Comparison of Subjective Methods for Quality Assessment of 3D Graphics in Virtual Reality

YANA NEHMÉ, Univ Lyon, LIRIS CNRS, France
JEAN-PHILIPPE FARRUGIA, Univ Lyon, LIRIS CNRS, France
FLORENT DUPONT, Univ Lyon, LIRIS CNRS, France
PATRICK LE CALLET, Univ Nantes, LS2N CNRS, France
GUILLAUME LAVOUÉ, Univ Lyon, LIRIS CNRS, France

Numerous methodologies for subjective quality assessment exist in the field of image processing. In particular, the Absolute Category Rating with Hidden Reference (ACR-HR), the Double Stimulus Impairment Scale (DSIS) and the Subjective Assessment Methodology for Video Quality (SAMVIQ) are considered three of the most prominent methods for assessing the visual quality of 2D images and videos. Are these methods valid/accurate to evaluate the perceived quality of 3D graphics data? Is the presence of an explicit reference necessary, due to the lack of human prior knowledge on 3D graphics data compared to natural images/videos? To answer these questions, we compare these three subjective methods (ACR-HR, DSIS and SAMVIQ) on a dataset of high-quality colored 3D models, impaired with various distortions. These subjective experiments were conducted in a virtual reality (VR) environment. Our results show differences in the performance of the methods depending on the 3D contents and the types of distortions. We show that DSIS and SAMVIQ outperform ACR-HR in terms of accuracy and point out a stable performance. In regard to the time-effort, DSIS achieves the highest accuracy in the shortest assessment time. Results also yield interesting conclusions on the importance of a reference for judging the quality of 3D graphics. We finally provide recommendations regarding the influence of the number of observers on the accuracy.

CCS Concepts: • **Computing methodologies** → **Appearance and texture representations**; *Perception*; *Mesh models*; Virtual reality.

Additional Key Words and Phrases: Visual quality assessment, 3D graphics, subjective methodologies, single stimulus, double stimulus, SAMVIQ, accuracy, time-effort

ACM Reference Format:
Yana Nehmé, Jean-Philippe Farrugia, Florent Dupont, Patrick Le Callet, and Guillaume Lavoué. 2020. Comparison of Subjective Methods for Quality Assessment of 3D Graphics in Virtual Reality. *ACM Transactions on Applied Perception* 18, 1, Article 2 (December 2020), 23 pages. https://doi.org/DOI:https://doi.org/10.1145/3427931

1 INTRODUCTION

Nowadays, three-dimensional (3D) graphics are widely used in many applications such as digital entertainment, architecture and scientific simulation. These data are increasingly rich and detailed; as a complex 3D scene may contain millions of geometric primitives, enriched with various appearance attributes such as texture maps designed to produce a realistic material appearance. These huge data tend to be visualized on various devices (e.g., smartphone, head mounted display) and possibly via the network. Therefore, to avoid latency or rendering issues, there is a critical need for the compression and simplification of these high quality 3D models. These processing operations may impact the visual quality of the 3D models and thus the quality of user experience (QoE). Thus to evaluate the visual quality as perceived by human observers, it is fundamental to resort to subjective quality

Authors' addresses: Yana Nehmé, yana.nehme@insa-lyon.fr, Univ Lyon, LIRIS CNRS, Lyon, France; Jean-Philippe Farrugia, Univ Lyon, LIRIS CNRS, Lyon, France, jean-philippe.farrugia@univ-lyon1.fr; Florent Dupont, Univ Lyon, LIRIS CNRS, Lyon, France, Florent.Dupont@liris.cnrs.fr; Patrick Le Callet, Univ Nantes, LS2N CNRS, Nantes, France, Patrick.Le-Callet@univ-nantes.fr; Guillaume Lavoué, Univ Lyon, LIRIS CNRS, Lyon, France, glavoue@liris.cnrs.fr.

2020. XXXX-XXXX/2020/12-ART2 $15.00
https://doi.org/DOI:https://doi.org/10.1145/3427931

assessment tests. In these tests, a group of human subjects is invited to judge the quality of a set of images subject to some predefined distortions. Such subjective experiments are also the most reliable way to create a ground-truth for judging the performance of objective quality metrics. However, selecting the best subjective methodology is not a trivial task since we should ensure that such method give valid and reliable results.

In the past years, several methodological guidelines have been defined for 2D image and video quality assessment by the International Telecommunication Union (ITU) [6, 7, 26]. In the field of computer graphics, previous subjective experiments were carried out to evaluate the visual quality of still and animated 3D models [9, 13, 22]. However, no comparison of subjective methodologies have been made for such 3D data. So, there is no consensus about the best methodology to adopt for quality assessment of 3D models. In this work, we propose to compare the performance of three of the most prominent methods, Absolute Category Rating with Hidden Reference (ACR-HR), Double Stimulus Impairment Scale (DSIS) and Subjective Assessment Methodology for Video Quality (SAMVIQ), on a dataset of high-quality colored 3D meshes. We also assess whether or not the presence of an explicit reference is necessary for evaluating the quality of such data. We chose to make the experiments in Virtual Reality (VR) using the HTC Vive Pro headset because VR is becoming a popular way of consuming and visualizing 3D content.

Our first psycho-visual experiment compares ACR-HR and DSIS. It is detailed in section 3 and the results are presented in section 4. Our second experiment, described in section 5, investigates the performance of SAMVIQ. In section 6, we provide the results of SAMVIQ as well as a comparison of the 3 tested methods. Finally, concluding remarks are outlined in sections 7 and 8. Note that, the only 3D representation used in this work is meshes, however, we believe that our results remain valid for other 3D representations, such as point clouds.

2 RELATED WORK

In this section we first review popular methodologies for subjective quality assessment of (natural) images and videos, and then focus on existing subjective tests conducted with 3D graphics. We finally discuss previous work that compares subjective methodologies. The reader is referred to [22] for a comprehensive survey of subjective quality assessment in computer graphics.

2.1 Methodologies for subjective quality assessment of images and videos

Several methodologies for 2D video/image quality assessment exist in the literature and have been standardized by the International Telecommunication Union [7]. Four subjective quality assessment methodologies are notably used nowadays: Absolute Category Rating (ACR), Double Stimulus Impairment Scale (DSIS), Subjective Assessment Methodology for Video Quality (SAMVIQ) and pairwise comparison (PC). The ACR method consists of presenting each impaired sequence individually to the observer and then asking him/her to rate its quality on a quality scale. In the DSIS method, the reference video is presented first, followed by the same video impaired. The observer is asked to rate, on an impairment scale, the degradation of the second video compared to its reference. These methods are categorical rating since they use a 5-level discrete scale [24]. They are dominant in video subjective quality tests [7, 26]. Furthermore, ACR with hidden reference (detailed in section 3) is notably used by the Video Quality Experts Group (VQEG) [39]. The pairwise comparison method (PC) is an alternative method in which two distorted videos are displayed, side by side, and the observer has to choose the one having the highest quality. The fourth method is SAMVIQ. It differs form the others in several aspects. SAMVIQ uses a multi-stimuli with random access approach [6, 15]. The test sequences are presented one at a time but the observer is able to review each video and modify the quality score multiple times. In addition, it uses a continuous quality scale (0-100). Note that for graphics applications requiring localized information on the distortion visibility, the methodology based on the local marking of visible distortions is commonly used [28, 41]. In such subjective experiments, observers manually mark the visible local artifacts in the impaired images.

2.2 Subjective quality assessment of 3D graphical models

When it comes to subjective tests involving 3D models, no specific standard or recommendation exist. Researchers have adapted existing image/video protocols, while considering different ways to display the 3D models to the observers (e.g., 2D still images, animated videos, interactive scenes). Lavoué et al. [21] and Corsini et al. [9] considered single stimulus protocols, derived from ACR, to assess the quality of impaired 3D meshes. The observers were able to freely interact with the 3D models and then had to rate the visibility of the distortions between 0 (invisible) and 10. Zerman et al. [43] considered ACR with Hidden Reference (ACR-HR) to compare the quality of two different representations (textured meshes and colored point clouds) for a volumetric video compression scenario. Despite these works, in the majority of existing experiments a double stimulus protocol (derived from DSIS) is used, with diverse modalities of display. Watson [40] used still screenshots to evaluate mesh simplification distortions, while Lavoué [20] considered free-viewpoint interactions for evaluating 3D meshes subject to smoothing and noise addition. Pan et al. [27] and Silva et al. [10] considered animations (e.g. low speed rotations) for assessing the quality of respective textured meshes and colored point clouds. Javaheri et al. [16] studied the impact of the different artifacts produced by point cloud codecs for different types of rendering. They generated 2D rendered videos from the original and decoded point clouds and selected DSIS method for their subjective test. According to them, DSIS allows to mitigate the impact of the density of original point clouds, as well as the impact of acquisition artifacts. Similarly, Alexiou et al. [3] assessed the influence of MPEG point cloud codecs. However, they adopted a real-time interactive evaluation protocol (rotation, translation, and zoom) to simulate realistic consumption. The authors used, in their subjective tests, the pairwise comparison (PC) approach when stimuli were nearly equal in quality and DSIS otherwise. Zerman et al. [42] also implemented both DSIS and PC to evaluate the quality of compressed volumetric videos represented as colored point clouds. Indeed, DSIS served to capture large differences introduced by compression, while PC was used to capture smaller differences related to point counts. Note that several recent works only used pairwise comparison methods [13, 38]. Moving to experiments in VR environments, it is worth mentioning the recent study of Subramanyam et al. [34], which assessed the quality of digital humans represented as dynamic point clouds, in two different VR viewing conditions enabling 3 and 6 degrees of freedom (DoF). This study was conducted using the ACR-HR method. In 2020, Alexiou et al. [4] developed a toolbox for subjective evaluation of point clouds in VR. It seems that most researchers intuitively felt that rating the absolute quality of a 3D graphical model (i.e., without the reference nearby) might be a difficult task for a naive observer (i.e. non-expert).

2.3 Comparison of subjective methodologies

Several works evaluate and compare the performance of the methodologies described above (mostly for natural image or video content). Péchard et al. [33] evaluated the impact of the video resolution on the behavior of both ACR and SAMVIQ methods. They found that, for a given number of observers, SAMVIQ is more accurate especially when the resolution increases. They also stated that the precision of the methods depends on the number of observers: 22 observers are required in ACR to obtain the same precision than SAMVIQ with 15 observers. Contrary to what the ITU recommends regarding the minimum number of subjects required for ACR (15), VQEG [39] and Brunnström et al. [5] recommended to use at least 24 observers. Nevertheless, the SAMVIQ method is considerably more time-consuming than an ACR (or DSIS) method.

Moving to double stimulus methods, the main difference between DSIS and ACR is the presence of explicit references. According to the ITU [26], DSIS ratings are less biased compared to ACR ratings. Indeed due to the presence of the references, subjects are able to detect shape and color impairments that they may miss with the ACR method. In addition, in DSIS, the scores are not influenced by the subjects opinion of the content. Surprisingly, Mantiuk et al.[24] denoted that for the experimental procedures, images and distortions used in their study, there was "no evidence that the double stimulus method is more accurate than the single stimulus method". They demonstrated that since the PC methodology is straightforward, it tends to be the most accurate from the 4 tested methods (single stimulus, double stimulus, forced choice pairwise comparison, and similarity judgments methods). However, despite the simplicity of the task of this method, it may become tedious if all sequences need to be tested (PC requires $\frac{n(n-1)}{2}$ trials to assess n sequences while ACR requires $n+1$ trials and DSIS requires n trials). Tominaga et al. [35] compared eight subjective assessment methods for mobile videos. They denoted that ACR, DSIS and SAMVIQ are the most reliable among the tested methods and showed that ACR is the most suitable method for quality assessment of mobile video services, in terms of total assessment time and ease of evaluation. In 2014, Kawano et al. [17] investigated the performance of ACR, DSIS and Double Stimulus Continuous Quality Scale (DSCQS) for assessing the quality of 2D and 3D Videos. They found that, ACR is the most time efficient and DSIS is the most stable. In terms of the discrimination ability, they stated that DSIS outperforms the others for low quality-video, while DSCQS is better for high-quality video. Recently, Alexiou et al. [2] extended their work [1] on quality assessment of point cloud geometry by comparing the results of the ACR and DSIS experiments, in which subjects were able to visualize the point clouds on a screen and interact with them. They found that, the DSIS method is more consistent in identifying the level of impairment. Singla et al. [32] evaluated the performance of the DSIS and a Modified Absolute Category Rating (M-ACR) method for omnidirectional (360°) videos using an Oculus Rift. They denoted that M-ACR is statistically slightly more reliable than DSIS since DSIS gave larger confidence intervals.

