IEICE TRANSACTIONS

on Communications

VOL. E101-B NO. 3
MARCH 2018

A PUBLICATION OF THE COMMUNICATIONS SOCIETY

The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Performance Comparison of Subjective Quality Assessment Methods for 4k Video

Kimiko KAWASHIMA[†a)], **Kazuhisa YAMAGISHI**[†], ***and*** **Takanori HAYASHI**[†], ***Members***

SUMMARY Many subjective quality assessment methods have been standardized. Experimenters can select a method from these methods in accordance with the aim of the planned subjective assessment experiment. It is often argued that the results of subjective quality assessment are affected by range effects that are caused by the quality distribution of the assessment videos. However, there are no studies on the double stimulus continuous quality-scale (DSCQS) and absolute category rating with hidden reference (ACR-HR) methods that investigate range effects in the high-quality range. Therefore, we conduct experiments using high-quality assessment videos (high-quality experiment) and low-to-high-quality assessment videos (low-to-high-quality experiment) and compare the DSCQS and ACR-HR methods in terms of accuracy, stability, and discrimination ability. Regarding accuracy, we find that the mean opinion scores of the DSCQS and ACR-HR methods were marginally affected by range effects, although almost all common processed video sequences showed no significant difference for the high- and low-to-high-quality experiments. Second, the DSCQS and ACR-HR methods were equally stable in the low-to-high-quality experiment, whereas the DSCQS method was more stable than the ACR-HR method in the high-quality experiment. Finally, the DSCQS method had higher discrimination ability than the ACR-HR method in the low-to-high-quality experiment, whereas both methods had almost the same discrimination ability for the high-quality experiment. We thus determined that the DSCQS method is better at minimizing the range effects than the ACR-HR method in the high-quality range.

key words: *4k video, subjective quality assessment, range effect*

1. Introduction

In recent years, 4k video services have been attracting attention for their use in next-generation video services. Some electronics manufacturers have been providing 4k cameras and televisions (TVs) [1], [2]. Spanning the world, 4k broadcast trials were conducted from the FIFA World Cup in Brazil and the Sochi (Russia) Winter Olympics in 2014. To provide high-quality 4k video services, a video service needs to be designed and managed on the basis of the end-user quality of experience [3]. To achieve quality-based service design and management, a methodology is needed for assessing the quality of 4k video services.

The International Telecommunication Union (ITU) has standardized many subjective quality assessment methods such as the absolute category rating (ACR, [4]), ACR with hidden reference (ACR-HR, [4]), degradation category rating (DCR, [4], [5]), double stimulus continuous quality-scale (DSCQS, [5]), and pair comparison (PC) methods [4]. These methods have different specific features. Therefore, experimenters need to select a method in accordance with the aim of the planned subjective assessment experiment. The performances of these subjective quality assessment methods have been sufficiently investigated and compared in terms of correlation [6]–[14], stability [10], [11], discrimination ability [11], [12], [15], and efficiency [10], [11], [13], [14].

In addition, it is often argued that these subjective quality assessment methods are affected by subject bias [16], [17] and range effects [12], [18], [19]. Subject bias means the overall shift between a subject's scores and the true value [16]. To exclude the effects of subject bias, the subjective model for the ACR method [16] and that for the PC method [17] have been proposed, and these methods provide experimenters with the knowledge for determining the number of subjects and the number of stimuli to exclude the effects of subject bias.

On the other hand, the range effects have not been sufficiently investigated or compared [12], [18], [19]. Range effect means that viewers unconsciously adjust their ratings on the basis of the total range of qualities in the test [18]. The effects are based on the video quality distribution for the assessment sequences. To compare two subjective quality assessment results with two different video quality distributions, it is necessary to clarify whether the subjective quality assessment method is affected by the different video quality distributions for assessment sequences. Therefore, range effects need to be compared.

In this study, we focused on range effects with the DSCQS and ACR-HR methods. The reason is that they are said to be the two best methods to minimize range effects because their mean opinion scores (MOSs) are calculated by the difference in video quality between the reference and assessment videos. There are no studies that compare range effects between an experiment using high-quality assessment videos (defined hereafter as a high-quality experiment) and an experiment using assessment videos that range in quality from low to high (defined hereafter as a low-to-high-quality experiment) with these methods. In addition, as viewers expect high quality from 4k video services, it is important to investigate the range effects of the subjective quality assessment methods for high-quality assessment videos. Therefore, our aim in this study is to investigate range effects in a high-quality experiment using the DSCQS and ACR-HR methods.

To do this, we conducted high-quality and low-to-high-

Manuscript received March 31, 2017.
Manuscript revised July 25, 2017.
Manuscript publicized August 29, 2017.

[†]The authors are with NTT Network Technology Laboratories, NTT Corporation, Musashino-shi, 180-8585 Japan.
a) E-mail: kawashima.kimiko@lab.ntt.co.jp

DOI: 10.1587/transcom.2017EBP3123

quality experiments. We compared the DSCQS and ACR-HR methods from the viewpoints of accuracy, stability, and discrimination ability. To examine their accuracy, we first examined the correlation between Pearson's correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC) for the MOSs of the DSCQS method (defined hereafter as DSCQS values) and for those of the ACR-HR method (defined hereafter as ACR-HR values), both derived from the high- and low-to-high-quality experiments. Next, to examine their stability, we compared the confidence intervals (CIs) of DSCQS values and ACR-HR values. Then to examine their discrimination ability, we used a Student's t-test. Finally, from these results, we determined which subjective quality assessment method better minimizes the range effects in the high-quality range in terms of accuracy, stability, and discrimination ability.

This paper is organized as follows. Previous studies related to the range effects of subjective quality assessment methods are explained in Sect. 2. The methods compared in this paper are explained in Sect. 3. Then, we describe the experimental conditions in Sect. 4, and discuss the experimental results in Sects. 5 and 6. Finally, in Sect. 7, we present the conclusion and discuss further studies.

