaluates 16 symptoms grouped in 3 factors: oculomotor, nausea, and disorientation. Each symptom is evaluated using a four-grade scale (0 = none, 1 = slight, 2 = moderate, and 3 = severe). In addition to global scores for each factor, a total score can be computed.
- Vertigo Scale [83]: The single-question method proposed by Pérez *et al.*, which evaluates simulator sickness stating the question "Are you feeling any sickness or discomfort now?" and using a five-grade scale (from "no problem" to "unbearable").
- Short-SSQ [84]: Another single-question method proposed by Tran *et al.*, which evaluates simulator sickness in terms of dizziness using the question "How is your level of dizziness or nausea?" and a five-grade scale (from "absolutely not dizzy" to "very dizzy").

These questionnaires were filled by the participants (not wearing the HMDs) in various moments during the test session (see details in Subsection III-E), so it was possible to analyze the evolution of the symptoms. In all those moments, each participant filled the full SSQ and one of the single-item questionnaires (always the same), which were randomly assigned to obtain balanced samples.

[6][Online]. Available: https://git.gti.ssr.upm.es/pub/Miro360

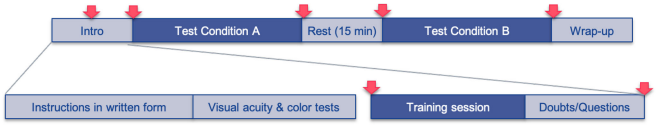


Fig. 3. Diagram of the structure of the test session.

D. Environment and Equipment

The tests were carried out by ten laboratories at Wuhan University (China), AGH University of Science and Technology (Poland), Roma TRE University (Italy), Centrum Wiskunde & Informatica (The Netherlands), Nokia Bell-Labs (Spain), Universidad Politécnica de Madrid (Spain), Ghent University (Belgium), RISE Research Institutes of Sweden (Sweden), TU Ilmenau (Germany), and University of Surrey (United Kingdom). The tests were conducted in controlled environments in all laboratories, where the observers were seated in a swivel chair, so they could rotate freely to explore the 360° videos.

To study the influence of the HMD, four different devices were used in the cross-lab tests: Samsung GearVR, a mobile solution based on attaching a smartphone to an HMD support with a resolution of 1280x1440 pixels per eye and a refresh rate of 60 Hz; Oculus Rift and HTC Vive, consumer desktop solutions with resolutions of 1080x1200 pixels per eye and 80 Hz and 90 Hz, respectively; and HTC Vive Pro, high-resolution (1440x1600 pixels per eye and 90 Hz) solution available both tethered and untethered.

E. Session Structure

As can been seen in Table I, two test conditions were evaluated in each lab. The evaluation of the two test conditions was done by the same participants, following the session structure depicted in Fig. 3, which was followed by all laboratories. Firstly, an introductory session was performed with the participants, where instructions for the test were provided, visual screening was performed, and training video samples were shown to appropriately adjust the HMD and familiarize them with the test methodology. Also, consent forms and background questionnaires were filled. At the end of this session, any doubts or questions from the participants were clarified. Then, the participants evaluated the test stimuli corresponding to the first test condition and, after a break of 15 minutes, they evaluated the corresponding ones for the second test condition. At the end of the test, the participants were requested to answer some more general questions about it.

As aforementioned, the participants were asked to fill questionnaires to evaluate simulator sickness, which was done various times during the test session. As depicted by the red arrows in Fig. 3, these questionnaires were filled before the training session (1), after the training session and just before starting the evaluation of the first test condition (2), after the evaluation of the first test condition (3), after the training and just before the evaluation of the second condition (4), and at the end of the test (5).

The test sessions lasted less than 90 minutes, and the evaluation of each test condition did not last more than 25 minutes, approximately. In those cases in which DCR methodology was used, and longer test sequences were evaluated, a subset of the test stimuli was considered to satisfy those time limits. In particular, the source contents NokiaDojo, CheerLeading and OculusBeach (marked with * in Table II) were not considered to generate the test stimuli used in test conditions B (AGH), C (Roma3) and D (CWI), and in this last case, the non-uniform coding configurations using 8x5 tiling patterns were also not used (marked with ** in Fig. 2).

F. Observers

A total of 306 participants took part in the cross-lab test (38.9% women, 61.1% men), with ages ranging between 18 and 79 (average of 28.8). Vision screening was carried out before the tests, to assure that observers had a standard or corrected-to-normal vision in terms of visual acuity and colour vision. The participants were also asked to fill a background questionnaire in which they had to indicate their experience using VR/AR headsets. All details by lab and in total are reported in Table III. A total of 60 participants performed the tests in UPM & Nokia (Test F), who were organized so that each observer evaluated two HMDs, thus, each HMD was evaluated by 40 participants.

IV. Results

A. Audiovisual Quality

Fig. 4 and Fig. 5 show some of the results obtained for audiovisual quality in terms of Mean Opinion Scores (MOSs) and 95% confidence intervals. On the one side, Fig. 4 shows the results from one lab that evaluated all test sequences: Test A, performed at Wuhan, to study the impact of sequence duration (10 s vs. 20 s) using ACR. On the other side, Fig. 5 shows the results from all laboratories for a characteristic SRC video (VSenseLuther), as an example to show the behavior of the obtained results in all laboratories. These two figures provide illustrative examples of the results obtained in all laboratories for all Processed Video Sequences (PVSs) and test conditions, which can be found in the supplemental material.

The following subsections present the statistical analysis of the results obtained for the evaluated test conditions related with the main contributions provided to ITU-T Rec. P.919 [9]. In particular, when comparing conditions evaluated within the same laboratory, we used mixed-model analysis [102] and post-hoc tests (applying Bonferroni corrections for multiple comparisons when required). The reference significance level considered in all analyses is $\alpha = 0.05$.

In essence, mixed-effect modelling allows to study the datasets with both fixed-effect factors (described below) and random-effect factors (subject and PVS) and the conclusions can be generalized to the populations sampled by the random-effect factors [103]. The reason for using mixed models analysis is to compare means (despite the scale used in the test being discrete), which would technically not allow using classical linear regression. Classical regression analysis assumes the normality, homoscedasticity, and serial independence of regression residuals [104]. However, this analysis does not focus on residuals and

TABLE III
NUMBER, AGE DISTRIBUTION, AND EXPERIENCE WITH VR/AR HEADSETS OF THE OBSERVERS. ONE PARTICIPANT FROM ROMA3 DID NOT REPORT HIS/HER EXPERIENCE

Lab	Test ID	Number of Observers			Age			Experience with VR Headsets				
		Total	Female	Male	Min	Max	Avg	Times=1	Times<5	5<Times<20	Times> 20	Every day
Wuhan	A	30	15	15	20	30	24.5	8	15	7	0	0
AGH	B	40	13	27	18	79	28.5	13	17	8	2	0
Roma3	C	30	8	22	21	57	30.6	7	10	2	8	2
CWI	D	28	14	14	21	60	27.6	2	12	5	6	3
Surrey	E	31	10	21	19	44	25.9	13	12	3	2	1
UPM & Nokia	F	60	25	35	20	31	23.2	18	32	9	1	0
Ghent	G	30	4	26	23	45	31.6	3	14	7	5	1
RISE	H	28	16	12	22	66	41.6	3	16	8	1	0
TU Ilmenau	I	29	14	15	20	37	25.9	4	18	4	3	0
Total		306	119 38.9%	187 61.1%	18	79	28.8	71 23.20%	146 47.71%	53 17.32%	28 9.15%	7 2.29%