To conclude, while many methodological guidelines have been defined for natural video/image quality assessment (using a screen), no similar standards exist for quality evaluation of 3D graphics. No consensus has emerged toward the best methodology for such data, especially in a virtual or mixed reality environment. One particular open question is whether or not a reference is necessary. In this context, we compare the performance of the Double-Stimulus Impairment Scale method (DSIS), the Absolute Category Rating with Hidden Reference method (ACR-HR) and the Subjective Assessment Methodology for Video Quality (SAMVIQ) for assessing the quality of 3D graphics. We consider a VR context using the HTC Vive Pro headset, a high-end virtual reality headset. [1] The present work attempts to make a first step toward standardizing a methodology for assessing the quality of 3D graphics.

3 SUBJECTIVE EXPERIMENT 1

The purpose of this experiment is to determine the impact of the explicit reference on the quality assessment of 3D meshes. Thus, we conducted a psycho-visual experiment that compares the performance of 2 methods: ACR-HR (with hidden reference) and DSIS (with explicit references). We involved in this study 2 groups of observers. Each group did the 2 tests in different orders. The experiment was conducted in an immersive virtual environment, using the HTC Vive Pro, in the fixed position mode. Note that, no color calibration was performed for the headset. This section provides the details of this first subjective study.

3.1 Experimental methodologies

In this experiment, we investigate two categorical rating methods: a single and a double stimulus methods. The selected methodologies are presented below and illustrated in Figure 1.

- **Absolute Category Rating with Hidden Reference (ACR-HR)**: also known as single stimulus categorical rating, in which the impaired stimuli are presented one at a time in addition to the original unimpaired stimuli (references), without informing the subjects of their presence. The observers are asked to evaluate the quality of the stimulus shown using a Likert-type scale ranged from 1 to 5 (or five-level scale), where the discrete levels correspond to bad, poor, fair, good, and excellent. Note that some methods favour continuous rather than categorical scales to avoid quantizing errors [7]. According to ITU-R BT.500-13, the presentation time for

[1] https://www.vive.com

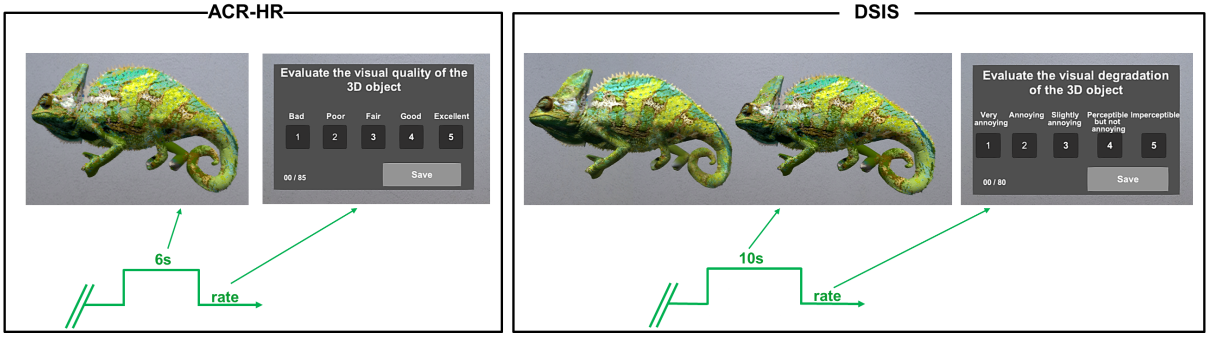


Fig. 1. Illustration and timeline of the two subjective quality assessment methods explored in this study.

the stimulus should be 10s. It may be reduced or increased according to the content of the test sequence [26]. In our pilot study (pretests), we found that 6s presentation is sufficient to assess the quality of the presented 3D model.

- **Double Stimulus Impairment Scale (DSIS)**: also called Degradation Category Rating (DCR), in which the viewer sees an unimpaired reference model, then the same model impaired. Following that, the subject is asked to rate the impairment of the second stimulus in relation to the reference [26] using the following five-level impairment scale: Imperceptible(5), Perceptible but not annoying(4), Slightly annoying(3), Annoying(2),Very annoying (1). Similarly to ACR, 10s presentation time is recommended per stimulus (≈20s/pair). However, this methodology slows-down the experiment too much since it requires at least twice as much time as ACR method. The overall length of the experiment affects the efficiency of the experimental method especially in virtual reality where most of the subjects are not used to the VR headset and tend to exhibit symptoms of cybersickness both during and after the VE experience [19]. To avoid these issues, we chose to display the reference and the test stimulus simultaneously side by side in the same scene. In this way, the number of presentations is halved. In addition, using simultaneous presentation makes the evaluation of the differences between the stimuli easier for the subjects [26]. Note that this "simultaneous" version of DSIS is what is preferred in most subjective tests involving 3D content [10, 20, 22, 27, 40]. For this methodology, we increased the presentation time to 10s, since, comparing to ACR-HR, 6s is not sufficient to observe the 2 stimuli displayed in the scene, compare them and assess their impairments.

3.2 Experiment design

For this experiment, observers were divided into two groups and were asked to rate the quality of a set of 80 distorted models (from 5 references), using both methodologies. ACR-HR and DSIS tests were made in different order according to the groups. Details about our study are described below.

3.2.1 Stimuli Generation.

We selected five triangle meshes of high resolution, each having a vertex color map: "Aix", "Ari", "Chameleon", "Fish", "Samurai" (see Figure 2). These 3D models are considered to be "good" or "excellent" quality. The number of vertices of the five models ranges from 250000 to 600000. They belong to very different semantic categories (human statues, animal, art) and have different shapes and colors (monocolor, warm cool and dull colors) (Figure 2). These reference models have been corrupted by 4 types of distortions, each applied with four different strengths:

- Uniform Geometric quantization (QGeo): applied on the geometry.
- Uniform LAB color quantization (QCol): applied on the vertex colors.

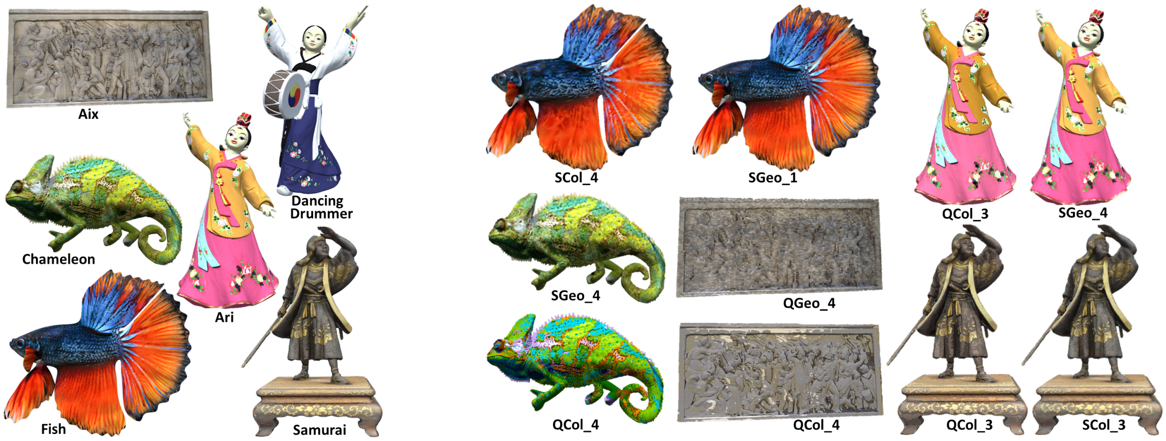


Fig. 2. Illustration of the 3D graphic reference models (Left) and some examples of distorted models (Right). Acronyms for distorted models refer to Type_Strength.

- "Color-ignorant" simplification (SGeo): mesh simplification algorithm that takes into account the geometry only [12].
- "Color-aware" simplification (SCol): mesh simplification algorithm that takes into account both geometry and color [23].

The strength of these distortions was adjusted manually in order to span the whole range of visual quality from imperceptible levels to high levels of impairment. For this task, a large set of distortions was generated and viewed by the authors, and a sub-set of them spanning the desired visual quality (i.e. “Excellent,” “Good,” “Fair,” and “Poor”) was chosen to be included in the database (as in [13, 31]). Thus, we generated 80 distorted models (5 reference models × 4 distortion types × 4 strengths). Figure 2 illustrates some visual examples.

3.2.2 *Rendering parameters.*

In designing our subjective experiment, we had to choose whether we select static or dynamic scenes. In fact, deciding the way the 3D models are displayed to the observers is a crucial problem. No standardized procedures exist for subjective evalution of the quality of 3D objects and current studies show a lack of generalization in the methodology that should be used [9, 13, 30]. Rogowitz et al. [30] proved that the perceived degradation of still images may not be adequate to evaluate the perceived degradation of the equivalent 3D model. Indeed, still images may mask both artifacts and the effect of light and shading. Following this approach, Corsini et al.[9] allowed the subject to interact with the model by rotating and zooming it. While it is important for the observer to have access to different viewpoints of the 3D object, the problem of allowing free interaction is the cognitive overload which may alter the results. Hence, we decided to control the interaction between the subject and the stimulus displayed on the scene. So, based on the principle of pseudo-videos and as in Guo et al. [13], we used animations. For each object in our database, we selected the viewpoint that covers most of the shape. We then applied a slow rotation of 15 degrees around the vertical axis in clockwise and then in counterclockwise directions (i.e. total rotation of 30 degrees). These dynamic stimuli are rendered in a virtual scene (using a perspective projection) at a viewing distance fixed to 3 meters from the observer and rotate in real-time. Note that, in DSIS test, the reference and the distorted model were specifically oriented in order to show exactly the same vertices of the 2 models at the same time. Stimulus size is approximately 36.87 degrees of visual angle. Its material type complies with the Lambertian reflectance model (diffuse surfaces). The apparent brightness of such a surface to an observer is the same regardless of the observer’s angle of view/position in the scene.

They are visualized in a neutral room (light gray walls) without shadows and under a directional light (all the vertices are illuminated as if the light is always from the same direction. It simulates the sun). We aimed to design a neutral room so that the experimental environment does not influence the users (quality-) perception of the stimulus. The default color calibration of the HTC Vive Pro was used.

3.2.3 Experimental procedure.

The goal of this experiment is to evaluate the impact of explicit references/test methodologies (ACR-HR, DSIS) on the user quality assessment. For this purpose, we divided our experiment into 2 sessions, one for each methodology i.e. one session consisted of presenting the stimuli using ACR-HR and the other session presented them using DSIS. In addition, in order to study whether a methodology has an influence over the other and if the order of the methodologies matters, we divided the subjects into 2 groups (G1 and G2). G1 refers to the participants who completed the ACR-HR test before DSIS and G2 refers to those who passed the DSIS session first then the ACR-HR session. None of these sessions took place on the same day in order to reduce the learning effect between stimuli. Thus, these two sessions occurred at least two days apart. In each session, the stimuli were displayed in a random order (3D models, distortions types and levels all mixed) to each observer. Each stimulus (for ACR-HR) or pair of stimuli (for DSIS) was presented once; the observer was not able to replay/review the objects.