2. Related Work

In this section, we describe studies on the performance, subject bias, and range effects of subjective quality assessment methods.

2.1 Performance Comparison

The performance of subjective quality assessment methods has been sufficiently investigated and compared in terms of correlation, stability, discrimination ability, and efficiency [6]–[15].

Various studies [6]–[14] have revealed correlations between MOSs of many subjective quality assessment methods.

Their stability has also been investigated [10], [11]. Tominaga et al. [10] compared five subjective quality assessment methods (DSCQS, double stimulus impairment scale (DSIS), ACR-HR, ACR, and ACR with an 11-grade scale) for both High Definition (HD) and Quarter Video Graphics Array (QVGA) resolutions. The DSIS and ACR methods outperformed the other methods for HD and QVGA, respectively. Kawano et al. [11] compared three subjective quality assessment methods (ACR, DCR, and DSCQS) for both 2D and 3D videos. They found that the stability of the DCR method is most suitable for low-quality video and that of the DSCQS and DCR methods is high for high-quality video.

Discrimination ability has also been investigated [11], [12], [15]. Narita [12] found that the Modified double-stimulus impairment scale (EBU) method has higher discrimination ability than the EBU and DSCQS methods when the impairment range is limited to the high-quality range. Kawano et al. [11] found that the DSCQS method has higher discrimination ability than the ACR and DCR methods; otherwise, the DCR method performed the best at evaluating 3D videos. Lee et al. [15] found that the PC method has higher discrimination ability than the single stimulus continuous quality scale (SSCQE) method.

Assessment time has also been investigated [10], [11], [13], [14]. Tominaga et al. [10] showed that the ACR, ACR-HR, and ACR with an 11-grade scale were the most efficient subjective quality assessment methods. Kawano et al. [11] showed that the ACR method is the most efficient based on an investigation of the correlation between subjective quality assessment methods. Huynh-Thu and Ghanbari [13] showed that the ACR-HR method is more efficient than the DSCQS method because, although there is a high correlation between both methods, the assessment time of the ACR-HR method is a quarter of that of the DSCQS method. Pinson and Wolf [14] proposed SSCQE with hidden reference removal and showed its advantages in terms of a shorter assessment time than the DSCQS and double stimulus comparison scale methods. As described above, the performances of these subjective quality assessment methods have been sufficiently investigated and compared.

2.2 Subject Bias

The effect of subject bias has also been sufficiently investigated [16], [17]. Janowski and Pinson [16] found that subjective ratings are affected by subject bias, subject inaccuracy, and stimulus scoring difficulty when using the ACR method. Subject inaccuracy means subject error. They [16] thus proposed a "theoretical subject model" based on these subjective rating behaviors. Lee [17] studied the PC method because it does not have well established guidelines unlike the single or double stimulus method. For example, the number of subjects for obtaining a reliable subjective score has not been determined. Therefore, Lee [17] proposed models based on the relationship between the number of subjects, the number of stimuli to be compared, and the convergence of the quality levels estimated from the paired comparison results via simulation. Using these models, experimenters can determine the number of subjects and the number of stimuli and can obtain reliable subjective data that excludes the effect of subject bias. As described above, the effect of subject bias has been sufficiently investigated.

2.3 Range Effects

The range effects have not been sufficiently investigated or compared [12], [18], [19]. Some studies have focused on the range effects with the DSCQS method in terms of accuracy, but few studies have focused on range effects in terms of stability and discrimination ability.

Speranza et al. [18] compared the DSCQS and PC methods by analyzing their results in terms of accuracy. They found a marginal range effect with the DSCQS method when a subjective assessment experiment was conducted with non-expert viewers with only high-quality assessment videos and a very small range effect for the PC method with expert

viewers. Narita [12] compared the Modified EBU with the EBU and DSCQS methods in full and limited impairment ranges using only high-quality videos. He analyzed his results in terms of accuracy and discrimination ability and found that the scores for sequences obtained using the DSCQS method did not disperse when the impairment range was limited, whereas the scores obtained using the Modified EBU and EBU methods dispersed. Therefore, the Modified EBU method has higher discrimination ability than the DSCQS method. We [19] examined the range effects for the DSCQS method by analyzing the results in terms of accuracy and stability. We found that DSCQS values from a high-quality experiment were almost the same as those from a low-to-high-quality experiment, whereas the DSCQS values of some low-quality videos were affected. Among the previous studies, many focused on the range effects of DSCQS methods. However, although almost all of them focused on range effects in terms of accuracy, few focused on range effects in terms of stability and discrimination ability. Therefore, range effects with the DSCQS method need to be investigated in terms of stability and discrimination ability.

There are few studies on the ACR-HR method even though it is considered to minimize range effects [13]. Huynh-Thu and Ghanbari [13] compared the DSCQS method with the ACR-HR method in an experiment with assessment videos in the low-quality range. They analyzed their results in terms of accuracy and found that the correlation between the DSCQS and ACR-HR methods was high. They focused on assessment videos in the low-quality range because they aimed to evaluate the quality of video sequences transmitted over 3G mobile networks. High-quality 4k video broadcasting and IPTV services have recently started [1], [2], but there are no studies focused on range effects in only the high-quality range. Therefore, to investigate range effects in the high-quality range, the DSCQS and ACR-HR methods need to be compared in terms of accuracy, stability, and discrimination ability.

As described above, there are some issues that need to be addressed. First, range effects with the DSCQS method need to be investigated in terms of stability and discrimination ability. Second, the range effects on the DSCQS and ACR-HR methods in the high-quality range have not been clarified. Therefore, to clarify them, subjective assessment tests need to be conducted using high-quality assessment videos and low-to-high-quality assessment videos for the DSCQS and ACR-HR methods. In addition, both methods need to be compared in terms of accuracy, stability, and discrimination ability.

3. Subjective Quality Assessment Methods

In this section, we explain the procedures of the DSCQS and ACR-HR methods.