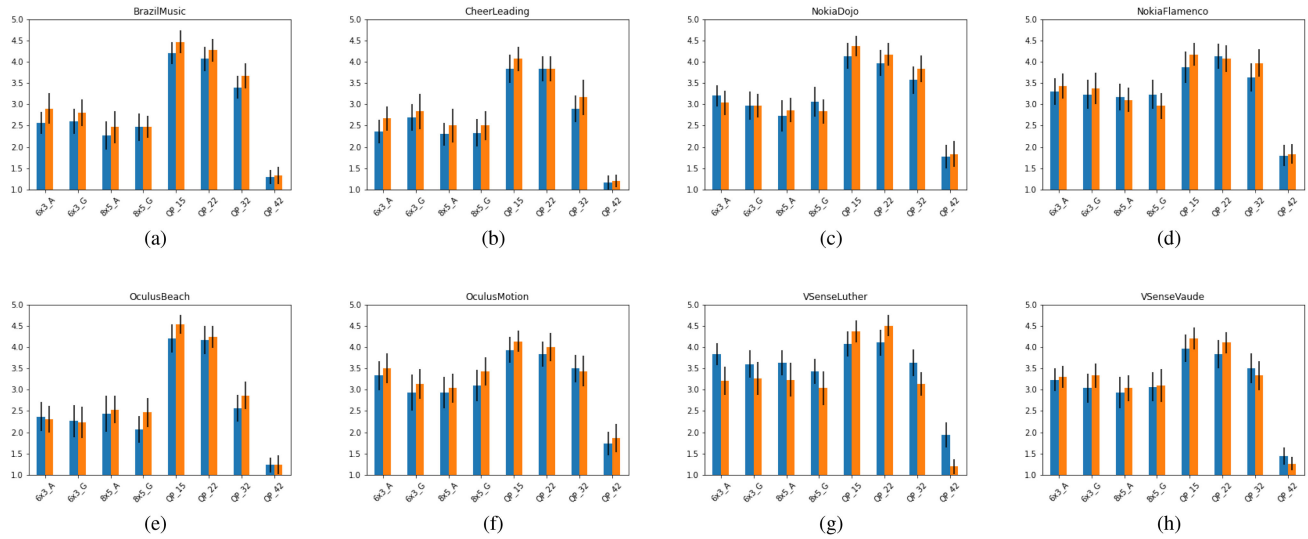


Fig. 4. Results of MOSs from Test A (Wuhan) using ACR with videos of 10 s (blue) and 20 s (orange). Uniform encoding schemes are indicated with the QP, non-uniform ones are named by the tiling division and transition (A: Abrupt, G: Gradual).

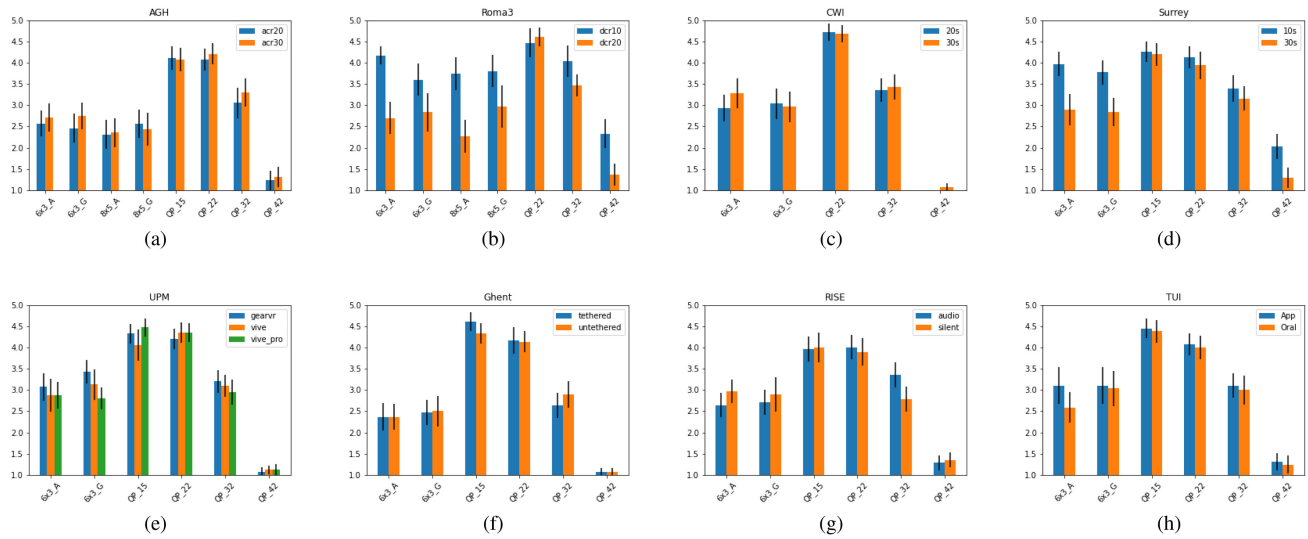


Fig. 5. Results of MOS from all laboratories (considering the tested conditions) for VSenseLuther. Charts for the rest of SRCs can be found in the supplemental material.

residual normality, but a normal distribution can approximate the differences in MOS and means since the central limit theorem can be applied. However, the central limit theorem could fail for specific voting behaviors or at the end of the scale. Nevertheless, those are the corner cases, and we are interested in the general differences among the test conditions (see the considered test conditions in Table I).

1) Influence of Methodology: In principle, the two methodologies employed in the test, namely ACR and DCR, were not directly compared by any of the laboratories involved. Nonetheless, as the same conditions were employed in different labs to test the influence of the sequence length for ACR and DCR, it is possible to perform an inter-lab analysis to understand the influence of the selected methodology on the final scores. In particular, we compare the results obtained in Test A (ACR: 10 s vs 20 s) and Test C (DCR: 10 s vs 20 s), as well as the ones obtained in Test B (ACR: 20 s vs 30 s) and Test D (DCR: 20 s vs 30 s). In our analysis, we exclude any sequence that was not present in both the test sessions under exam, to ensure a fair comparison. Results of the Mann-Whitney's U test show a significant effect of test methodology for Test A with respect to Test C ($z = -6.6370, p < 0.001, r = 0.1024$), as well as for Test C with respect to Test D ($z = -3.2416, p = 0.0012, r = 0.0560$), albeit with a smaller effect size. To further understand whether the sequence length might affect the differences among methodologies, we compare the two methodologies separately per sequence length. To do so, we aggregate the results obtained in Test A, B, C, D, and E, while considering only the lowest common group of contents and distortions. Mann-Whitney's U test shows a significant effect of methodology for sequence length of 10 s (Test A, Test C, and Test E: $z = -8.1081, p < 0.001, r = 0.1700$) and 20 s (Test A, B, C, and D: $z = -4.9043, p < 0.001, r = 0.0870$), whereas no significant effect of methodology was observed for sequence length of 30 s (Test B, Test D, and Test E: $z = -1.6306, p = 0.1030, r = 0.0329$). Results indicate that the choice of methodology might have an impact on the distribution of the scores, especially for certain sequence lengths, as MOS values are on average 0.24 higher when using the DCR methodology as opposed to the ACR methodology (for Tests A, B, C, D, and E, and all sequence lengths: $z = -8.5471, p < 0.001, r = 0.0962$). However, the effect sizes we obtain in our comparisons imply that the effect, if existing, is quite small. In addition, the patterns of the results obtained in the involved labs (i.e., expected decreasing quality when increasing uniform QPs and no big differences among the non-uniform configurations considered in the tests, as shown in Fig. 4 and Fig. 5, and in the supplemental material), validate the use of ACR and DCR methodologies for subjective assessment of coding quality for 360° video. Thus, these two methodologies were included in the ITU-T Rec. P.919 [9].