Rating interface. We opted to ensure in our test a user experience and quality of experience (QoE) in fully immersive virtual environment (VE). So, we integrated a rating billboard in the VE of our experiment (see Figure 1). This board is adapted to each methodology and is displayed after the presentation time of each stimulus. There is no time limit to vote and the stimulus to rate is not shown during that time. The same neutral room (light gray walls) utilized to show the stimulus was used in the rating environment. To vote, the subject selects and saves the score using the trigger of the HTC Vive controller. As in [29], to facilitate the interaction with the rating panel, we attached a raycast beam to the controller.

Training. As recommended in ITU-R BT.500-13 [7], both sessions started with a training in which observers could familiarize themselves with the virtual environment and the task. We selected a training 3D model not included in our original test set: "Dancing Drummer" (see Figure 2) and generated 11 distorted models that span the whole range of distortions. At the beginning of each session, the training models are shown in the same manner (single or pairwise) and with the same time (6s or 10s) adopted in the upcoming session. After each stimulus, the rating panel, with the corresponding scale, is displayed for 5s. The score attributed/assigned to this distortion is highlighted. We added a practice trials stage at the end of the training: we displayed 2 extra stimuli and asked the subject to rate the quality or the impairment, according to the session. The results of these stimuli were not recorded. This stage was used to allow the observer to familiarize with the experimentation, to focus appropriately and to ensure that observers fully understand the task of the experiment.

Duration. No session took longer than 30 minutes to avoid fatigue and boredom: the total time was 18 minutes for the ACR-HR session (informed consent/instructions + 11 training stimuli × (6s display + 5s Rating) + 85 Test stimuli × (6s display + ≈4s rating)) and 23 minutes for DSIS session (informed consent/instructions + 11 training stimuli × (10s display + 5s Rating) + 80 Test stimuli × (10s display + ≈4s rating)).

The whole experience was developed in Unity3D using c# scripting. Snapshots of the experimental environment are provided in the supplementary material.

3.2.4 Participants.

As mentioned in the previous section, the stimuli were rated by 30 subjects divided into 2 groups of 15. The participants were students and professionals at the University of Lyon and LIRIS laboratory. 27 males and 3 females, aged between 19 and 45, they were naive about the purpose of the experiments. All observers had a normal or corrected to normal vision. In order to avoid the effect of the temporal sequencing factor, the order of stimuli was randomly generated so that each participant views the stimuli in a different order.

4 RESULTS AND ANALYSIS OF SUBJECTIVE EXPERIMENT 1

The following sections analyze and discuss the results of the experiment described above.

4.1 Screening observers

Before starting any analysis, participants were screened using the ITU-R BT.500-13 recommendation [7]. Applying this procedure on our data, we did not find any outlier participant from group 1 (G1). However, one subject from group 2 was rejected (G2) by reason of reporting implausible scores in the DSIS session (the first session for G2).

4.2 Computing the mean ratings MOS/DMOS

The first step of the analysis of the results is the calculation of the mean score for each of the stimuli [7].
For ACR-HR, it is advised to compute the difference scores between hidden reference and test stimuli instead of using directly the raw rating results. Indeed, studies [24, 37] show that subjects tend to assign a different quality scale for each object. Their rating is influenced by their opinion of the content (whether they like or dislike the object). Therefore, assessing differences in quality allows to take into account this variability in the use of the rating scale:

$$d_i^j = s_i^{ref(j)} - s_i^j \tag{1}$$

s_i^j refers to the score assigned by observer i to the stimulus j. *ref*(j) is the reference of stimulus j. The difference scores for the reference stimuli ($d_i^{ref(j)} = 0$) are removed from the collected data in the ACR-HR session for/of both groups G1 and G2. Finally, we computed the Difference Mean Opinion Score (DMOS) of each stimulus for both groups:

$$DMOS_j = \frac{1}{N} \sum_{i=1}^{N} d_i^j \tag{2}$$

N denotes the remaining subjects after screening observers i.e, N=15 for G1 and N=14 for G2.
For DSIS, we don't need to compute the DMOS since DSIS is based on the comparison between the reference and test models. Hence, we can directly use the rating results and compute the MOS.

$$MOS_j = \frac{1}{N} \sum_{i=1}^{N} s_i^j \tag{3}$$

4.3 Resulting MOS/DMOS

First, as recommended by VQEG [39], we computed the pairwise MOS correlation coefficient for the 2 groups of subjects, for each method. The (Pearson, Spearman rank order) correlation coefficient between G1's and G2's (D)MOS are (0.95, 0.92) for ACR-HR and (0.97, 0.94) for DSIS. Correlation values between subjects of the two groups are relatively high for both methods. Nevertheless, G1 and G2 subjects seem slightly more correlated with the DSIS method. We then further explored the intraclass correlation coefficient (ICC, Type (A,k) coefficients for two-way random effects model) [25] that analyzes the absolute agreement among the (D)MOS attributed to the stimuli by the two groups of subjects. The estimated ICC(A,k) for ACR-HR and DSIS are 0.89 and 0.96 respectively. Obviously, the correlation coefficients do not state everything, so we illustrate, in Figure 3, the results of ACR-HR and DSIS tests for all stimuli, averaged over all screened observers. To ensure a better readability in interpreting the results, we show the MOS (instead of the DMOS) for ACR-HR. Note that DMOS are used in the statistical tests presented in section 4.4. We show, in the supplementary material, a comparison of G1's and G2's DMOS, the computed (D)MOS and their confidence intervals separately for each group, as well as their box plots.

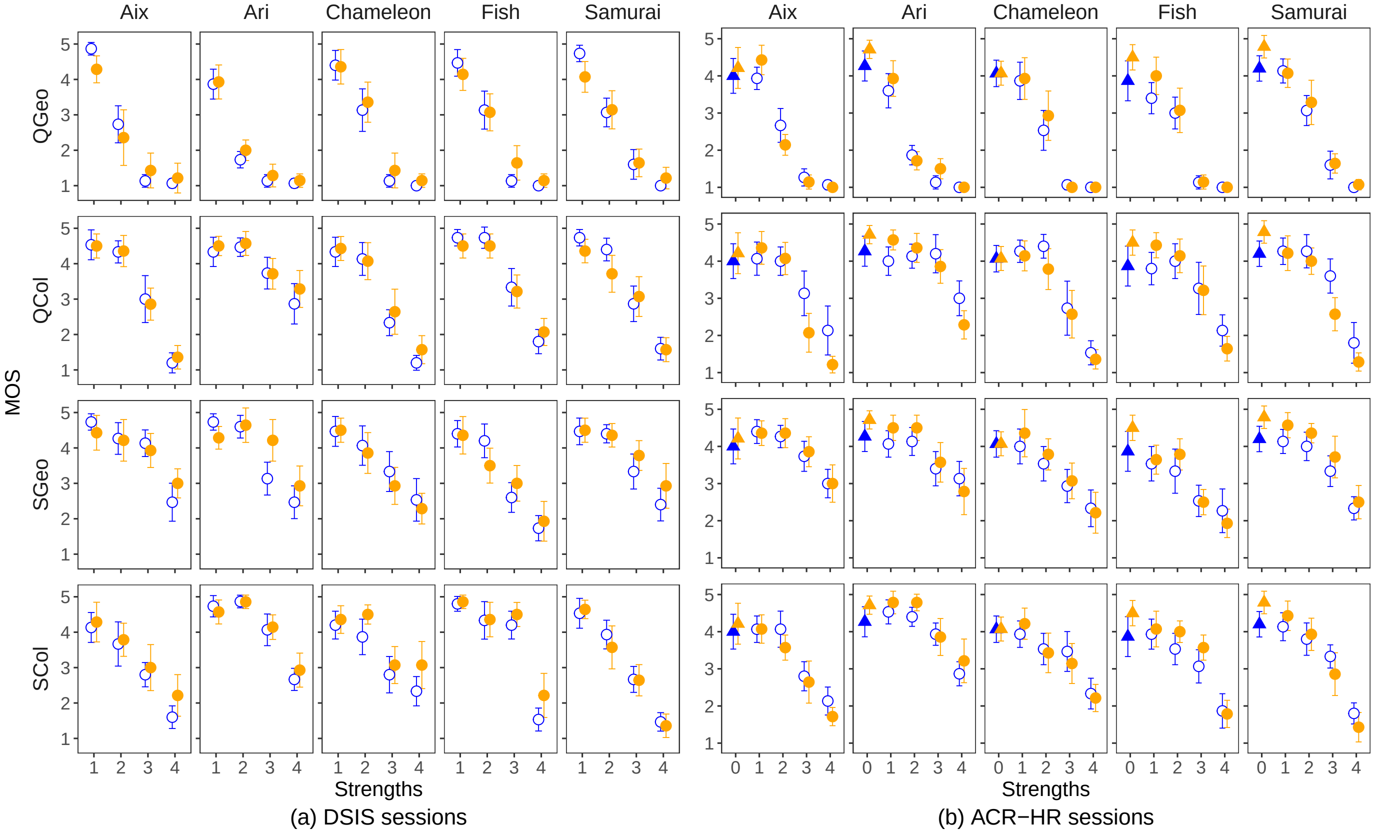


Fig. 3. Comparison of the G1's and G2's mean scores of the ACR-HR and DSIS experiments for all the stimuli (the blue and orange dots refer to the MOS of G1 and G2 respectively). For a given distortion strength, the dots are horizontally spaced apart to avoid overlapping.

As expected the MOS decrease as the distortion strengths increase. For the DSIS method (Figure 3.a), we can notice a strong consistency between the two groups and a good use of the entire rating scale. Indeed the observers of both groups showed almost the same behavior for each stimulus and the rating scores reach the scale limits.

For the ACR-HR method (Figure 3.b), we can notice some differences between the rating scores of the two groups. In fact, observers of G1 tend to downrate the reference stimuli, i.e. the rating scores given by G2 observers to almost all the references, except the *Chameleon*, exceed those of G1 observers. As a consequence, the amplitude of the rating scale is reduced. The specificity of the *Chameleon* model will be discussed in the next subsection. Moreover, we note that G2 observers were able to detect some distortions that G1 observers missed, notably the color distortions: e.g. *QCol* distortion with high strength (strengths ≥ 3) for *Aix*, *Ari* and *Samurai* (row 2) obtained better scores in G1 than in G2.

These first results reveal some differences in the performance and behavior of the methodologies. In the next section, we assess whether these differences are statistically significant and we attempt to provide explanations for their causes.

4.4 Quantitative analysis

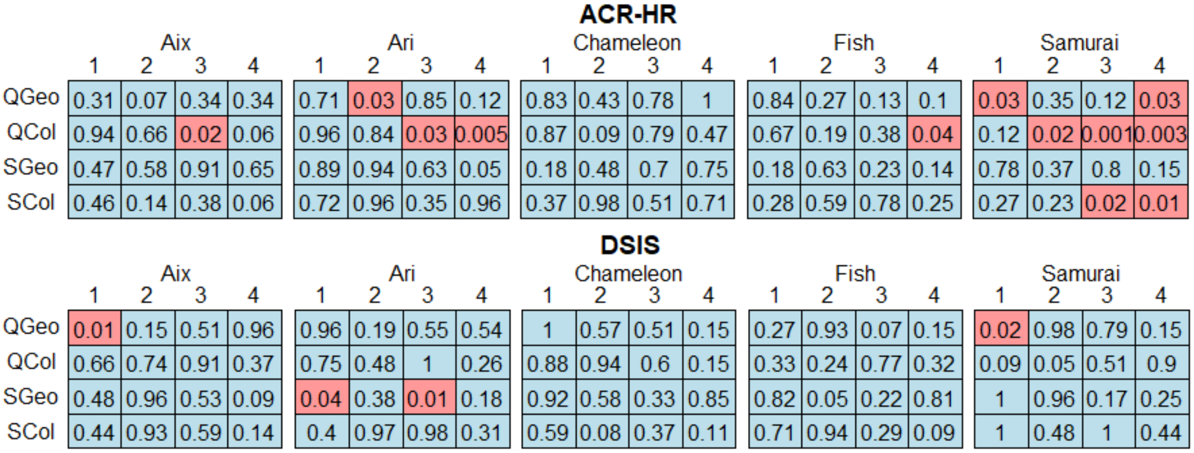
In this section, we analyze and compare quantitatively the results of both methodologies. In particular, we evaluate if the orders of the ACR/DSIS sessions have an impact on their results and why, and we demonstrate which methodology provides the most accurate results.