3.1 DSCQS Method

In the DSCQS method, a pair of a reference video and a test/assessment video with artifacts such as video coding is presented, as shown in Fig. 1. These videos are presented twice, and assessment is performed when the second videos are presented. The videos are presented in random order and the participants are not told which the reference video is. The participants assess the videos on a continuous quality assessment scale on the basis of five grades, as shown in Fig. 2. The assessment scale is normalized to the range 0–100 (maximum value: 100, minimum value: 0), and the difference in video assessment values, called "video quality differential values," is calculated for the reference and assessment videos in each pair. These video quality differential values are averaged across all the participants to yield a DSCQS value. These values were calculated as follows.

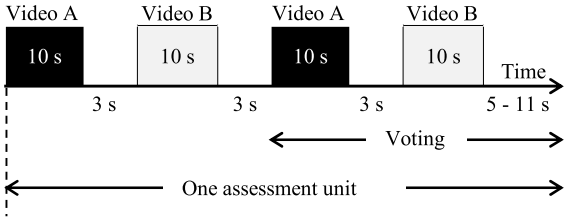

Fig. 1 Flow of DSCQS method.

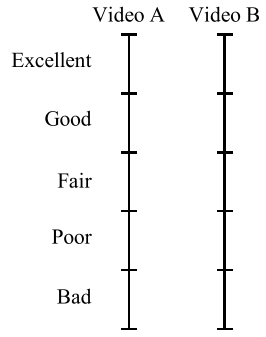

Fig. 2 Rating scale with DSCQS method.

$$OS(i,j) = OS_{Ref}(i,j) - OS_{Test}(i,j), \tag{1}$$

$$DSCQS(i) = \frac{1}{J} \times \sum_{j=1}^{J} OS(i,j), \tag{2}$$

where i is the index of the processed video sequence (PVS), j is the index of the participant, J is the number of participants, $OS(i,j)$ is the video quality differential value, $OS_{Ref}(i,j)$ is the opinion score of the reference video, $OS_{Test}(i,j)$ is the opinion score of the assessment video, and $DSCQS(i)$ is the average of the opinion scores of all participants. Because the DSCQS value is calculated from the difference in video quality, a smaller value indicates a higher quality (closer to the reference video), and a larger value indicates a lower quality. To obtain reliable data, participants are screened after subjective quality data is collected. The method for screening participants has been standardized in ITU-R Recommendation BT.500 [5].

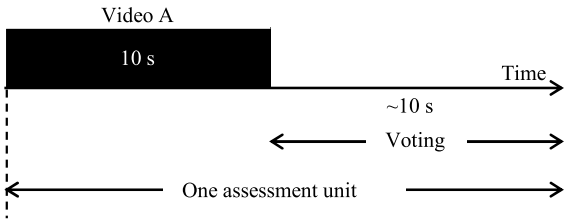

Fig. 3 Flow of ACR-HR method.

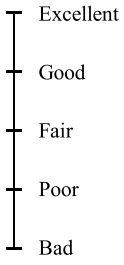

Fig. 4 Rating scale with ACR-HR method.

3.2 ACR-HR Method

In the ACR-HR method, test videos including reference videos are presented one at a time, as shown in Fig. 3, and are rated independently on the category scale shown in Fig. 4. Video quality is assessed by scoring it on the basis of a discrete scale of five categories (5: Excellent, 4: Good, 3: Fair, 2: Poor, 1: Bad). Then, the video quality differential values are averaged across all the participants to yield an ACR-HR value. These values were calculated as follows.

$$OS(i,j) = OS_{Test}(i,j) - OS_{Ref}(i,j) + 5, \quad (3)$$

$$ACR\text{-}HR(i) = \frac{1}{J} \times \sum_{j=1}^{J} OS(i,j), \quad (4)$$

where i is the index of the PVS, j is the index of the participant, J is the number of participants, $OS(i,j)$ is the video quality differential value, $OS_{Ref}(i,j)$ is the opinion score of the reference video, $OS_{Test}(i,j)$ is the opinion score of the assessment video, and $ACR\text{-}HR(i)$ is the average of the opinion scores of all participants. When the assessment video has a higher score than the reference video, the $OS(i,j)$ is greater than 5. In accordance with ITU-T Recommendation P.910 [4], to prevent $OS(i,j)$ that is greater than 5 from unduly affecting $ACR\text{-}HR(i)$, we calculated $clipped\_OS$ as:

$$clipped\_OS(i,j) = (7 \times OS(i,j))/(2 + OS(i,j)). \quad (5)$$

To obtain reliable data, participants are screened on the basis of ITU-R Recommendation BT.500 [5] in the same way as the DSCQS method, because there is no standardized screening method for the ACR-HR method.

4. Subjective Tests

As described in Sects. 1 and 2, to investigate the range effects in the high-quality experiment for the DSCQS and ACR-HR methods, we conducted four subjective assessment tests, listed in Table 1. We prepared two video sets (video sets 1 and 2). Video set 1 included low-to-high-quality assessment videos, and video set 2 included only high-quality assessment videos. The details of these video sets are described in Sect. 4.2.

Table 1 Subjective assessment experiment.

Experiment	Video set	Subjective assessment method
1	1	DSCQS method
2	2	DSCQS method
3	1	ACR-HR method
4	2	ACR-HR method

Table 2 Source reference circuits (SRCs).

Number	Title	Source
SRC01	Tokyo Skytree	AQUA Geo Graphic
SRC02	Falls	AQUA Geo Graphic
SRC03	Horse	AQUA Geo Graphic
SRC04	Tropical fish	ASTRODESIGN
SRC05	Flower	ASTRODESIGN
SRC06	Nagigama Forest	NHK
SRC07	Grove with snow	NHK
SRC08	Fallen tree	NHK
SRC09	Onbashira Festival	NHK
SRC10	Festival crowd	NHK
SRC11	Train	NHK
SRC12	Swimming	NHK
SRC13	Nebuta Festival	NHK
SRC14	Rice fields	NHK
SRC15	Children in rice fields	NHK
SRC16	Playing in water	NHK

Table 3 Video conditions.