2) Influence of Sequence Duration: Regarding the influence of sequence duration, we present in Table IV the p-values obtained for the different tests considering these conditions. As we can see, all compared conditions, except ACR 20 s vs. 30 s, are statistically significant but with different significance level. Since the obtained results are aggregated over different conditions and SRCs, the results' visual investigation is necessary.

TABLE IV
p-VALUES FOR A MIXED MODEL AND DIFFERENT TEST CONDITIONS. FOR CONDITIONS INVOLVING SEQUENCE DURATION ALSO p-VALUE WITHOUT VSENSELUTHER SEQUENCE IS PRESENTED

ID	Lab	Test Condition	p-value with VSenseLuther	p-value without VSenseLuther
A	Wuhan	ACR, 10s vs. 20s	0.005	6.4e-06
B	AGH	ACR, 20s vs. 30s	0.326	0.754
C	Roma3	DCR, 10s vs. 20s	9.4e-09	0.089
D	CWI	DCR, 20s vs. 30s	0.014	0.001
E	Surrey	ACR: 10s vs. 30s	9.03e-06	0.035
F	UPM & Nokia	GearVR vs. Vive GearVR vs. Vive Pro Vive vs. Vive Pro	0.1087 0.2230 0.0014	N/A
G	Ghent	Tethered vs. untethered HMD	0.562	N/A
H	RISE	With vs. w/o audio	0.006	N/A
I	TUI	Scoring app vs. voice	0.046	N/A

Further, inspection shows that one of the sequences (VSenseLuther) showed unexpected results compared to the other videos. For Wuhan (Test A), the 20-second sequences have, most often, higher MOS (see Fig. 4), except for the sequence VSenseLuther. For Roma3 (Test C), again for VSenseLuther, we obtain a significant decrease in the quality, as observed in Fig. 5(c) (see also all the results from this lab in the supplemental material). That might have been caused by the new scene in this particular sequence, that is not displayed for the first 10 seconds. Thus, in addition we present results obtained without the VSenseLuther sequence (see Table IV). The new results showed the higher statistical significance of Wuhan (Test A) and CWI (Test D), while for Roma3 (Test C) the results stopped being significant. For Surrey (Test E) the significance was reduced, and for AGH (Test B) the results were still not statistically significant. For Wuhan, ACR 20 s comes with higher scores. It is not an effect distinctly visible for a single scene. However, the mixed model's analysis allows us to see all the sequences together, also normalizing each sequence quality's influence. Since in Wuhan (Test A) general differences between MOS for 10 s and 20 s can be observed (see Fig. 4), the overall result shows the statistical significance, and it was shadowed by VSenseLuther reverse influence. After removing this sequence, we conclude that 20-second sequences obtained higher MOS by 0.12 than 10 s ($\chi^2(1) = 20.3, p = 6.4e-06$). Also for CWI (Test D), the effect without the VSenseLuther sequence is more substantial, and again more extended sequences obtain higher MOS by 0.14 ($\chi^2(1) = 10.2, p = 0.001$). The effect observed for Roma3 (Test C) is mainly, or even only, caused by the extreme difference obtained for the VSenseLuther sequence. Thus, after removing it, the effect is not observed anymore ($\chi^2(1) = 2.90, p = 0.089$). Again removing this sequence is necessary since it is not consistent for the first and last 10 seconds. It should be noted, that apart from Roma3 (Test C), AGH (Test B) also did not gather significantly different results, which, in this case, this could be caused by the subjects inconsistency. Since there are two contradicting subject removal algorithms described in ITU-R BT.500 [6] and ITU-T P.913 [7], we decided to not use any of them and leave for the further research this particular condition.

To go one step further and analyze for which test stimuli there were significant differences, Wilcoxon Signed-Rank tests (non-parametric tests for related samples) were computed, after checking the non-normality of the gathered scores, and applying Bonferroni corrections for multiple comparisons. Only significantly different pairs were identified with VSenseLuther: one pair (QP42, $p = 0.0002$) among 64 for Test A (Wuhan), 3 pairs (6x3-abrupt with $p = 8.6e-06$, 8x5-abrupt with $p = 8.4e-05$, QP42 with $p = 0.0003$) among 35 for Test C (Roma 3), and 2 pairs (6x3 gradual with $p = 0.0007$ and abrupt with $p = 0.0002$) among 48 for Test E (Surrey). No significantly different pairs were found for Test B (AGH) among 40 pairs and Test D (CWI) among 25 pairs.

These results evidence that no systematic effects of the sequence duration on the quality ratings are generally observed, while, as expected, differences can be obtained when using characteristic videos with changing properties during time (e.g., VSenseLuther). Thus, subjective tests of coding degradations with 360° videos can be done with sequences of 10 seconds, taking into account these effects, as reported in the ITU-T Rec. P.919 [9].

3) Influence of HMD: On the one side, three different HMDs were compared in UPM & Nokia (Test F). The mixed-model analyses showed no significant differences comparing the mobile Samsung GearVR HMD with the desktop HMDs ($\chi^2(1) = 1.48$ and $p = 0.2230$ for Vive Pro, $\chi^2(1) = 2.57$ and $p = 0.1087$ for Vive), although, surprisingly, slightly higher MOSs were obtained with GearVR. However, significant differences were found comparing HTC Vive and HTC Vive Pro ($\chi^2(1) = 10.16, p = 0.0014$), with better MOSs for the HTC Vive, which provides a lower resolution than HTC Vive Pro. However, the Wilcoxon Signed-Rank tests did not show any significantly different pair among all the possible comparisons (144) among the HMDs for all test videos.

On the other side, the comparison between HTC Vive Pro with and without cables (Test G performed in Ghent) did not show any significant differences, neither from the mixed-model analysis nor from the post-hoc tests.

These results evidence that any commercial HMD (tethered or untethered) can be used in visual quality tests with 360° videos, provided that it has enough resolution and refresh rate to represent the content that is going to be tested, as included in the recommendation ITU-T P.919 [9].