4.4.1 Normality and dependency analysis / preliminary tests.

The statistical analysis is affected by the dependencies between the samples. In our experiment, two groups of observers (G1 and G2) rated the same stimuli. The only difference for the two groups was the order of the ACR-HR/DSIS sessions. We aimed to test whether there are differences in scores between the two groups so we could evaluate whether a methodology has an influence over the other. Hence, for the analysis, the raw rating scores are independent and thus we could have used unpaired two-sample t-tests. However before using a parametric test, it is important to make sure that the data follow a normal distribution. We applied several normality tests, on the rating scores; such as Shapiro-Wilk's test, Lilliefors's test, Anderson-Darling's test. All these tests ascertained that the distribution of our data is not-normal (p-value $\ll 0.05$). Hence, for our data analysis, we have opted for the unpaired two-samples Wilcoxon test (also known as Wilcoxon rank-sum test or Mann-Whitney test). It is a non-parametric alternative to the unpaired two-samples t-test.

4.4.2 Consistency across the groups.

To assess whether, for a given methodology, there are significant differences in rating scores between the two groups of observers, we conducted for each stimulus the unpaired two-samples Wilcoxon test on the scores s_i^j (for DSIS) or the differential scores d_i^j (for ACR-HR) of the 2 groups. The null hypothesis (H0) is that, for a given stimulus, the rating scores of the G1's observers are equal to those of the G2's observers at the 95% confidence level. The alternative hypothesis (H1) is that the scores of G1 are greater (or lesser) than the scores of G2. The p-values are presented in Figure 4. The red boxes (p-value<0.05) indicate that the corresponding stimuli have been rated significantly different by the two groups of subjects.



ACR-HR

	Aix				Ari				Chameleon				Fish				Samurai			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
QGeo	0.31	0.07	0.34	0.34	0.71	0.03	0.85	0.12	0.83	0.43	0.78	1	0.84	0.27	0.13	0.1	0.03	0.35	0.12	0.03
QCol	0.94	0.66	0.02	0.06	0.96	0.84	0.03	0.005	0.87	0.09	0.79	0.47	0.67	0.19	0.38	0.04	0.12	0.02	0.001	0.003
SGeo	0.47	0.58	0.91	0.65	0.89	0.94	0.63	0.05	0.18	0.48	0.7	0.75	0.18	0.63	0.23	0.14	0.78	0.37	0.8	0.15
SCol	0.46	0.14	0.38	0.06	0.72	0.96	0.35	0.96	0.37	0.98	0.51	0.71	0.28	0.59	0.78	0.25	0.27	0.23	0.02	0.01

DSIS

	Aix				Ari				Chameleon				Fish				Samurai			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
QGeo	0.01	0.15	0.51	0.96	0.96	0.19	0.55	0.54	1	0.57	0.51	0.15	0.27	0.93	0.07	0.15	0.02	0.98	0.79	0.15
QCol	0.66	0.74	0.91	0.37	0.75	0.48	1	0.26	0.88	0.94	0.6	0.15	0.33	0.24	0.77	0.32	0.09	0.05	0.51	0.9
SGeo	0.48	0.96	0.53	0.09	0.04	0.38	0.01	0.18	0.92	0.58	0.33	0.85	0.82	0.05	0.22	0.81	1	0.96	0.17	0.25
SCol	0.44	0.93	0.59	0.14	0.4	0.97	0.98	0.31	0.59	0.08	0.37	0.11	0.71	0.94	0.29	0.09	1	0.48	1	0.44

Fig. 4. p-values computed between the rating scores of the two groups for all stimuli of both methodologies (the red color indicates a significant difference between the scores of G1 and G2).

For the ACR-HR method, we noticed that the scores of the two groups are not consistent (i.e. differ significantly) for 12 stimuli, out of 80; especially for the LAB quantization (QCol) of all the models excluding the *Chameleon*. This is coherent with the results observed in section 4.3. Our hypothesis is that this is due to the absence of explicit references. Indeed for G1's observers, as they did the ACR-HR test first, the assessment was absolute. Thus, it was difficult for them to detect the distortions of some models especially the color impairments, notably for Samurai and Ari (10 red boxes out of 12). The reason is that, for statues like Ari and Samurai, they have no prior knowledge of the exact color of the model. This is not the case for G2's observers since they had already seen the references during the DSIS session. Hence, they were able to detect the distortions (even the color distortions) that G1 observers might miss. For the *Chameleon*, there is no significant difference between the 2 groups. We believe that this is related to the fact that people have strong prior knowledge about this model: the chameleon/iguana is an animal known worldwide and everyone has an idea of its shape, color and geometry characteristics.

We observe, for certain models and distortion types notably the color quantization (Qcol), a better consistency/agreement among the subjects of the two groups, for the DSIS method. This confirms the fact that the presence of the reference makes the DSIS methodology more consistent across the groups and independent of the sessions order. The absence of explicit reference in the ACR-HR method makes it difficult for observers to assess certain distortions (e.g. color quantization), especially when they do not have prior knowledge about the models.

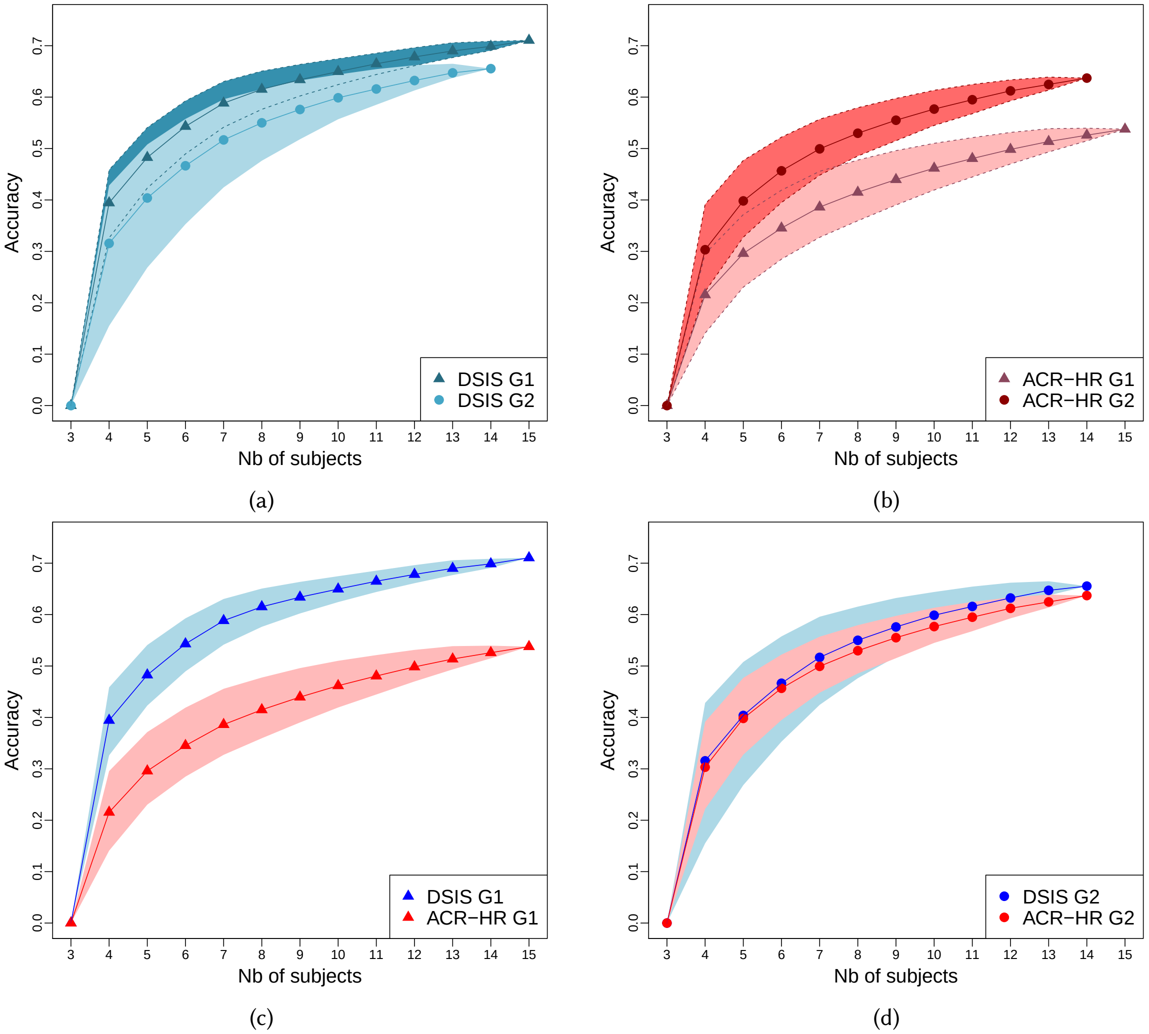
4.4.3 Accuracy of the quality scores.

As stated by Mantiuk et al. [24]: *« A more accurate method should reduce randomness in answers, making the pair of compared conditions more distinctive. A more accurate method should result in more pairs of images whose quality can be said to be different under a statistical test. »*. To assess the accuracy of the methodologies, we thus computed the number of pairs of stimuli rated significantly different by G1 and G2 subjects. For this task, we conducted unpaired two-samples Wilcoxon tests between rating scores of each possible pairs of stimuli. We conducted 80 x 79/2 = 3160 tests. The α levels used here is 0.05.

In order to study the behavior of this accuracy according to the number of subjects, we made these tests for different numbers of subjects and assessed the evolution of the number of pairs of stimuli significantly different. For each number N of subjects, we considered all possible combinations (without repetition) (with $3 \leq N \leq 15$ for G1 and $3 \leq N \leq 14$ for G2) and averaged the number of pairs significantly different over all these combinations of observers. Results are shown in Figure 5. The numbers of pairs in y-axis are given in percentages of the total number (i.e., 3160).

From Figure 5.a, it can be noticed that, for the DSIS method, the accuracy does not evolve much from G1 to G2. Hence, double stimulus methodology seems, once again, stable and independent of the sessions order. However, this is not the case of the ACR-HR method since the accuracy undergoes a large increase for G2 compared to G1 (Figure 5.b). This demonstrates anew that the method without explicit reference is not consistent across the groups. G2's subjects –who completed the ACR-HR test in the 2^{nd} session- were more familiar with the stimuli than G1's subjects since they had already seen the models and their references in the 1^{st} session (the DSIS test). Therefore, they are capable of distinguishing/detecting the degradations/loss in the visual quality of the stimuli more easily than the G1's observers. Beyond this better consistency observed for DSIS, Figures 5.c and 5.d clearly show that the DSIS method is more accurate than the ACR-HR method. This is valid even for G2, in which ACR-HR was conducted after DSIS. This finding corroborates previous results by Kawano et al. [17] obtained for stereoscopic 3D Videos. However, this result is inconsistent with comparative studies conducted with images and videos, including omnidirectional videos [32], in which M-ACR was slightly more reliable than DSIS. Our hypothesis is that people have more prior knowledge about the quality of 2D (natural) images/videos than 3D graphics, and therefore, the presence of references seems not necessary to assess the quality of these data.