Parameter	Value
Video format	3840 × 2160p
Frame rate	60 fps
YUV format	4:2:0
Bit depth	8 bits

4.1 Source Reference Circuits

In all experiments, we used the 16 4k source reference circuits (SRCs) listed in Table 2. These videos lasted 10 seconds. Since the original video format of SRCs 06-16 was 8k (7680 × 4320) at 60p, we converted it to 4k, i.e., 3840 × 2160 at 60p, using a bicubic method. The 4k format is listed in Table 3. The video SRCs in Table 2 were selected because they are distributed widely without overlapping in terms of temporal perceptual information measurement (TI) or spatial perceptual information measurement (SI). ITU-T Recommendation P.910 [4] defines TI as the maximum standard deviation of the motion difference feature and SI as the maximum standard deviation of the pixels in each Sobel-filtered frame. However, we defined TI and SI using the average value instead of the maximum value because the maximum value is not suitable for determining the charac-

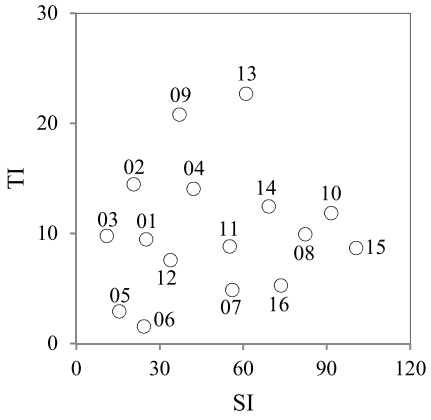

Fig. 5 Distributions of TI and SI.

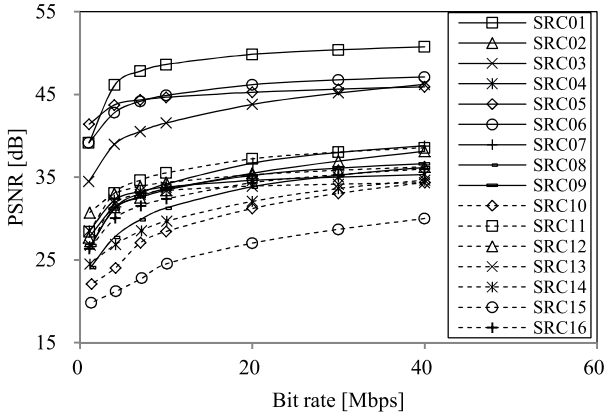

Fig. 6 Relationships between PSNR and bit rate.

teristics of the entire scene. The distributions of TI and SI of these 4k videos are shown in Fig. 5, where each number indicates the index of the source video listed in Table 2. In addition, to demonstrate the content dependency on the encoding difficulty, Fig. 6 shows the relationship between peak signal to ratio (PSNR) and bit rate per SRC. As shown in Fig. 6, PSNR values were distributed from low to high at a certain bit rate. This means that the SRCs are distributed widely in terms of encoding difficulty.

4.2 Hypothetical Reference Circuits

We used eight hypothetical reference circuits (HRCs) including seven compressed conditions and uncompressed condition for video set 1 and three HRCs including two compressed conditions and uncompressed condition for video set 2. The bit rates of compressed conditions of video sets 1 and 2 are listed in Table 4. For video set 1, we chose low-to-high-quality bit rates in order to use different 4k video qualities. We chose the lowest bit rate for video set 2 in accordance with the assessment results of video set 1 to use assessment videos in the high-quality range. To set the quality range to be similar to that of Narita [12], we defined the lowest bit rate as 30 Mbps because the DSCQS values of 30 Mbps were lower than 25 in the results for video set 1. We selected bit rates of

Table 4 Bit rate conditions.

Video set	Bit rate [Mbps]
1	1, 4, 7, 10, 20, 30, 40
2	30, 40

Table 5 Codec settings.

Parameter	Value
CU size/depth	64/4
Motion search range	64
Intra period	32
GOP size	8

Table 6 Specifications of display.

Parameter	Value
Panel size	56 inches
Resolution	3840 × 2160
Backlight	CCFL
Brightness	350 cd/m2
Contrast	950:1
Color depth	10 bits

30 and 40 Mbps as the high-quality video set (video set 2) in order to compare the discrimination abilities of the DSCQS and ACR-HR methods in the high-quality range. We selected these bit rates on the basis of the hypothesis that participants can distinguish the difference in quality between videos encoded at 30 and 40 Mbps when using the DSCQS method but cannot when using the ACR-HR method. The compressed videos were encoded in H.265/HEVC (HMv11.0) [20]. The codec and coding parameters are listed in Table 5. In addition, condition was used in our experiment. The reason we used the conditions in Table 5 is that these settings are usually used for video services over IPTV using H.265/HEVC [21] and broadcasting using H.265/HEVC [22].

4.3 Environmental Conditions

The 4k monitor we used was DM-3410-A produced by Astrodesign. The display specifications are listed in Table 6. We used the same 4k monitor in all the experiments. Two participants viewed a single monitor simultaneously. They were centered evenly in front of the monitor and viewed each video at a distance of 1.5 H (1.5 H is about 105 cm; H indicates the picture height) from the monitor [23]. The room luminance was 200 lux. These settings were defined on the basis of the subjective video-quality assessment experiments in ITU-R Report BT.2246 [23] and ITU-R Recommendation BT.500 [5].