4) Influence of Audio: To check the influence on quality assessment of watching the 360° videos with or without audio, the results from the Test H, carried out by RISE, were analyzed. The mixed model analysis shows that silent sequences obtained MOSs higher by 0.075 ($\chi^2(1) = 7.51, p = 0.006$). The measured difference is statistically significant but minimal, and visible only by analyzing all sequences. Analyzing the differences between all the pairs with Wilcoxon Signed-Rank tests (with Bonferroni corrections), no significant different pairs are detected among the 48 possible comparisons.

These results support that it is possible to use test stimuli either with or without audio to evaluate visual quality, as included in the ITU-T Rec. P.919 [9]. Nevertheless, it should be noted that no spatial audio was used in these tests, so it should be considered that, especially when dealing with non-uniform degradations, off-screen sound may influence audiovisual quality ratings.

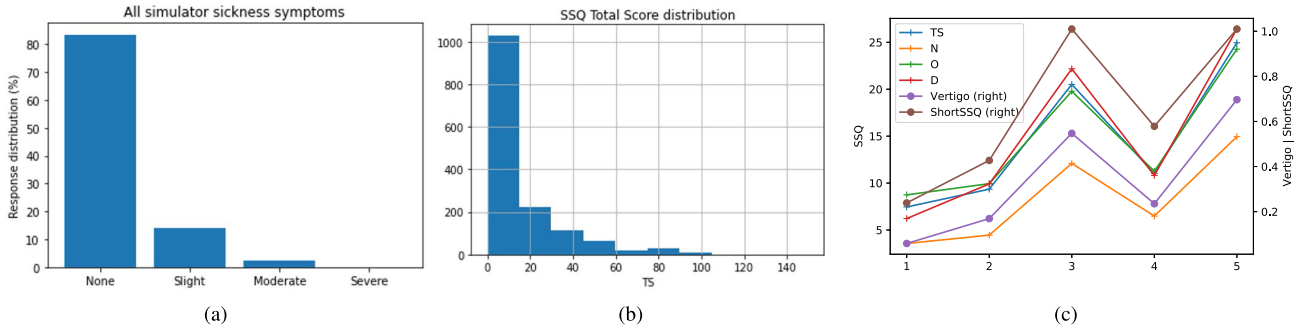
5) Influence of Method to Collect Ratings: To check the influence of the two tested methods to collect the observers' ratings (i.e., through the application and verbally), the results from the Test I, carried out by TU Ilmenau, were analyzed. The mixed model shows the border case with ($\chi^2(1) = 3.975396, p = 0.046$), which is theoretically statistically significant, but indicating a very similar performance of both methods. In fact, the post-hoc Wilcoxon Signed-Rank tests showed no significantly different pairs among the 48 test videos compared. Therefore, both voting interfaces or verbal voting are recommended in the ITU-T P.919 [9] for evaluations performed with 360° videos.

6) Minimum Number of Observers: To compute the minimum number of observers required per laboratory, we base our analysis on the desired statistical power $1 - \beta = 0.8$. Given the within-subject design and the assumed non-normality of the data, we consider the case of a one-tailed Wilcoxon signed-rank statistical test aiming to determine whether one distortion leads to higher MOS scores concerning another. Assuming a type I error probability $\alpha = 0.05$, and an effect size of $r = 0.5$ (in our test, the observed range was $r = [0.46, 0.62]$), we use the free software G*Power [105] to obtain a minimum sample size of $N = 28$. This is in line with an estimation as outlined in Brunnström, and Barkowsky [106], using VQEGNumSubjTool.[7] For this, we considered a within-subject design with the same statistical power of 0.8, a standard deviation of 0.9 (which is a bit higher than we can expect in regular 2D video quality test), and a MOS difference of 1. Considering that the number of PVSs in each sub-experiment is about 50, and that we are looking at all possible comparisons (i.e., $50 \cdot 49/2 = 1225$), the result was also $N = 28$. This calculation is based on the t-test, which is more efficient as it relies on parametric statistics and would give a lower number, but considers multiple comparisons with an overall $\alpha = 0.05$ for each experiment. These results supported the recommendation, included in ITU-T P.919 [9], to have at least 28 participants in similar subjective tests with 360° videos.

B. Simulator Sickness

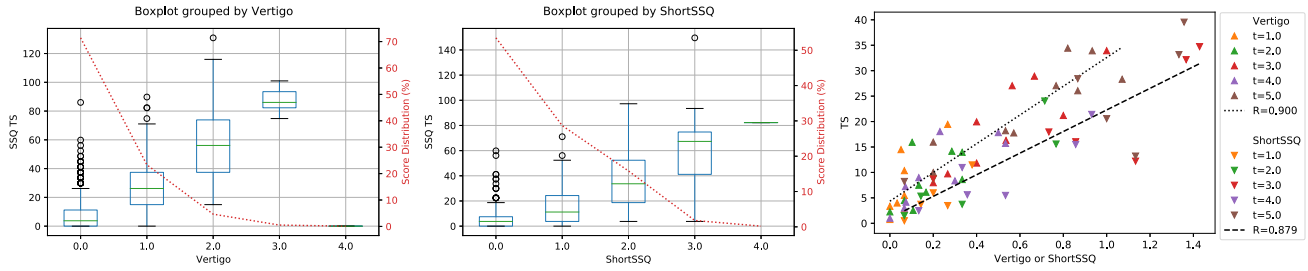
1) Test Methodology: The scores collected from the widely-used SSQ [76] can be considered a ground truth for simulator sickness measurement. Thus, these results are used to analyze whether the implemented test methodologies are appropriate for simulator sickness. The distribution of all the symptoms shown in Fig. 6(a), evidence that the simulator sickness of the participants was low, with only some slight/moderate symptoms. The distribution of the total scores also confirms it (computed form the evaluated symptoms according to [76]) shown in Fig. 6(b), since mainly low scores were obtained. Regarding the evolution of simulator sickness during the test session, the results shown in Fig. 6(c), demonstrate a positive effect of the break and no significant differences between the symptoms before and after the training.

[7] [Online]. Available: https://slhck.shinyapps.io/number-of-subjects/



(a) (b) (c)

Fig. 6. Global results of simulator sickness: (a) Distribution of all symptoms, (b) Distribution of the total score, (c) Results on each measurement point.



(a) (b) (c)

Fig. 7. Simulator sickness results from single-item questionnaires: (a) Boxplot of total scores grouped by the Vertigo scale [83], (b) Boxplot of total scores grouped by the Short-SSQ [84], (c) Total scores vs. Vertigo/Short-SSQ scores (average in each lab for each measurement point) and Pearson correlation coefficient.