The accuracy of a method is related to the agreement between raters, as an accurate method is intended to reduce random scores. Thus, we assessed, using the ICC (type (A,k) coefficients for two-way random effects models), the inter-rater reliability of each group, in the DSIS and ACR-HR tests: i.e. for a given method, the scores of a group of raters were compared to each other to evaluate how close the raters were in terms of their scores.



(a)

(b)

(c)

(d)

Fig. 5. Variation of the accuracy according to the number of subjects for both methodologies and both groups (G1's subjects did the ACR-HR session 1^{st} followed by the DSIS session, while G2's subjects did the DSIS session 1^{st} and then the ACR-HR session). The accuracy (y-axis) is defined as the percentage of pairs of stimuli whose qualities were assessed as statistically different. Curves represent mean values of these percentages and areas around curves represent 2.5th - 97.5th percentiles.

Results, shown in Figure 6, denote that the degree of agreement among raters is higher for DSIS than for ACR-HR. Moreover, subjects' agreement increased during the second session for both methods, yet this increase is larger for ACR-HR. We provide, in the supplementary material, an intra-rater reliability analysis among the 2 methods.

In Figure 7, we determined the number of subjects required in both methodologies to obtain the same accuracy. As can be seen in the figure, ACR-HR requires almost twice as many subjects as DSIS for G1. For instance, for a given number of observers unfamiliar with the test stimuli, ACR-HR requires minimum 15 observers to get a discrimination with an overall level of 53% while DSIS requires only 6 observers. Recommendations regarding

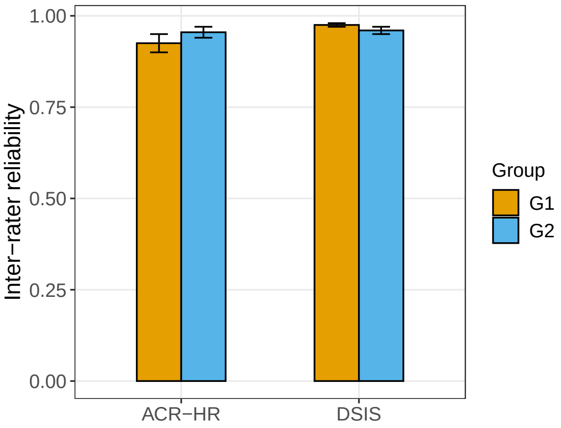


Fig. 6. Inter-rater reliability (as measured by ICC), of each group, in the DSIS and ACR-HR tests.

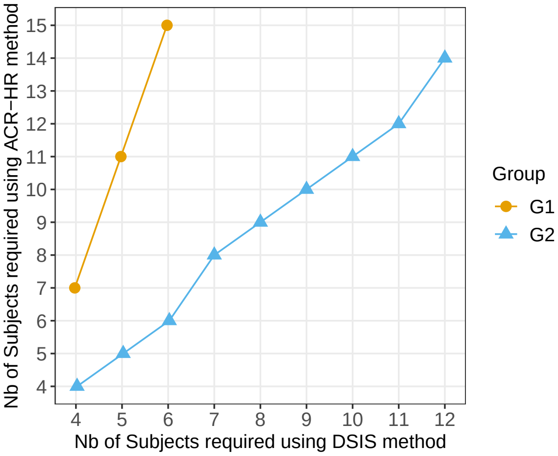


Fig. 7. Number of observers required to obtain the same accuracy with ACR-HR and DSIS methods.

the required number of observers for such experiments and the level of accuracy sought (considered acceptable) are discussed in section 7 below.

4.4.4 *Confidence intervals.*

Another way to evaluate the accuracy of the methodologies is to compute the 95% confidence intervals of the obtained MOS/DMOS. We thus computed these 95% confidence intervals (CI) for both groups and methodologies, in order to determine the "true" mean score (i.e. the interval in which the MOS/DMOS will reside if we have an ∞ number of observers) [7]. Then, we explored the evolution of the width of these intervals for both methodologies according to the number of subjects.

The curves of Figure 8 were obtained by averaging the width of CI over all the possible combinations of subjects. Note that for a given reference model and type of distortion, we average the widths of the CI over the four strengths of the distortion. We can observe that width of CI increases as the sample size decreases. For G1 (see Figure 8.a), we notice that, for most stimuli, the CI of the ACR-HR experiment are much larger than the CI given by the DSIS experiment, implying a strong dispersion of the ACR-HR scores across the G1's subjects. This disagreement is due to the fact that the references of the models are unknown for G1's subjects. This disagreement is not so apparent for G2 where the widths of CI given by the ACR-HR method are closer to the CI of the DSIS method (see Figure 8.b). These results confirm that DSIS is more accurate than ACR-HR, regardless the group. We illustrate, in the supplementary material, that there is almost no difference between the CI of G1 and G2 involved in DSIS, while the CI of G1's ACR-HR test are always superior to those of G2, except color quantization distortions of the *Chameleon*. Figure 8.c illustrates the confidence intervals of the *Chameleon*. As explained in section 4.4.2, the strong prior knowledge of observers on the color of this animal increases their accuracy, even without the presence of the explicit reference.

5 SUBJECTIVE EXPERIMENT 2

According to the results presented above, the presence of an explicit reference seems to be a necessity to improve not only the accuracy of the method but also to obtain lower confidence intervals. Therefrom, in order to find the best (the most suitable) methodology to adopt for quality assessment of 3D meshes, we decided to compare DSIS with another subjective quality assessment method which also presents explicit references. We chose SAMVIQ

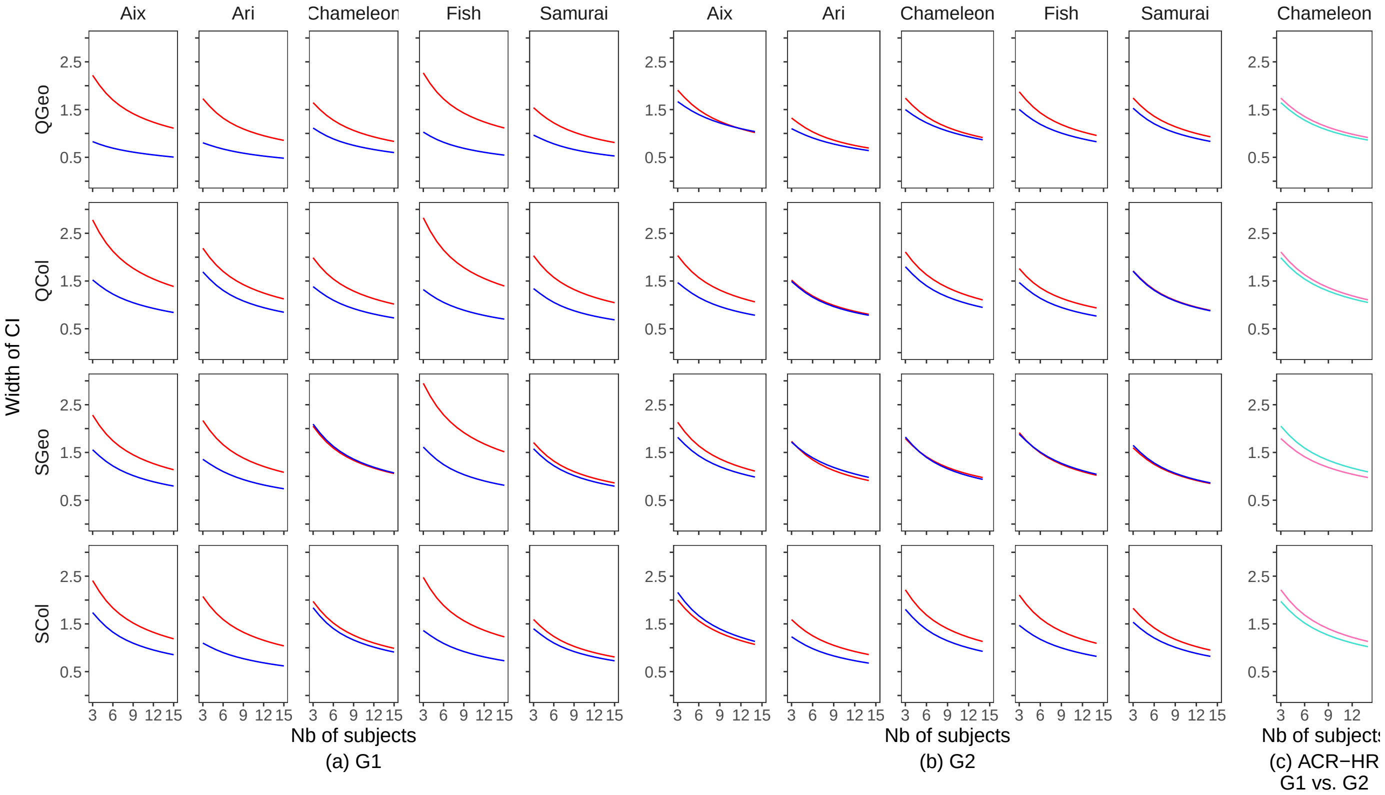


Fig. 8. Width of confidence intervals (CI) for both ACR-HR (red curves) and DSIS (blue curves) methodologies as a function of the number of observers involved in both groups (G1's subjects did the ACR-HR session 1^{st} followed by the DSIS session, while G2's subjects did the DSIS session 1^{st} and then the ACR-HR session). For (c), the turquoise and violet curves refer to CI of G1 and G2 respectively.

method, standardized within ITU-R [6]. Indeed in 2018, 3GPP[2] decided, in a study of VR media services [36], to start from the SAMVIQ method, since there is no existing standardized approach for subjective quality assessment in immersive environments. Note that, we did not consider the pairwise comparison method (PC), yet it is a widely used approach for assessing image quality, because this method has no explicit reference. Modifying the PC method to include an explicit reference would require the display of 3 stimuli in the scene (the reference and a pair of distorted stimuli to compare) which poses problems in a VR setting.

In this context, we conducted a second psychovisual experiment to investigate the performance of this method for assessing the visual quality of 3D graphics. As for the previous experiment involving ACR and DSIS, we considered a VR context using the HTC Vive Pro headset. This section provides the details of our second subjective study.

5.1 Experimental methodology

As mentioned, the selected methodology for the second experiment is SAMVIQ. It is presented below and illustrated in Figure 9.

[2] The 3rd Generation Partnership Project

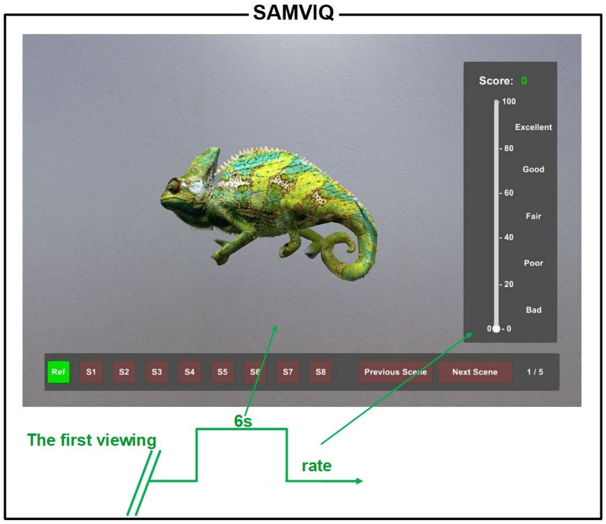


Fig. 9. Illustration of the SAMVIQ method explored in the second experiment of this study.

The Subjective Assessment Methodology for Video Quality (SAMVIQ) is a multiple stimuli assessment methodology [6]. This means that each stimulus can be seen and assessed as many times as the observers want: subjects are allowed to access the stimuli and adjust their scores, as appropriate.