4.4 Flow of the Experiments

The DSCQS method was used in experiments 1 and 2, and the ACR-HR method was used in experiments 3 and 4. The procedures of the DSCQS and ACR-HR methods are described in Sect. 3. In each experiment, 32 participants were divided into 8 groups. Each group was divided into two sub-groups, groups A and B. Both groups performed the

assessment tests alternately. At the beginning of each experiment, participants performed a practice session to become accustomed to evaluation with the subjective assessment. In the practice sessions of all experiments, we performed an instruction session and a demonstration session in accordance with ITU-R Recommendation BT.500 [5] for the DSCQS method and ITU-T Recommendation P.910 [4] for the ACR-HR method. In the instruction session, we explained the details of the experiment using an instruction sheet in order for participants to receive exactly the same information. The instruction sheet explained what participants were going to watch, voting procedures, kinds of impairments, and rating scales. We showed examples of impairments such as blur, distortion, shaky movements, and noise. In addition, we showed that a clear video without impairment is sometimes presented. After the instruction session, we performed the demonstration session. In the demonstration session, participants experienced both voting procedures and kinds of impairments. At the end of the practice session, if participants had questions about voting procedures and the rating scales, the experimenter answered them. After the practice session, participants performed eight evaluation sessions in experiments 1 and 3 and three evaluation sessions in experiments 2 and 4. Finally, participants evaluated 128 PVSs in experiments 1 and 3 and 48 PVSs in experiments 2 and 4. Therefore, in each session, participants evaluated 16 PVSs.

4.5 Participants

For all experiments, we screened the participants for visual acuity and color vision. In total, 128 participants who passed the screening tests participated in our experiments. Thirty-two participants (16 males, 16 females) ranging from 20 to 29 years old took part in each experiment.

5. Experimental Results

In this section, we compare the results for the DSCQS and ACR-HR methods from the viewpoints of accuracy, stability, and discrimination ability. In this paper, accuracy represents the proximity of results obtained from the high-quality experiment to the results obtained from the low-to-high-quality experiment. Stability represents the dispersion of individual ratings, and discrimination ability means the ability to discriminate one PVS from other PVSs. We defined $DSCQS1(i)$ as the DSCQS values of experiment 1 using video set 1, $DSCQS2(i)$ as the DSCQS values of experiment 2 using video set 2, $ACR\text{-}HR1(i)$ as the ACR-HR values of experiment 3 using video set 1, and $ACR\text{-}HR2(i)$ as the ACR-HR values of experiment 4 using video set 2. Here, i means the index of the PVS.

First, to obtain reliable data, we screened participants on the basis of ITU-R Recommendation BT.500 [5]. The number of screened participants is shown in Table 7. In this study, we used the *clipped_OS* as the ACR-HR values. In our experiments, especially those using video set 2, participants evaluated the assessment videos to have similar quality in the high-quality range. Therefore, the score of an assessment video was sometimes higher than that of the reference video. However, there was no significant difference between $ACR\text{-}HR(i)$ using $OS(i, j)$ and $ACR\text{-}HR(i)$ using *clipped_OS* with a significance level of 5% for any PVSs.

Here, we show the quality range of assessment videos in video sets 1 and 2. Figures 7 and 8 show the average assessment results of the 16 4k videos per bit rate to show the characteristics of each experiment. The error bars in these figures mean the 95% CIs. In particular, the ranges of the DSCQS and ACR-HR values from video set 1 were larger than those from video set 2. This means that we were able to set the quality range of the assessment videos in video sets 1 and 2 as expected. In addition, to verify the correctness of our hypothesis that participants can distinguish the small difference in quality when using the DSCQS method but cannot when using the ACR-HR method, we performed the Student's t-test with a significance level of 5% between the $DSCQS2$ of 30 Mbps and $DSCQS2$ of 40 Mbps for each SRC. There was no significant difference between them in

Table 7 Number of screened participants.

Experiment	1	2	3	4
Number of screened participants	2	3	1	1

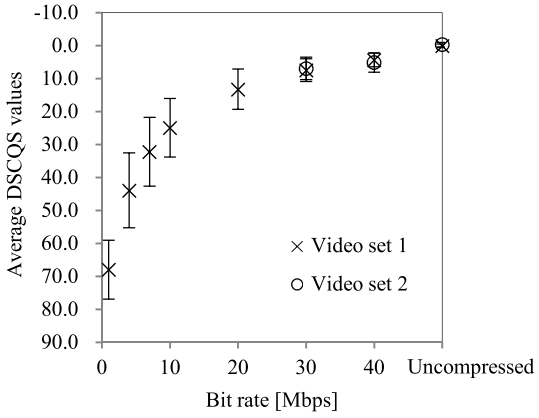

Fig. 7 Average DSCQS values from video sets 1 and 2.

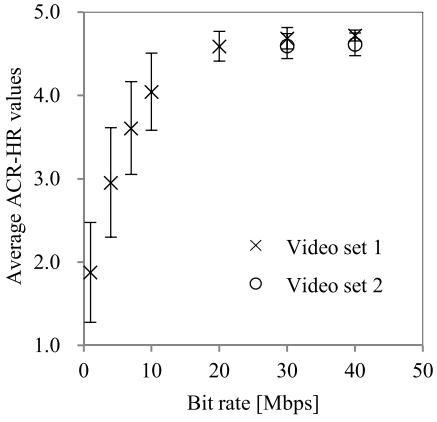

Fig. 8 Average ACR-HR values from video sets 1 and 2.

12 of 16 SRCs. Similarly, we performed the Student's t-test with a significance level of 5% between $ACR\text{-}HR2$ of 30 Mbps and $ACR\text{-}HR2$ of 40 Mbps for each SRC. There was no significant difference between them for any of the 16 SRCs. This means that participants can distinguish the difference in quality between videos encoded at 30 and 40 Mbps when using the DSCQS method but cannot when using the ACR-HR method. Therefore, we were able to select the bit rate of video set 2 as we expected. As shown in Eq. (1), the DSCQS values of uncompressed videos are not always 0 because uncompressed videos are assessed as both reference and assessment videos. However, as shown in Eq. (3), the ACR-HR values of uncompressed videos are always 5 because the reference video is evaluated once. Therefore, uncompressed videos are analyzed with the DSCQS method but not with the ACR-HR method.