2) Long Vs. Short Questionnaires: To analyze the performance of the single-item questionnaires used in the test, their results are compared to those obtained with the long SSQ, serving as ground truth. Fig. 7(a) and (b) show the boxplots of the total scores (obtained from the long SSQ) grouped by the Vertigo scale [83] and by the Short-SSQ [84], respectively. In both cases, the differences among the single-item levels 0 to 3 are statistically significant ($p < 0.05$) after computing Kruskal-Wallis and post-hoc Mann-Whitney (with Bonferroni correction for multiple comparisons) tests. Also, the dotted lines represent the score distribution. As it can be seen, while the Short-SSQ provides a bit wider scores distribution (more scores in bins 1 and 2), the Vertigo scale covers a broader range of SSQ Total Score (bins 0-3 are more separated). Also, Fig. 7(c) shows the correlation coefficient of the average total scores from the long SSQ with the Vertigo and Short-SSQ average scores (per lab and measurement point), 0.90 and 0.88, respectively. These results show that: *(i)* single-item questionnaires provide valid coarse-level information about simulator sickness; *(ii)* to compute the "Mean Sickness Score" for a test session (no individual scores needed), they can safely replace the full SSQ; and *(iii)* these two properties do not depend on the specific single-item questionnaire used.

To test whether all 16 symptoms of SSQ are needed to have a good understanding of simulator sickness for 360° video, three alternative sub-samplings were evaluated: the Virtual Reality Sickness Questionnaire (VRSQ) [79], the CyberSickness Questionnaire (CSQ) [80], and new factor analysis (New-FA) performed on the SSQ results of the cross-lab experiments to be used for benchmarking purposes. To obtain a similar number of items and factors as CSQ and VRSQ, New-FA considered 2-factor decomposition with *oblimin* rotation, keeping the eight symptoms with loadings greater than 0.5. The Pearson correlation coefficients between the SSQ total score and the rest of the total scores are greater than 0.9, as shown in the Table V. The correlation coefficients between VRSQ and SSQ scores for the factors disorientation and oculomotor, and the total score are 0.910, 0.960 and 0.958, respectively. These results evidence that VRSQ can be a good shorter alternative to the SSQ for scenarios addressing 360° video.

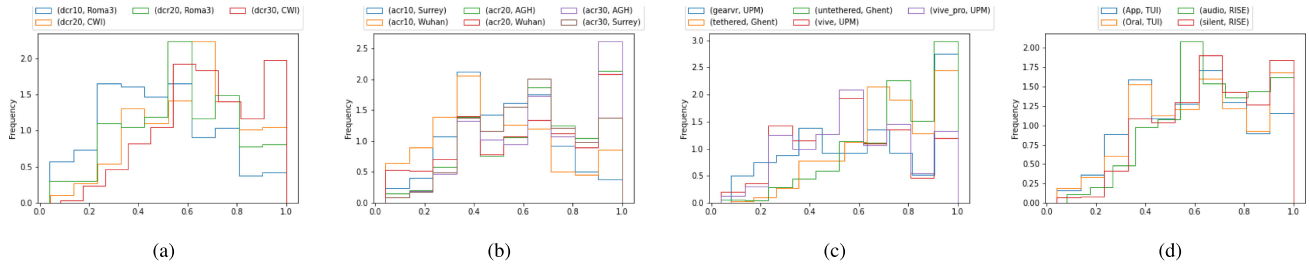
TABLE V
PEARSON CORRELATION BETWEEN SSQ TOTAL SCORE AND THE REST OF TOTAL SCORES

Questionnaire	SSQ	VRSQ	CSQ	New-FA
SSQ	1.000	0.958	0.918	0.951
VRSQ	0.958	1.000	0.870	0.905
CSQ	0.918	0.870	1.000	0.878
New-FA	0.951	0.905	0.878	1.000

Therefore, both Vertigo scale [83] and VRSQ [79] have been included in the recommendation ITU-T P.919 [9] as alternatives to the SSQ [76].

C. Exploration Behavior

The head rotation movements recorded through the HMD sensors while the participants watched the 360° videos allow the analysis of exploration behaviors depending on the different test



(a) (b) (c) (d)

Fig. 8. Results of the participants' exploration (histograms of covered portions of the longitudinal range) of the test sequences.

conditions addressed in the experiment. The coverage results are shown Fig. 8, which provides information on the degree of horizontal exploration of the test contents by the participants. So, the abscissa axis represents the fraction of the sphere longitude that has been visited by them, while the ordinate axis represents how many times (as normalized frequencies) a certain portion of the sphere was visited, accounting for all participants and test videos. Thus, the right end of the abscissa axis (value "1.0") reflects the probability that the entire horizontal range is explored.

Fig. 8(a) shows the coverage related to test conditions involving DCR methodology and different sequence duration. As expected, the participants explored more longer videos, as shown by the higher frequencies achieved for the exploration of the whole longitudinal range with 30-second sequences. On the contrary, with 10-second sequences the participants explored mainly less than half of the range. Generally, similar results can be seen with ACR methodology in Fig. 8(b). Furthermore, the coverage related to conditions comparing different HMDs are depicted in Fig. 8(c), showing that untethered devices (e.g., Samsung GearVR and HTC Vive Pro without cables) allow a wider exploration of the test sequences. Finally, Fig. 8(d) shows the coverage related to test conditions involving sequences with and without audio and the two rating methods (i.e., rating app and verbal voting). On the one side, the participants explored more the silent sequences, which can be due to the fact that in those cases audio is not leading the participants' attention, especially in certain videos with characters speaking (e.g., VSense-Vaude). On the other side, providing the ratings orally may allow a wider exploration of the sequences thanks to not holding the controllers, letting the participants to move more comfortably.

V. Conclusion

This paper presents a cross-lab study on subjective quality assessment of 360° video that was carried out within the IMG of the VQEG involving ten laboratories and more than 300 participants. The obtained results were instrumental on the development of the ITU-T recommendation P.919. These tests allowed to analyze the influence on the visual quality ratings, simulator sickness, and exploration behavior of several factors. In particular, the tests have shown the validity of ACR and DCR methodologies for subjective assessment of short 360° videos, the possibility of using 10-second sequences (with or without audio) for quality assessment tests and any commercial HMD (that satisfies minimum resolution and refresh rate requirements), and the adequacy of both VR voting interfaces and verbal rating to provide the evaluations. Statistical analyses have shown that a minimum number of 28 participants is recommended for this type of tests. Also, methods to assess simulator sickness have been analyzed, recommending the most appropriate ones for tests with 360° videos. Finally, this work has resulted in the generation and publication of a dataset of subjectively assessed 360° content to foster future research. Future work will focus on: 1) obtaining more outcomes from the gathered subjective results with deeper analyses, 2) the study of the performance of objective metrics and the development of new models, and 3) the research on methodologies to assess other influencing factors not covered in these test, which require the use of longer 360° sequences for an appropriate evaluation, such as immersion and presence.