The SAMVIQ test is divided into *scenes*. The content of a scene has to be homogeneous: a scene presents an explicit reference model as well as several versions of the same model with different impairments. Each stimulus is presented on its own and rated using a continuous quality scale. The continuous scale is graded from 0 to 100 (typically represented by a slider) and divided into five equal portions: Bad (0 to 20), Poor (20 to 40), Fair (40 to 60), Good (60 to 80), Excellent (80 to 100). The associated terms categorizing the different levels correspond to the basic ITU-R BT.500-13 five-level quality scale and are included for general guidance [11].

In SAMVIQ, the subjects are asked to assess the overall quality of the stimuli. Each stimulus (including the explicit reference) must be fully viewed at least once and then, the observer rates it. During the first viewing, all the other stimuli access buttons including the sliding rating scale are disabled. Once the current stimulus has been graded, the subject can access the previous rated stimuli to adjust their scores if needed. The last score of each stimulus remains recorded. Note that, the number of distorted stimuli is limited to ten per scene to avoid boredom and fatigue. All the stimuli of the current scene must be scored before the assessor can proceed to the next scene or visit the previous scene. To finish the test, all the stimuli of all the scenes must be scored.

This method is functionally similar to single stimulus method (i.e. ACR-HR) with random access, nevertheless a subject can view the explicit reference whenever he wants and compare it directly to the impaired stimulus. This makes SAMVIQ similar to methods that use a reference (i.e. DSIS)[6]. Following the ITU-R BT.1788 recommendations [6], the maximum presentation time for a stimulus is in the range of 10 to 15 s. Since SAMVIQ method provides a global score, like single stimulus methods (e.g. ACR), we set the presentation time to 6s, as in the ACR-HR test (section 3.1).

5.2 Experiment design

In order to be able to compare the results with those of the ACR-HR and DSIS tests, we consider the reference models (Aix, Ari, Chameleon, Fish, Samurai) and their distorted versions (16 impaired stimuli/reference) presented previously in section 3.2.1. Our experiment is organized such that each scene presents one reference model and its distortions. Thus the test consists of 5 scenes. The SAMVIQ methodology is limited to a maximum of 10 stimuli/conditions for each scene, excluding the references. Therefore, we divided the test into 2 sessions. Each test session contains the 5 reference models (=5 scenes), each corrupted by 2 types of distortions, applied with 4 strengths ((Ref + 8 impaired stimuli)/scene). Thus we ensure that both tests cover the entire quality range and have a balanced representation of visual qualities. We note that these two sessions occurred at least two days apart. For both sessions, the presentation order of the scenes was randomized across viewers.

For rendering, we proceed as described in section 3.2.2. The stimuli are loaded/displayed in a virtual room at 3 meters from the observers and animated in real-time with a slow rotation. The neutral room used is the same as that used for the ACR-HR and DSIS tests. As seen in the SAMVIQ's snapshot (Figure 9), we implemented a vertical slider directly on the right side of the test objects. It is a typical way used in SAMVIQ test to allow the subject to rate the quality according to the continuous scale (from 0 to 100). The subject selects the score (drag the slider to the chosen position) using the pad of the HTC Vive Controller and switches between stimuli and scenes using the controller trigger.

The SAMVIQ test sessions started with a training. The training procedure was identical to that of ACR-HR and DSIS training: we chose the "Dancing Drummer" model and 5 of its distorted models. The training models were presented for 6 s. Then the assigned score was marked on the slider for 10 s. We insisted in the training on the possibility of switching between stimuli to correct the scores. The training was followed by a practice stage (see section 3.2.3).

Each session lasted approximately 20 minutes to 30 minutes. It depended on the subject and how many times they viewed each test stimulus. The experience was developed in Unity3D.

5.3 Participants

A total of 17 non-expert subjects participated in the experiment: 4 females and 13 males, aged between 22 and 31. They were interns, PhD students and engineers at LIRIS laboratory. All subjects reported to have corrected to normal vision.

6 RESULTS AND ANALYSIS OF SUBJECTIVE EXPERIMENT 2 AND COMPARISON OF PROTOCOLS

In this section, we analyze the results of the SAMVIQ experiment. We also compare the performance of this method with that of ACR-HR and DSIS in terms of accuracy and time-effort. The following analysis/comparisons were carried out using the ACR-HR scores of G1 and the DSIS scores of G2. We chose these scores because the observers of G1 and G2 first performed the ACR-HR and DSIS sessions respectively and therefore the models were unknown for these subjects, as for the subjects of the SAMVIQ experiment. We provide, in the supplementary material, a table which summarizes all the experimental details of the three methods compared. Note that for the quantitative analysis of ACR-HR and SAMVIQ (section 6.3 and 6.4), we assessed the difference scores between references and test stimuli instead of using directly the raw scores (as explained in section 4.2).

6.1 Screening SAMVIQ Observers

The SAMVIQ screening procedure differs from the procedure described in Recommendation ITU-R BT.500-13. The SAMVIQ rejection criteria is based on a correlation of individual scores against corresponding mean scores from all the observers [6]. Applying this procedure, we find 2 outliers. Hence, only the data/scores of the 15 remaining subjects will be used in our data analysis.

6.2 Resulting MOS/DMOS

In this section, we compare the ACR-HR's, DSIS's and SAMVIQ's MOS for all the stimuli. In order to facilitate the comparison between the methods, we converted the ratings of SAMVIQ (ranged from 0 to 100) to those of

ACR and DSIS method (1 to 5 scale) as proposed in [14, 33]:

$$S'_{1-5} = \frac{S_{0-100} - 10}{20} + 1 \tag{4}$$

This linear mapping transforms the SAMVIQ scores to align the labels of the ACR-HR scale. Figures 10 and 11 show the results. Note that, for ACR-HR and SAMVIQ, we present the MOS (instead of the DMOS) for a better legibility of the results. We provide, in the supplementary material, the CIs associated with the (D)MOS of the 3 methodologies separated.

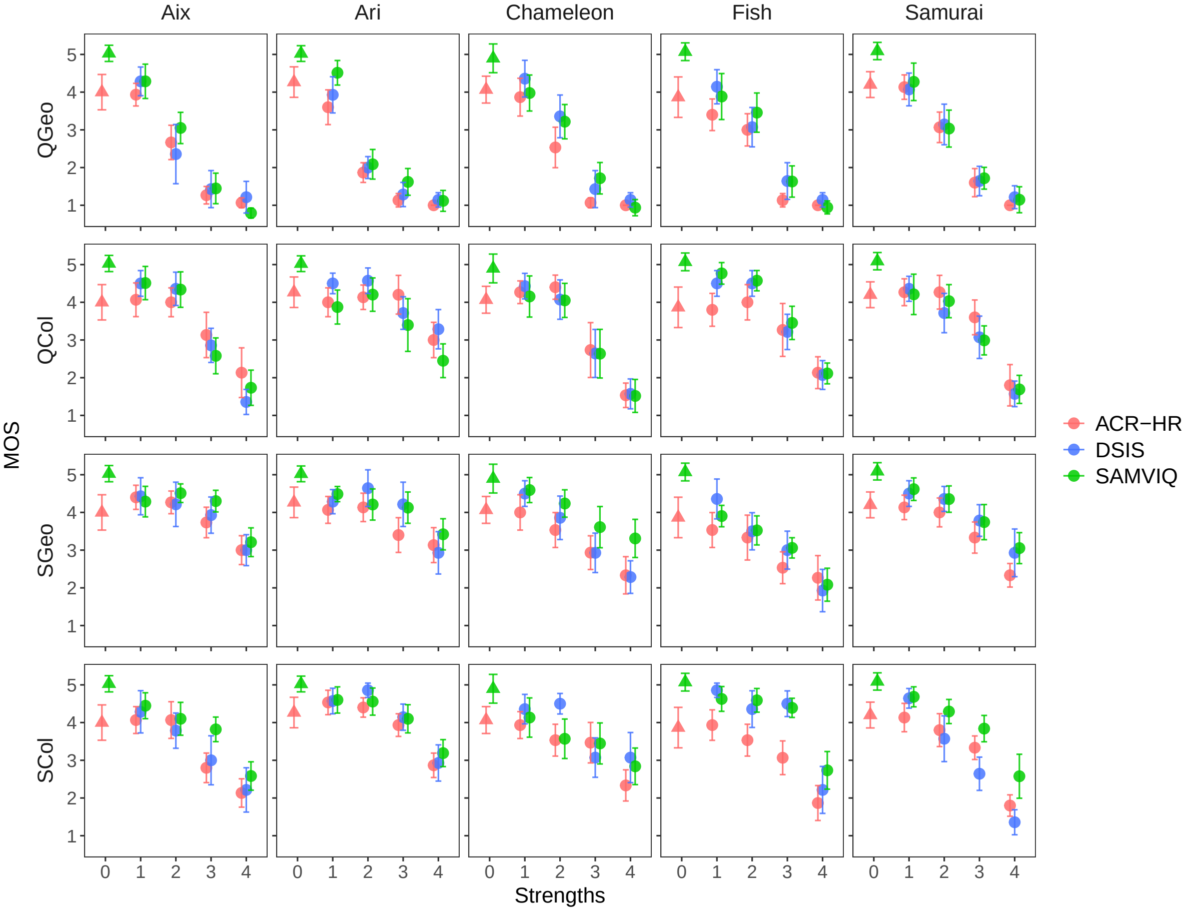


Fig. 10. Comparison of the mean scores of the ACR-HR, DSIS, and SAMVIQ experiments for all the stimuli. For a given distortion strength, the dots are horizontally spaced apart to avoid overlapping.

The 3 methodologies show almost the same behavior. Indeed, the pairwise (D)MOS correlation analysis indicated that the correlations of (D)MOS between pairs of tested methods are high, similarly to what was obtained by Tominaga et al. [35]. Table 1 summarizes the results. Nevertheless, as can be seen in Table 1 and Figure 10, the results of SAMVIQ seem slightly more correlated with those of DSIS than with those of ACR-HR: there is better consistency between the subjects of the SAMVIQ test and those of the DSIS test, notably for *Aix*

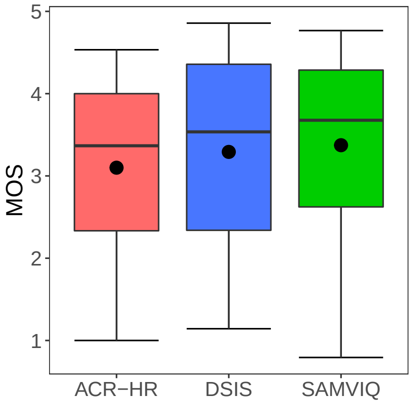


Fig. 11. Boxplots of MOSs obtained for tested methods.

	DSIS-ACR-HR	SAMVIQ-ACR-HR	SAMVIQ-DSIS
Pearson Correlation	0.943	0.937	0.946
Spearman Rank Order Correlation	0.885	0.883	0.906

Table 1. (D)MOS Correlation matrices for ACR-HR, DSIS, and SAMVIQ.