5.1 Accuracy

To investigate the accuracy between the subjective assessment values of the different experiments, we compared 48 PVSs that had the same common conditions (three conditions: uncompressed, 30 Mbps, and 40 Mbps of 16 videos) for $DSCQS1$ and $DSCQS2$, and 32 PVSs that had the same common conditions (two conditions: 30 and 40 Mbps of 16 videos) for $ACR\text{-}HR1$ and $ACR\text{-}HR2$. Figure 9 shows the relationship between $DSCQS1$ and $DSCQS2$. Figure 10 shows the relationship between $ACR\text{-}HR1$ and $ACR\text{-}HR2$. There was a high correlation between $DSCQS1$ and $DSCQS2$ and between $ACR\text{-}HR1$ and $ACR\text{-}HR2$. We calculated the PCC and SRCC. The former is an index that shows the correspondence-related strength of DSCQS or ACR-HR values, and the latter is a measure that shows the correlation degree of rank for both values. The formulas for PCC and SRCC are as follows.

$$PCC = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\overline{y})^2}}, \quad (6)$$

$$SRCC = 1 - \frac{6\sum_{i=1}^{n}(x_i-y_i)^2}{n^3-n}, \quad (7)$$

where x_i is $DSCQS1(i)$, y_i is $DSCQS2(i)$, and n is the number of common PVSs, i.e., 48 in this paper. Similarly, x_i is $ACR\text{-}HR1(i)$, y_i is $ACR\text{-}HR2(i)$, and n is the number of common PVSs, i.e., 32 in this paper. By using Eqs. (6) and (7), we calculated the PCC and SRCC of the DSCQS and ACR-HR values, given in Table 8. The table shows that there was a high correlation between $DSCQS1$ and $DSCQS2$ and between $ACR\text{-}HR1$ and $ACR\text{-}HR2$.

To investigate the relationship between encoding difficulty and PCC/SRCC, we divided SRCs into two groups, i.e., groups 1 and 2, on the basis of the average PSNR among seven bit rates for video set 1 described in Table 4. Group 1 contains eight SRCs, i.e., 01-03, 05-07, and 11-12. Their average PSNR is relatively high. Group 2 contains eight SRCs, i.e., SRCs 04, 08-10, and 13-15. Their average PSNR is relatively low. Table 9 shows PCC and SRCC in the common conditions for video sets 1 and 2. Table 9 shows that PCC/SRCC of group 2 was larger than that of group 1. Therefore, it is considered that videos that are difficult to encode contribute to improve the performance in PCC/SRCC.

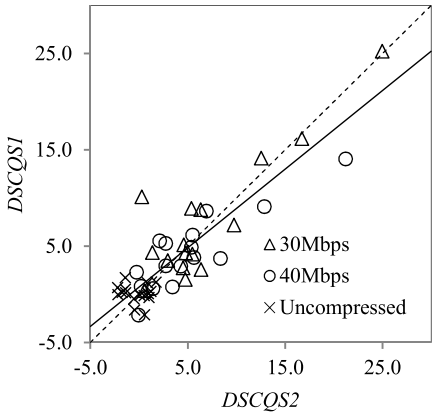

Fig. 9 Correlation analysis between $DSCQS1$ and $DSCQS2$.

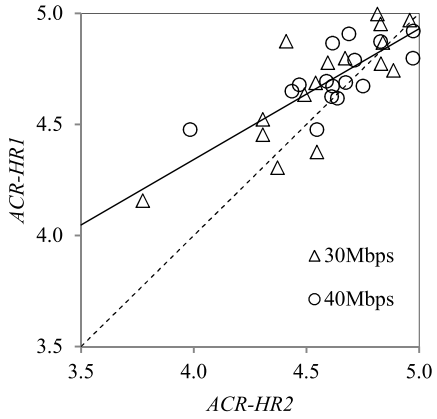

Fig. 10 Correlation analysis between $ACR\text{-}HR1$ and $ACR\text{-}HR2$.

Second, we calculated the slope and intercept of the approximate lines of Figs. 9 and 10, given in Table 10. The table shows that the values of the slopes of the DSCQS and ACR-HR were under 1.0 and that those of the intercept were not 0. In addition, as shown in Table 10, 95% of the CIs of the slope of both methods were under 1.0, and 95% of the CIs of the intercept of both methods were over 0. This means that the DSCQS and ACR-HR values were slightly affected by the range effects. Third, we performed the Student's t-test with a significance level of 5% between $DSCQS1(j)$ and $DSCQS2(j)$ and between $ACR\text{-}HR1(j)$ and $ACR\text{-}HR2(j)$, where j means the index of common PVS. From the results, 2 of the common 48 PVSs had significant differences between $DSCQS1(j)$ and $DSCQS2(j)$, and 2 of the common 32 PVSs had a significant difference between $ACR\text{-}HR1(j)$ and $ACR\text{-}HR2(j)$. This means that $DSCQS1(j)$ and $DSCQS2(j)$ were almost

Table 8 Correlation analysis between video sets 1 and 2.

	DSCQS	ACR-HR
PCC	0.88	0.78
SRCC	0.78	0.73

Table 9 Comparison in PCC/SRCC between groups 1 and 2.

	DSCQS		ACR-HR	
	Group 1	Group 2	Group 1	Group 2
PCC	0.79	0.88	0.67	0.78
SRCC	0.72	0.78	0.54	0.74

Table 10 Regression analysis between video sets 1 and 2.

	DSCQS		ACR-HR	
	Average	95% of CIs	Average	95% of CIs
Slope	0.82	0.13	0.59	0.17
Intercept	0.76	0.89	1.99	0.81

the same and that $ACR\text{-}HR1(j)$ and $ACR\text{-}HR2(j)$ were almost the same. As described above, the DSCQS and ACR-HR values were marginally affected by range effects, but the differences between $DSCQS1(j)$ and $DSCQS2(j)$ and between $ACR\text{-}HR1(j)$ and $ACR\text{-}HR2(j)$ were within the range of the confidence intervals.