References

[1] Digi-Capital, "Augmented/Virtual reality report Q4 2017," Digi-Capital, Tech. Rep., 2017.
[2] J. Feltham, "Nvidia predicts 50 million VR headsets sold by 2021," *Upload VR*, 2018.
[3] J. G. Apostolopoulos *et al.*, "The road to immersive communication," *Proc. IEEE Proc. IRE*, vol. 100, no. 4, pp. 974–990, Apr. 2012.
[4] A. Perkis *et al.*, "Qualinet white paper on definitions of immersive media experience (IMEx)," 2020, *arXiv:2007.07032*.
[5] M. Xu, C. Li, S. Zhang, and P. Le Callet, "State-of-the-art in 360° video/image processing: Perception, assessment and compression," *IEEE J. Sel. Top. Signal Process.*, vol. 14, no. 1, pp. 5–26, Jan. 2020.
[6] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," Recommendation BT. 500–14, Oct. 2019.
[7] ITU-T, "Methods for the subjective assessment of video quality, audio quality and audiovisual quality of internet video and distribution quality television in any environment," Recommendation P.913, Mar. 2016.
[8] ITU-T, "Subjective assessment methods for 3D video quality," *Recommendation P.915*, Mar. 2016.
[9] ITU-T, "Subjective test methodologies for 360° video on head-mounted displays," Recommendation P. 919, Oct. 2020.
[10] ITU-T, "Subjective video quality assessment methods for multimedia applications," Recommendation P.910, Apr. 2008.
[11] S. Möller and A. Raake, *Quality of Experience: Advanced Concepts, Applications and Methods*. Berlin, Germany: Springer, 2014.
[12] P. Le Callet, S. Moller, and A. Perkis, "Qualinet white paper on definitions of quality of experience (QoE)," in *Proc. Output 5th Qualinet Meeting*, 2013.
[13] ITU-T, "Vocabulary for performance, quality of service and quality of experience," Recommendation P.10/G. 100, Nov. 2017.
[14] F. Hammer, S. Egger-Lampl, and S. Möller, "Quality-of-user-experience: A position paper," *Qual. User Experience*, vol. 3, no. 1, pp. 1–15, Dec. 2018.
[15] U. Reiter *et al.*, "Factors influencing quality of experience," in *Proc. Qual. Experience: Adv. Concepts, Appl. methods, Ser. T-Labs Ser. in Telecom. Services*, S. Möller and A. Raake, Eds. Berlin, Germany: Springer, 2014, ch. 4, pp. 45–60.
[16] ITU-T, "Influencing factors on quality of experience (QoE) for virtual reality (VR) services," Recommendation G.1035, May 2020.

[17] J. Häkkinen, M. Pölönen, J. Takatalo, and G. Nyman, "Simulator sickness in virtual display gaming: A comparison of stereoscopic and non-stereoscopic situations," in *Proc. Conf. Hum.-Comput. Interact. Mobile Dev. Serv.*, 2006, pp. 227–230.

[18] ISO/IEC JTC1/SC29/WG1(JPEG) & WG11(MPEG), "Technical report of the joint ad hoc group for digital representations of light/sound fields for immersive media applications," Geneva, Switzerland, 2016.

[19] IEEE, "IEEE SA - VRAR - virtual reality and augmented reality working group," P 2048, 2017.

[20] IEEE, "IEEE SA - 1918.1.1 - haptic codecs for the tactile internet," 1918.1.1, 2017.

[21] IEEE, "IEEE SA - HFVE - human factors for visual experiences working group," P3333.1, 2017.

[22] A. M. Rohaly *et al.*, "Video quality experts group: Current results and future directions," in *Proc. Vis. Communi. Image Process.*, vol. 4067, May 2000, pp. 742–753.

[23] J. Gutiérrez and K. Brunnström, "Standards column: VQEG," *ACM SIGMultimedia Records*, vol. 12, no. 2, Jun. 2020.

[24] Q. Huynh-Thu, A. Webster, K. Brunnström, and M. Pinson, "VQEG: Shaping standards on video quality," in *Proc. Int. Conf. Adv. Imag.*, 2015, pp. 1–4.

[25] K. Brunnström, D. Hands, F. Speranza, and A. Webster, "VQEG validation and ITU standardisation of objective perceptual video quality mmetrics," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 96–101, Apr. 2009.

[26] A. Singla, S. Fremerey, F. Hofmeyer, W. Robitza, and A. Raake, "Quality assessment protocols for omnidirectional video quality evaluation," *Electron. Imag.*, vol. 2020, no. 11, pp. 69–1-69-7, Jan. 2020.

[27] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 63, no. 11, pp. 1–52, Nov. 2020.

[28] F. Lopes, J. Ascenso, A. Rodrigues, and M. P. Queluz, "Subjective and objective quality assessment of omnidirectional video," in *Proc. Appl. Digit. Image Process. XLI*, vol. 10752, Sep. 2018, Art. no. 107520P.

[29] E. Upenik, M. Reřábek, and T. Ebrahimi, "Testbed for subjective evaluation of omnidirectional visual content," in *Proc. Picture Coding Symp.*, 2016, pp. 1–5.

[30] A. Singla, S. Fremerey, W. Robitza, and A. Raake, "Measuring and comparing QoE and simulator sickness of omnidirectional videos in different head mounted displays," in *Proc. Int. Conf. Qual. Multimedia Experience*, 2017, pp. 1–7.

[31] M. Orduna, P. Pérez, C. Díaz, and N. García, "Evaluating the influence of the HMD, usability, and fatigue in 360VR video quality assessments," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces Abstr. Workshops*, 2020, pp. 682–683.

[32] H. Duan *et al.*, "Perceptual quality assessment of omnidirectional images," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2018, pp. 1–5.

[33] S. Fremerey, F. Hofmeyer, S. Göring, and A. Raake, "Impact of various motion interpolation algorithms on 360° video QoE," in *Proc. Int. Conf. Qual. Multimedia Experience*, 2019, pp. 1–3.

[34] S. Fremerey, F. Hofmeyer, S. Göring, D. Keller, and A. Raake, "Between the frames—Evaluation of various motion interpolation algorithms to improve 360° video quality," in *Proc. IEEE Int. Symp. Multimedia*, 2020, pp. 65–73.

[35] A. Singla, S. Fremerey, W. Robitza, P. Lebreton, and A. Raake, "Comparison of subjective quality evaluation for HEVC encoded omnidirectional videos at different bit-rates for UHD and FHD resolution," in *Proc. Thematic Workshops ACM Multimedia*, 2017, pp. 511–519.

[36] A. Singla, W. Robitza, and A. Raake, "Comparison of subjective quality test methods for omnidirectional video quality evaluation," in *Proc. Int. Workshop Multimedia Signal Process.*, 2019, pp. 1–6.

[37] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[38] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, 2015, pp. 31–36.

[39] S. Yule, A. Lu, and Y. Lu, "WS-PSNR for 360 video objective quality evaluation," in *Proc. JVET-D0040, 4th Meeting*: Chengdu, China, 15–21, Oct. 2016.

[40] B. Vishwanath, Y. He, and Y. Ye, "AHG8: Area weighted spherical PSNR for 360 video quality evaluation," in *Proc. JVET-D0072, 4th Meeting*: Chengdu, China, 15–21, 2016.

[41] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," in *Proc. Optics Photonics Information Processing X*, vol. 9970, Sep. 2016, Art. no. 99700C.

[42] P. Gao, P. Zhang, and A. Smolic, "Quality assessment for omnidirectional video: A spatio-temporal distortion modeling approach," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2020.3044458.