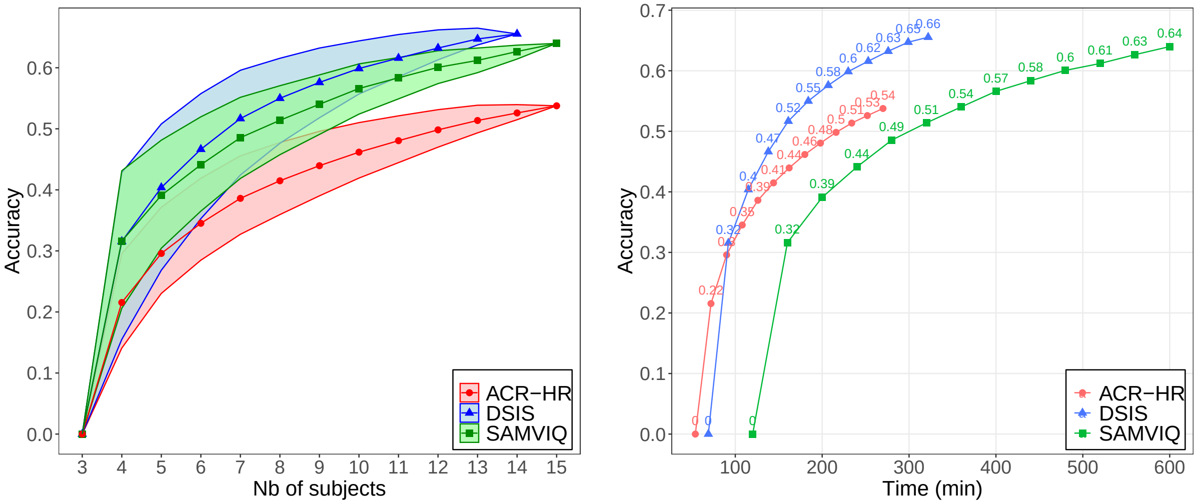
and *Fish* color quantized with low strengths (*QCol*, strengths ≤ 2), *Samurai* geometrically simplified (*SGeo*), and *Fish* color simplified (*SCol*). Concerning the scores attributed to the reference models (strength=0), SAMVIQ does not seem to downrate them, like ACR-HR does, since references are explicit in SAMVIQ. Moreover, Figures 10 and 11, show a difference in the use of the quality scale of each method. For a continuous scale (i.e. the SAMVIQ scale), subjects tend to avoid extremities and thus tend to use a smaller range of values, overall. This is known as the "Saturation Effect" [8, 14]. This effect is less visible for DSIS, since it uses a discrete categorical scale: no possible variations around best and worst qualities. In the following sections, we compare quantitatively these results.

6.3 Accuracy and time-effort of subjective assessment methods

First, we study the inter-rater reliability of SAMVIQ method. The degree of agreement among SAMVIQ raters is almost the same as that of DSIS raters (ICC(A,k)=0.96), and higher than that of ACR-HR raters (ICC(A,k)=0.93). We then further investigated the accuracy (defined in section 4.4.3) of SAMVIQ method and compare it to that of DSIS and ACR-HR: we computed the percentage of pairs of stimuli rated significantly different among all possible stimuli pairs and assessed its evolution according to the number of subjects. Note that, no-transformed ratings were used for SAMVIQ to compute the accuracy (no mapping at the same scale required). Figure 12.a shows the results.

The accuracy of the ACR-HR is smaller than that of SAMVIQ and DSIS. DSIS is slightly more accurate than SAMVIQ. Our hypothesis is that detecting the losses in the visual quality of the stimuli is easier and more obvious with the DSIS method than with the SAMVIQ method. In fact, we believe that the task of DSIS is simpler/more straightforward than that of SAMVIQ: the reference and the test stimulus are simultaneously displayed side by side in the scene and the subject is clearly asked to assess the impairments compared to the reference. However, SAMVIQ tends to be more complex since it uses a multi-stimuli with random access approach. Tominaga et al. [35] assessed the ease of evaluation of different methods and found that SAMVIQ, which has many grades on its quality scale, is more difficult than ACR-HR and DSIS.

To determinate the best methodology in subjective quality assessment tests, it is important to consider not only the accuracy of the methods, but also the time that observers need to complete the experiment. Ultimately, even less accurate methods may lead to smaller confident intervals (higher discrimination ability) if more data are collected [24]. Indeed, subjects may have difficulty maintaining their attentiveness throughout a long experiment



(a)

(b)

Fig. 12. Variation of the accuracy according to the number of subjects (a) and time-effort (b) for the 3 tested methodologies. The accuracy (y-axis) is defined as the percentage of pairs of stimuli whose qualities were assessed as statistically different. Curves represent mean values of these percentages and areas around curves represent 2.5th - 97.5th percentiles.

because of fatigue and boredom. This could skew the results of the experience. Thus, we compared the time-effort of each methodology. We determined the required time for these methods to reach a certain accuracy level. To do so, we multiplied the number of observers, used in abscissa of Figure 12.a, by the total time of each test session necessary to assess the whole dataset (80 distorted models + 5 references): 18 min for the ACR-HR session, 23 min for the DSIS session and 40 min for the SAMVIQ sessions. Note that for SAMVIQ, time may vary depending on how many times the subject viewed the stimuli. We considered the fastest scenario (≈ 20 min to assess 45 objects). Results are presented in Figure 12.b. DSIS is the most time-efficient method. SAMVIQ takes almost twice as long as DSIS to achieve the same accuracy: SAMVIQ requires minimum 600 min to get a discriminative power of 65% while DSIS requires only 300 min. Thus, SAMVIQ is considerably more time-consuming than DSIS (and ACR).

6.4 Confidence intervals

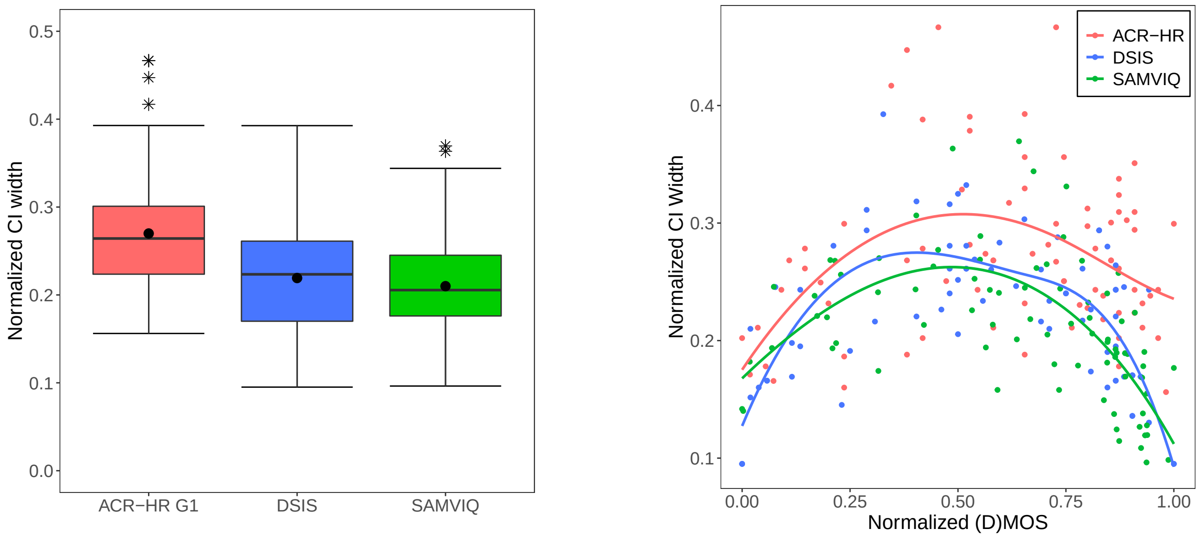
In this section, we evaluate the results of the subjective methods in terms of the dispersion of individual ratings (the standard deviation of the subjective scores). Thus, we compared the 95% CIs of the MOS (for DSIS) and DMOS (for ACR-HR and SAMVIQ) among the methods.

As in [18], we normalized the MOS and DMOS values as follows so that 0 means the lowest quality and 1 means the highest quality:

$$nMOS_i = \frac{MOS_i - min\{MOS_1 \ldots MOS_N\}}{max\{MOS_1 \ldots MOS_N\} - min\{MOS_1 \ldots MOS_N\}} \tag{5}$$

$$nDMOS_i = \frac{DMOS_i - max\{DMOS_1 \ldots DMOS_N\}}{min\{DMOS_1 \ldots DMOS_N\} - max\{DMOS_1 \ldots DMOS_N\}} \tag{6}$$

where i is the index of the stimulus and N is the total number of stimuli. We also normalized the CIs by expressing them as a percentage of the scale range. Figure 13 shows the boxplots of CIs, as well as the CIs in relation to their (D)MOSs, for the 3 methodologies.



(a)

(b)

Fig. 13. (a) Boxplots of CI obtained for the tested methods, (b) Comparison of normalized CI of the tested methods as a function of normalized (D)MOS.

The CIs of ACR-HR are significantly larger than those of the other methods, implying that ACR-HR has strong dispersion between the scores of the observers. These results are consistent with those presented in section 4.4.4. It is interesting to notice that, overall, SAMVIQ CI's are smaller than those of DSIS. Still, this difference is slight, since we performed the Student's t-test with a significance level of 5% between the CIs of DSIS and SAMVIQ and found no significant difference between the CIs of these 2 methods. We believe that SAMVIQ provides smaller CIs due to the subject's ability to review stimuli and adjust scores. Note that, despite the slightly smaller CIs of SAMVIQ, DSIS produced more accurate results, because the amplitude/range of the SAMVIQ rating scale, actually used by the subjects, is reduced/limited (i.e. the subjects did not use the whole scale, in the SAMVIQ test) (see Figures 10 and 11).

We can observe, in Figure 13.b, that the CIs of SAMVIQ tend to be larger than those of DSIS on the extreme MOS values ($nMOS/nDMOS \approx 0$ or 1). This is due to the fact that for the DSIS scale, there is no possible variations around best and worst qualities. However, for the SAMVIQ scale, subjects tend to avoid extremities, since the choice of the worst and the top scores is not limited to 0 and 100 only (saturation effect presented in section 6.2). We can also notice that the CIs of ACR-HR approach those of SAMVIQ and DSIS for the worst MOS values. However for the high MOS values, the dispersion of the ACR-HR scores remains high because the observers have not seen the references and therefore they have no prior knowledge of the best possible quality of stimuli.

We assessed the evolution of the confidence intervals according to the number of subjects, for the 3 methodologies, as described in section 4.4.4. The results are provided in the supplementary material. They point out the findings of this section and section 6.3.

7 DISCUSSION AND RECOMMENDATION

This section summarizes the results obtained in this study. We found that, for the quality assessment of 3D graphics, the ACR-HR method has a poor accuracy and large CIs compared to those of DSIS and SAMVIQ, and thus requires more subjects. In fact, in ACR-HR, the assessment is absolute (absence of explicit references) and therefore observers, who had never seen the reference models before, are not able to detect all the distortions, especially the color impairments. Thus, they tend to be less discriminating than those who are familiar with the test stimuli. This is not the case for the DSIS and SAMVIQ methods since they present explicit references. These 2 methods showed almost the same performance in terms of accuracy and CIs. DSIS appears to be slightly more accurate, while SAMVIQ offers slightly less dispersion in subjective rating. In regards to the time-effort, DSIS shows a great advantage. It is the most time-efficient, whereas SAMVIQ is considerably the most time-consuming: SAMVIQ takes twice as long as DSIS to achieve the same accuracy. Furthermore, the observers' task in SAMVIQ experiment is more difficult than that of DSIS (and ACR-HR).

Based on our results, we recommend the use of DSIS for the quality assessment of 3D graphics. We also attempt to make recommendations about the required number of observers for this methodology. For this purpose, we aggregate the DSIS test's scores of the 2 groups (G1 and G2 of the first experiment) and thus obtain 30 subjects. This aggregation is possible since we previously demonstrated that DSIS scores are consistent among the two groups. Thus, We recompute the accuracy (as in section 4.4.3) according to the number of observers.

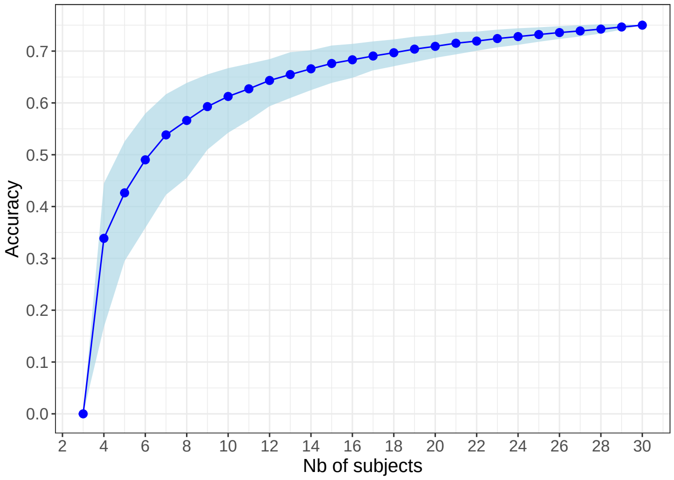


Fig. 14. Accuracy of the DSIS method according to the number of subjects.