5.2 Stability

To investigate the range effects in terms of the stability of DSCQS and ACR-HR values, we compared the 95% of the CIs since the CI values are based on the standard deviation of the DSCQS and ACR-HR values. To compare the stabilities of the values, we normalized the values as follows because ACR-HR values range from 1 to 5 and DSCQS values range from 0 to 100.

$$nDSCQS(i) = \frac{DS(i) - \max\{DS(i)\}}{\min\{DS(i)\} - \max\{DS(i)\}}, \quad (8)$$

$$nACR\text{-}HR(i) = \frac{AC(i) - \min\{AC(i)\}}{\max\{AC(i)\} - \min\{AC(i)\}}, \quad (9)$$

where $DS(i)$ means $DSCQS(i)$ and $AC(i)$ means $ACR\text{-}HR(i)$. For $nDSCQS(i)$ and $nACR\text{-}HR(i)$, 0 means the lowest quality and 1 means the highest quality. Figure 11 shows the normalized 95% of the CIs of $DSCQS1$ and $ACR\text{-}HR1$. We compared 112 PVSs (7 encoded conditions of 16 videos) for $DSCQS1$ and $ACR\text{-}HR1$. Figure 12 shows the normalized 95% of the CIs of $DSCQS2$ and $ACR\text{-}HR2$. We compared 32 PVSs (2 encoded conditions of 16 videos) for $DSCQS2$ and $ACR\text{-}HR2$. As shown in Fig. 11, the normalized 95% of the CIs of $DSCQS1$ and $ACR\text{-}HR1$ mostly overlapped. As shown in Fig. 12, those of $DSCQS2$ were smaller than those of $ACR\text{-}HR2$. In addition, we performed the Student's t-test with a significance level of 5% between 95% of the CIs of $DSCQS1$ and those of $ACR\text{-}HR1$ and between those of $DSCQS2$ and $ACR\text{-}HR2$. There was no significant difference between 95% of the CIs of $DSCQS1$ and those of $ACR\text{-}HR1$ but there was between those of $DSCQS2$ and $ACR\text{-}HR2$. This means that the stabilities of the DSCQS and ACR-HR methods were almost the same for the experiments using video set 1, which included low-to-high-quality assessment videos. In contrast, the stabilities of the DSCQS and ACR-HR were significantly different for the experiments using video set 2, which included only high-quality assessment videos. That is, the DSCQS method was more stable than the ACR-HR method for the high-quality experiment.

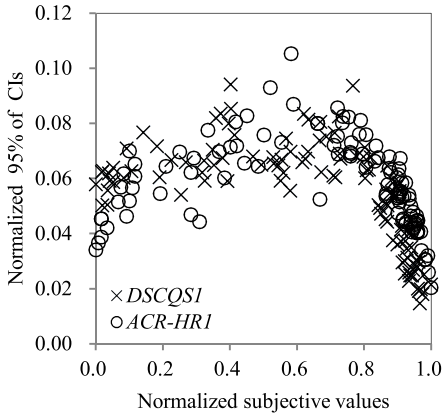

Fig. 11 Comparison of normalized 95% of CIs of $DSCQS1$ and $ACR\text{-}HR1$.

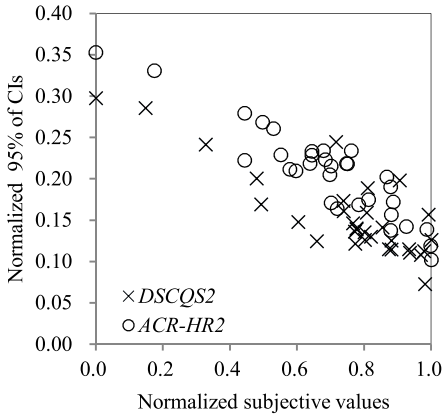

Fig. 12 Comparison of normalized 95% of CIs of $DSCQS2$ and $ACR\text{-}HR2$.

5.3 Discrimination Ability

To investigate the range effects in terms of the discrimination ability of the DSCQS and ACR-HR values, we performed the Student's t-test with a significance level of 5% among the subjective values of each method and each experiment. In the same way as Kawano et al. [11], we calculated the number of significant differences (N) between $DSCQS1(k)$ and $DSCQS1(l)$ ($k = 1, 2, \cdots, I, l \neq k$), where k and l represent the index of the PVS and I is 112 PVSs (7 encoded conditions of 16 videos) for the experiments using video set 1 and 32 PVSs (2 encoded conditions of 16 videos) for the experiments using video set 2. If the N of a certain subjective quality assessment method is higher than that of the others,

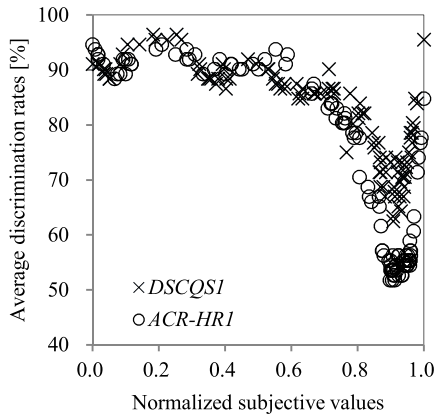

Fig. 13 Comparison of average discrimination rate of $DSCQS1$ and $ACR\text{-}HR1$.