[43] M. Orduna *et al.*, "Video multimethod assessment vusion (VMAF) on 360VR contents," *IEEE Trans. Consum. Electron.*, vol. 66, no. 1, pp. 22–31, Feb. 2020.

[44] W. Sun *et al.*, "MC360IQA: The multi-channel CNN for blind 360-degree image quality assessment," in *Proc. IEEE Int. Symp. Circuits Syst.*, 2019, pp. 1–5.

[45] J. Song *et al.*, "A fast FoV-switching DASH system based on tiling mechanism for practical omnidirectional video services," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2366–2381, Sep. 2020.

[46] J. v. der Hooft, M. Torres Vega, S. Petrangeli, T. Wauters, and F. D. Turck, "Optimizing adaptive tile-based virtual reality video streaming," in *Proc. IFIP/IEEE Symp. Integr. Netw. Service Manage.*, 2019, pp. 381–387.

[47] L. Muñoz *et al.*, "Methodology for fine-grained monitoring of the quality perceived by users on 360VR contents," *Digit. Signal Process.*, vol. 100, May 2020, Art. no. 102707.

[48] F. Hofmeyer, S. Fremerey, T. Cohrs, and A. Raake, "Impacts of internal HMD playback processing on subjective quality perception," in *SPIE Hum. Vis. Electron. Imag.*, vol. 7, Jan. 2019, pp. 219-1–219-7.

[49] Y. Zhang *et al.*, "Subjective panoramic video quality assessment database for coding applications," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 461–473, Feb. 2018.

[50] P. Pérez and J. Escobar, "MIRO360: A tool for subjective assessment of 360 degree video for ITU-T P.360-VR," in *Proc. Int. Conf. Qual. Multimedia Experience*, 2019, pp. 1–3.

[51] A. Singla, W. Robitza, and A. Raake, "Comparison of subjective quality evaluation methods for omnidirectional videos with DSIS and modified ACR," *Electron. Imag.*, vol. 2018, no. 14, pp. 1–6, Jan. 2018.

[52] E. Dima *et al.*, "Joint effects of depth-aiding augmentations and viewing positions on the quality of experience in augmented telepresence," *Qual. User Experience*, vol. 5, no. 1, pp. 1–7, Feb. 2020, Art. no. 2.

[53] A. Singla *et al.*, "Subjective quality evaluation of tile-based streaming for omnidirectional videos," in *Proc. ACM Multimedia Syst. Conf.*, 2019, pp. 232–242.

[54] M. H. Pinson, L. Janowski, and Z. Papir, "Video quality assessment: Subjective testing of entertainment scenes," *IEEE Signal Process. Mag.*, vol. 32, no. 1, pp. 101–114, Jan. 2015.

[55] W. J. Adams, "How people watch television as investigated using focus group techniques," *J. Broadcast. Electron. Media*, vol. 44, no. 1, pp. 78–93, Mar. 2000.

[56] A. Raake and S. Egger, "Quality and quality of experience," in *Proc. Qual. Experience: Adv. Concepts, Appl. methods, Ser. T-Labs Ser. Telecommunication Serv.*, S. Möller and A. Raake, Eds. Berlin, Germany: Springer, 2014, ch. 2, pp. 11–33.

[57] D. Ariely and G. Loewenstein, "When does duration matter in judgment and decision making?" *J. Exp. Psychol.: Gen.*, vol. 129, no. 4, pp. 508–523, 2000.

[58] N. Staelens *et al.*, "Assessing quality of experience of IPTV and video on demand services in real-life environments," *IEEE Trans. Broadcast.*, vol. 56, no. 4, pp. 458–466, Dec. 2010.

[59] P. Frohlich *et al.*, "QoE in 10 seconds: Are short video clip lengths sufficient for Quality of Experience assessment?" in *Proc. Int. Workshop Qual. Multimedia Experience*, Jul. 2012, pp. 242–247.

[60] S. Tavakoli, K. Brunnström, J. Gutiérrez, and N. García, "Quality of experience of adaptive video streaming: Investigation in service parameters and subjective quality assessment methodology," *Signal Process.: Image Commun.*, vol. 39, pp. 432–443, Nov. 2015.

[61] K. De Moor *et al.*, "Chamber QoE: A multi-instrumental approach to explore affective aspects in relation to quality of experience," in *SPIE Hum. Vis. Electron. Imag.*, B. E. Rogowitz, T. N. Pappas, and H. de Ridder, Eds., vol. 90140U, Feb. 2014.

[62] D. Zegarra Rodríguez, J. Abrahão, D. Coaquira Begazo, R. Lopes Rosa, and G. Bressan, "Video quality subjective assessment considering cognitive criteria and user preferences on video content," in *Proc. Braz. Symp. Multimedia Web*, 2012, Art. no. 269.

[63] E. J. David, J. Gutiérrez, A. Coutrot, M. P. DaS., and P. Le Callet, "A dataset of head and eye movements for 360 videos," in *Proc. ACM Multimedia Syst. Conf.*, 2018, pp. 432–437.

[64] H. Jun, M. R. Miller, F. Herrera, B. Reeves, and J. N. Bailenson, "Stimulus sampling with 360-Videos: Examining head movements, arousal, presence, simulator sickness, and preference on a large sample of participants and videos," *IEEE Trans. Affect. Comput.*, to be published, doi: 10.1109/TAFFC.2020.3004617.

[65] J. G. Beerends and F. E. De Caluwe, "The influence of video quality on perceived audio quality and vice versa," *J. Audio Eng. Soc*, vol. 47, no. 5, pp. 355–362, May 1999.

[66] X. Min, G. Zhai, J. Zhou, M. C. Q. Farias, and A. C. Bovik, "Study of subjective and objective quality assessment of audio-visual signals," *IEEE Trans. Image Process.*, vol. 29, pp. 6054–6068, Apr. 2020.

[67] D. S. Hands, "A basic multimedia quality model," *IEEE Trans. Multimedia*, vol. 6, no. 6, pp. 806–816, Dec. 2004.

[68] X. Min, G. Zhai, Z. Gao, C. Hu, and X. Yang, "Sound influences visual attention discriminately in videos," *Proc. in Int. Workshop Qual. Multimedia Experience*, 2014, pp. 153–158.

[69] X. Min, G. Zhai, C. Hu, and K. Gu, "Fixation prediction through multimodal analysis," *Vis. Commun. Image Process.*, Dec. 2015.

[70] X. Min *et al.*, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Trans. Image Process.*, vol. 29, pp. 3805–3819, Jan. 2020.

[71] F. Y. Chao *et al.*, "Audio-visual perception of omnidirectional video for virtual reality applications," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2020, pp. 1–6.

[72] A. Tse *et al.*, "Was I there? Impact of platform and headphones on 360 video immersion," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst.*, 2017, pp. 2967–2974.

[73] S.-Y. Kim, B.-S. Shin, and H. Choi, "Virtual reality based education with mobile device platform," *Mobile Inf. Syst.*, vol. 2019, Jan. 2019, Art. no. 6971319.

[74] N. Dużmańska, P. Strojny, and A. Strojny, "Can simulator sickness be avoided? A. review on temporal aspects of simulator sickness," *Front. Psychol.*, vol. 9, Nov. 2018, Art. no. 2132.