From Figure 14, we observe that at least 19 test subjects are required to be able to discriminate 70% of all possible pairs of stimuli. With 15 observers, the recommended number by the ITU, we obtain an accuracy of 67%. However, with 25 subjects the discrimination increases to 73% and reaches 75% with 30 subjects. As a conclusion, and with regard to the shape of the curve, 24 subjects seem to be a good compromise.

8 CONCLUSION

In this study, we designed two psycho-visual experiments that compare three of the most prominent subjective methodologies, with and without explicit reference (ACR-HR, DSIS and SAMVIQ). We compare these methods for the quality assessment of 3D graphics in a VR environment. Results assert that the presence of an explicit reference is necessary to improve the accuracy and the stability of the method. This conclusion is not consistent with recent comparative studies conducted with images and videos. We believe that this is due to the fact that people have less prior knowledge of 3D graphics quality than of (natural) images. DSIS seems to be the most suitable method to assess the quality of 3D graphics. It is the most accurate and mainly the most time-efficient. We recommend to use groups of at least 24 observers for the DSIS methodology.

This study makes the first step toward standardizing a methodology for assessing the quality of 3D graphics in virtual reality.

ACKNOWLEDGMENTS

This work was supported by French National Research Agency as part of ANR-PISCo project (ANR-17-CE33-0005).

REFERENCES

[1] E. Alexiou and T. Ebrahimi. 2017. On subjective and objective quality evaluation of point cloud geometry. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX).* 1–3.

[2] Evangelos Alexiou and Touradj Ebrahimi. 2017. On the performance of metrics to predict quality in point cloud representations. In *Applications of Digital Image Processing XL*, Andrew G. Tescher (Ed.), Vol. 10396. International Society for Optics and Photonics, SPIE, 282 – 297. https://doi.org/10.1117/12.2275142

[3] Evangelos Alexiou, Irene Viola, Tomás M. Borges, Tiago A. Fonseca, Ricardo L. de Queiroz, and Touradj Ebrahimi. 2019. A comprehensive study of the rate-distortion performance in MPEG point cloud compression. *APSIPA Transactions on Signal and Information Processing* 8, e27. https://doi.org/10.1017/ATSIP.2019.20

[4] E. Alexiou, N. Yang, and T. Ebrahimi. 2020. PointXR: A Toolbox for Visualization and Subjective Evaluation of Point Clouds in Virtual Reality. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX).* 1–6.

[5] Kjell Brunnström and Marcus Barkowsky. 2018. Statistical quality of experience analysis: On planning the sample size and statistical significance testing. *Journal of Electronic Imaging* 27, 1. https://doi.org/10.1117/1.JEI.27.5.053013

[6] ITU-R BT.1788. 2007. Methodology for the subjective assessment of video quality in multimedia applications. *International Telecommunication Union.*

[7] ITU-R BT.500-13. 2012. Methodology for the subjective assessment of the quality of television pictures BT Series Broadcasting service. *International Telecommunication Union.*

[8] Philip Corriveau, Christina Gojmerac, Bronwen Hughes, and Lew Stelmach. 1999. All Subjective Scales Are Not Created Equal: The Effects of Context on Different Scales. *Signal Process.* 77, 1, 1–9. https://doi.org/10.1016/S0165-1684(99)00018-3

[9] Massimiliano Corsini, Elisa Drelie Gelasca, Touradj Ebrahimi, and Mauro Barni. 2007. Watermarked 3-D mesh quality assessment. *IEEE Transactions on Multimedia* 9, 247–256.

[10] Luis A. da Silva Cruz, Emil Dumic, Evangelos Alexiou, Joao Prazeres, Rafael Duarte, Manuela Pereira, Antonio Pinheiro, and Touradj Ebrahimi. 2019. Point cloud quality evaluation : Towards a definition for test conditions. *International Conference on Quality of Multimedia Experience*, 6. http://infoscience.epfl.ch/record/264995

[11] EBU. 2003. SAMVIQ - Subjective Assessment Methodology for Video Quality. *Tech. Rep. BPN 056, European Broadcasting Union.*

[12] Michael Garland and Paul S. Heckbert. 1997. Surface simplification using quadric error metrics. In *ACM Siggraph.* 209–216.

[13] Jinjiang Guo, Vincent Vidal, Irene Cheng, Anup Basu, Atilla Baskurt, and Guillaume Lavoue. 2016. Subjective and Objective Visual Quality Assessment of Textured 3D Meshes. *ACM Transactions on Applied Perception* 14, 1–20. https://doi.org/10.1145/2996296

[14] Q. Huynh-Thu, M. Garcia, F. Speranza, P. Corriveau, and A. Raake. 2011. Study of Rating Scales for Subjective Quality Assessment of High-Definition Video. *IEEE Transactions on Broadcasting* 57, 1, 1–14.

[15] Quan Huynh-Thua and Martlesham Heath. 2007. Examination of the SAMVIQ methodology for the subjective assessment of multimedia quality. *Third Inter. Workshop on Video Processing and Quality Metrics for Consumer Electronics.*

[16] Alireza Javaheri, Catarina Brites, Fernando Pereira, and Joao Ascenso. 2019. Point Cloud Rendering after Coding: Impacts on Subjective and Objective Quality. arXiv:eess.IV/1912.09137

[17] Taichi KAWANO, Kazuhisa YAMAGISHI, and Takanori HAYASHI. 2014. Performance Comparison of Subjective Assessment Methods for Stereoscopic 3D Video Quality. *IEICE Transactions on Communications* E97.B, 4, 738–745. https://doi.org/10.1587/transcom.E97.B.738

[18] Kimiko Kawashima, Kazuhisa Yamagishi, and Takanori Hayashi. 2018. Performance Comparison of Subjective Quality Assessment Methods for 4k Video. *IEICE Transactions* 101-B, 933–945.

[19] Joseph J. LaViola. 2000. A discussion of cybersickness in virtual environments. *ACM SIGCHI Bulletin* 32, 47–56. https://doi.org/10.1145/333329.333344

[20] Guillaume Lavoué. 2009. A local roughness measure for 3D meshes and its application to visual masking. *ACM Transactions on Applied Perception (TAP)* 5, 4.

[21] Guillaume Lavoue, Elisa Drelie Gelasca, Florent Dupont, Atilla Baskurt, and Touradj Ebrahimi. 2006. Perceptually driven 3D distance metrics with application to watermarking. *Proceedings of SPIE - The International Society for Optical Engineering* 6312. https://doi.org/10.1117/12.686964

[22] G. Lavoué and R.K. Mantiuk. 2015. Quality assessment in computer graphics. In *Visual Signal Quality Assessment: Quality of Experience (QoE).* Springer, 243–286. https://doi.org/10.1007/978-3-319-10368-6_9

[23] Ho Lee, Guillaume Lavoué, and Florent Dupont. 2012. Rate-distortion optimization for progressive compression of 3D mesh with color attributes. *The Visual Computer* 28, 2, 137–153. https://doi.org/10.1007/s00371-011-0602-y

[24] Rafał K. Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. 2012. Comparison of Four Subjective Methods for Image Quality Assessment. *Computer Graphics Forum* 31, 8, 2478–2491. https://doi.org/10.1111/j.1467-8659.2012.03188.x

[25] Kenneth Mcgraw and S.P. Wong. 1996. Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods* 1, 30–46. https://doi.org/10.1037/1082-989X.1.1.30

[26] ITU-T P.910. 2009. Subjective video quality assessment methods for multimedia applications. *International Telecommunication Union.*

[27] Yixin Pan, I Cheng, and A Basu. 2005. Quality metric for approximating subjective evaluation of 3-D objects. *IEEE Transactions on Multimedia* 7, 2, 269–279. https://doi.org/10.1109/TMM.2005.843364

[28] Rafal Piórkowski, Radoslaw Mantiuk, and Adam Siekawa. 2017. Automatic Detection of Game Engine Artifacts Using Full Reference Image Quality Metrics. *ACM Transactions on Applied Perception* 14, 3, Article 14, 17 pages. https://doi.org/10.1145/3047407

[29] Georg Regal, Raimund Schatz, Johann Schrammel, and Stefan Suette. 2018. VRate: A Unity3D Asset for integrating Subjective Assessment Questionnaires in Virtual Environments. In *10th International Conference on Quality of Multimedia Experience, QoMEX 2018.* 1–3. https://doi.org/10.1109/QoMEX.2018.8463296

[30] Bernice E Rogowitz and Holly E Rushmeier. 2001. Are image quality metrics adequate to evaluate the quality of geometric objects?, In Photonics West 2001-Electronic Imaging. *Proc SPIE.* https://doi.org/10.1117/12.429504

[31] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K. Cormack. 2010. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing* 19, 1427–41. https://doi.org/10.1109/TIP.2010.2042111

[32] Ashutosh Singla, Werner Robitza, and Alexander Raake. 2018. Comparison of Subjective Quality Evaluation Methods for Omnidirectional Videos with DSIS and Modified ACR. *Electronic Imaging* 2018, 14. https://doi.org/10.2352/ISSN.2470-1173.2018.14.HVEI-525

[33] Stephane Péchard, Romuald Pepion, Patrick Le Callet. 2008. Suitable methodology in subjective video quality assessment: a resolution dependent paradigm. *International Workshop on Image Media Quality and its Applications, IMQA2008.*

[34] S. Subramanyam, J. Li, I. Viola, and P. Cesar. 2020. Comparing the Quality of Highly Realistic Digital Humans in 3DoF and 6DoF: A Volumetric Video Case Study. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR).* 127–136.

[35] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi. 2010. Performance comparisons of subjective quality assessment methods for mobile video. In *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX).* 82–87.

[36] 3GPP TR 26-918 V15.2.0. 2018. Virtual Reality (VR) media services over 3GPP. *Technical Specification Group Services and System Aspects.*

[37] Andre M. van Dijk, Jean-Bernard Martens, and Andrew B. Watson. 1995. Quality asessment of coded images using numerical category scaling, Vol. 2451. Proceedings of SPIE, 90–101. https://doi.org/10.1117/12.201231

[38] K. Vanhoey, B. Sauvage, P. Kraemer, and G. Lavoué. 2017. Visual quality assessment of 3D models: On the influence of light-material interaction. *ACM Transactions on Applied Perception* 15, 1.

[39] VQEG. 2007. Multimedia Test Plan 1.19.

[40] B Watson, A Friedman, and A McGaffey. 2001. Measuring and predicting visual fidelity. In *Proc. of SIGGRAPH 2001.* ACM, 213–220. https://doi.org/10.1145/383259.383283

[41] Krzysztof Wolski, Daniele Giunchi, Nanyang Ye, Piotr Didyk, Karol Myszkowski, Radoslaw Mantiuk, Hans-Peter Seidel, Anthony Steed, and Rafal Mantiuk. 2018. Dataset and Metrics for Predicting Local Visible Differences. *ACM Transactions on Graphics* 37, 1–14. https://doi.org/10.1145/3196493

[42] Emin Zerman, Pan Gao, Cagri Ozcinar, and Aljosa Smolic. 2019. Subjective and objective quality assessment for volumetric video compression. *electronic imaging* 2019, 323–1–323–7.

[43] E. Zerman, C. Ozcinar, P. Gao, and A. Smolic. 2020. Textured Mesh vs Coloured Point Cloud: A Subjective Study for Volumetric Video Compression. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX).* 1–6.