[75] H. Duan *et al.*, "Assessment of visually induced motion sickness in immersive videos," in *Proc. Adv. Multimedia Inf. Process.*, 2018, pp. 662–672.

[76] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness," *Int. J. Aviation Psychol.*, vol. 3, no. 3, pp. 203–220, Nov. 1993.

[77] K. Brunnström, K. Wang, S. Tavakoli, and B. Andrén, "Symptoms analysis of 3D TV viewing based on simulator sickness questionnaires," *Qual. User Experience*, vol. 2, no. 1, pp. 1–15, 2017.

[78] R. Vlad, O. Nahorna, P. Ladret, and A. Guerin, "The influence of the visualization task on the simulator sickness symptoms: A comparative SSQ study on 3DTV and 3D immersive glasses," in *Proc. 3DTV Conf.: Vis. Beyond Depth*, 2013, pp. 1–4.

[79] H. K. Kim, J. Park, Y. Choi, and M. Choe, "Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a VR environment," *Appl. Ergonom.*, vol. 69, pp. 66–73, May 2018.

[80] W. B. Stone III, "Psychometric evaluation of the simulator sickness questionnaire as a measure of cybersickness," Ph.D. dissertation, Iowa State University, 2017.

[81] B. Keshavarz and H. Hecht, "Validating an efficient method to quantify motion sickness," *Hum. Factors*, vol. 53, no. 4, pp. 415–426, Apr. 2011.

[82] A. H. Wertheim, J. E. Bos, and A. J. Krul, Predicting motion induced vomiting from subjective misery (MISC) ratings obtained in 12 experimental studies. Report TNO-TM-01-A066, Soesterbery, NL: TNO Human Factors Research Institute, 2001.

[83] P. Pérez, N. Oyaga, J. J. Ruiz, and A. Villegas, "Towards systematic analysis of cybersickness in high motion omnidirectional video," in *Proc. Int. Conf. Qual. Multimedia Experience*, 2018, pp. 1–3.

[84] H. T. T. Tran, N. P. Ngoc, C. T. Pham, Y. J. Jung, and T. C. Thang, "A subjective study on QoE of 360 video for VR communication," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2017, pp. 1–6.

[85] A. Singla, R. R. R. Rao, S. Göring, and A. Raake, "Assessing media QoE, simulator sickness and presence for omnidirectional videos with different test protocols," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces*, 2019, pp. 1163–1164.

[86] M. Elwardy, H.-J. Zepernick, and V. Sundstedt, "Annotated 360-Degree image and video databases: A comprehensive survey," in *Proc. Int. Conf. Signal Process. Commun. Syst.*, 2019, pp. 1–6.

[87] A. Abbas and B. Adsumilli, *AHG8: New GoPro Test Sequences for Virtual Reality Video Coding. JVET-D0026, 4th Meeting*: Chengdu, China, 15–21, Oct. 2016.

[88] E. Asbun, Y. He, Y. He, and Y. Ye, *AHG8: InterDigital Test Sequences for Virtual Reality Video Coding. JVET-D0039, 4th Meeting*: Chengdu, China, 15–21, Oct. 2016.

[89] H. Duan, G. Zhai, X. Yang, D. Li, and W. Zhu, "IVQAD 2017: An immersive video quality assessment database," in *Proc. Int. Conf. Syst., Signals Image Process.*, 2017, pp. 1–5.

[90] S. Fremerey, S. Göring, R. Rao, R. Huang, and A. Raake, "Subjective test dataset and meta-data-based models for 360° streaming video quality," in *Proc. Int. Workshop Multimedia Signal Process.*, 2020, pp. 1–6.

[91] B. J. Li, J. N. Bailenson, A. Pines, W. J. Greenleaf, and L. M. Williams, "A public database of immersive VR videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures," *Front. Psychol.*, vol. 8, Dec. 2017, Art. no. 2116.

[92] S. Fremerey, R. Huang, S. Göring, and A. Raake, "Are people pixel-peeping 360° videos?" *SPIE Electronic Imaging*, pp. 220-1–220-7(7), Jan. 2019.

[93] S. Knorr, C. Ozcinar, C. O. Fearghail, and A. Smolic, "Director's cut - a combined dataset for visual attention analysis in cinematic VR content," in *Proc. Eur. Conf. Vis. Media Prod.*, 2018, pp. 1–10.

[94] S. Fremerey, A. Singla, K. Meseberg, and A. Raake, "AVtrack360: An open dataset and software recording people's head rotations watching 360° videos on an HMD," in *Proc. ACM Multimedia Syst. Conf.*, 2018, pp. 403–408.

[95] X. Corbillon, F. De Simone, and G. Simon, "360-degree video head movement dataset," in *Proc. ACM Multimedia Syst. Conf.*, 2017, pp. 199–204.

[96] W. Lo *et al.*, "360 video viewing dataset in head-mounted virtual reality," in *Proc. ACM Multimedia Syst. Conf.*, 2017, pp. 211–216.

[97] A. T. Nasrabadi *et al.*, "A taxonomy and dataset for 360° videos," in *Proc. ACM Multimedia Syst. Conf.*, 2019, pp. 273–278.

[98] C. Cortés, P. Pérez, J. Gutiérrez, and N. García, "Influence of video delay on quality, presence, and sickness in viewport adaptive immersive streaming," in *Proc. Int. Conf. Qual. Multimedia Experience*, 2020, pp. 56–59.

[99] F. De Simone, J. Gutiérrez, and P. Le Callet, "Complexity measurement and characterization of 360-degree content," in *Proc. Hum. Vis. Electron. Imag.*, Jan. 2019, pp. 216–1–216-7.

[100] Y. Wang, Z. Chen, and S. Liu, "Equirectangular projection oriented intra prediction for 360-degree video coding," in *Proc. IEEE Int. Conf. Vis. Commun. Image Process.*, 2020, pp. 483–486.

[101] C. Cortés, P. Pérez, and N. García, "Unity3D-based app for 360VR subjective quality assessment with customizable questionnaires," in *Proc. IEEE Int. Conf. Consum. Electron.*, 2019, pp. 281–282.

[102] B. Winter, "Linear models and linear mixed effects models in r. with linguistic applications," 2013, *arXiv: 1308.5499*.

[103] R. H. Baayen, *Handbook of Laboratory Phonology*, A. C. Cohn, C. Fougeron, M. K. Huffman, Eds. Oxford U.K.: Oxford Univ. Press, pp. 668–677, Jan. 2012.

[104] J. Knaub, "Heteroscedasticity and homoscedasticity," in *Encyclopedia Meas. Statist.* Newbury Park, CA, USA: SAGE, 2007, pp. 431–432.

[105] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behav. Res. Methods*, vol. 39, no. 2, pp. 175–191, May 2007.

[106] K. Brunnström and M. Barkowsky, "Statistical quality of experience analysis: On planning the sample size and statistical significance testing," *J. Electron. Imag.*, vol. 27, no. 5, pp. 053 013–1-11, Sep. 2018.