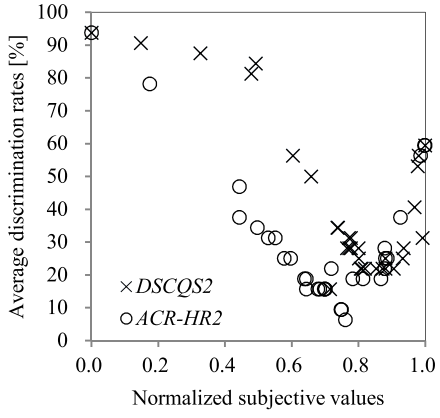

Fig. 14 Comparison of average discrimination rate of $DSCQS2$ and $ACR\text{-}HR2$.

the method with a higher N has a greater ability to identify differences in video quality. Similarly, we calculated the N between $DSCQS2(k)$ and $DSCQS2(l)$, $ACR\text{-}HR1(k)$ and $ACR\text{-}HR1(l)$, and $ACR\text{-}HR2(k)$ and $ACR\text{-}HR2(l)$. Figures 13 and 14 show the discrimination rate (N/I).

The discrimination rate shows a convex downward trend in Figs. 13 and 14. The highest-quality PVS is compared with lower-quality PVSs and vice versa. However, a middle-quality PVS is compared with both lower- and higher-quality PVSs. Therefore, the discrimination rates of the highest- and lowest-quality PVSs are higher than that of a middle-quality PVS, and the discrimination rate shows a convex downward trend.

Here, we explain this convex downward trend using $DSCQS2(k)$ of Fig. 14. First, we compared the difference in quality between the highest-quality $DSCQS2(1)$ (k = 1, k is the index of the PVS) and the second-highest-quality $DSCQS2(2)$, then compared the difference in quality between $DSCQS2(1)$ and the third-highest-quality $DSCQS2(3)$, and similarly compared the difference in quality between $DSCQS2(1)$ and other $DSCQS2(k)$ (k = 4, 5, ..., and 32) in turn. In this case, the difference in quality gradually increases. For example, when there is a difference in quality between $DSCQS2(1)$ and $DSCQS2(14)$ in the Student's t-test with a significance level of 5%, there is also a difference in quality between $DSCQS2(1)$ and $DSCQS2(k)$ (k = 15, 16, ..., and 32). In this case, there are 19 significant differences (N) between $DSCQS2(1)$ and $DSCQS2(k)$.

Table 11 p-values of comparison between DSCQS and ACR-HR methods for stability and discrimination ability.

	Stability	Discrimination ability
Video set 1	2.4×10^{-1}	4.1×10^{-7}
Video set 2	2.4×10^{-4}	5.1×10^{-2}

Second, when we compared the difference in quality between the lowest-quality $DSCQS2(32)$ and other $DSCQS2(k)$ (k=31, 30, ..., and 1) in turn, the difference in quality increases gradually. For example, when there is a difference in quality between $DSCQS2(32)$ and $DSCQS2(30)$ in the Student's t-test with a significance level of 5%, there is also a difference in quality between $DSCQS2(32)$ and $DSCQS2(k)$ (k = 29, 28, ..., and 1). In this case, N for $DSCQS2(32)$ is 30.

On the other hand, we compared the difference between the middle-quality $DSCQS2(16)$ and the other $DSCQS2(k)$ (k = 15, 14, ..., and 1), and similarly we compared the difference between the middle-quality $DSCQS2(16)$ and the other $DSCQS2(k)$ (k = 17, 18, ..., and 32). In both cases, the difference in quality also increases gradually. When there is a difference in quality between $DSCQS2(16)$ and $DSCQS2(3)$, there is a difference between $DSCQS2(16)$ and $DSCQS2(1)$. When there is a difference in quality between $DSCQS2(16)$ and $DSCQS2(27)$, there is a difference between $DSCQS2(16)$ and $DSCQS2(k)$ (k = 28, 29, ..., and 32). In this case, N of $DSCQS2(16)$ is 8. Similarly, we calculated N for the other $DSCQS2(k)$.

From these investigations, N is relatively high in the highest- and lowest-quality $DSCQS2$ but relatively low in the middle-quality $DSCQS2$. Therefore, the discrimination rate shows a convex downward trend.

As shown in Fig. 13, the discrimination rates of $DSCQS1$ were larger than those of $ACR\text{-}HR1$. In contrast, as shown in Fig. 14, the rates of $DSCQS2$ and $ACR\text{-}HR2$ appeared to be almost the same. In addition, we performed the Student's t-test with a significance level of 5% between the discrimination rates of $DSCQS1$ and $ACR\text{-}HR1$ and between those of $DSCQS2$ and $ACR\text{-}HR2$. There was a significant difference between the rates of $DSCQS1$ and $ACR\text{-}HR1$ but not between those of $DSCQS2$ and $ACR\text{-}HR2$. The p-value was 0.051 when we performed the t-test between the rates of $DSCQS2$ and $ACR\text{-}HR2$. This means that the DSCQS method has higher discrimination ability than the ACR-HR method for the experiments using video set 1, which included low-to-high-quality assessment videos. In contrast, the DSCQS and ACR-HR methods did not significantly differ for the experiments using video set 2, which included only high-quality assessment videos. That is, the discrimination abilities of both methods were almost the same for the high-quality experiments.

6. Discussion

As described in Sect. 5, we found that the DSCQS and ACR-HR values are marginally affected by range effects, although almost all of the common PVSs for high- and low-to-high-quality experiments have no significant differences. In addition, Table 11 summarizes the results of the comparison of the DSCQS and ACR-HR methods. The table also shows the p-values of the comparison of both methods with the Student's t-test in terms of stability and discrimination ability. In terms of stability, the DSCQS method was as stable as the ACR-HR method for video set 1 but more stable for video set 2, which included only high-quality videos. In terms of discrimination ability, the DSCQS method had higher discrimination ability than the ACR-HR method for video set 1, whereas both methods had almost the same discrimination ability for video set 2.

Here, we discuss the reason that both methods have almost the same discrimination ability despite the DSCQS method being more stable than the ACR-HR method for video set 2 with high-quality videos. To achieve high discrimination ability, two points need to be satisfied. One is that the 95% of the CIs of $MOS(k)$ and those of $MOS(l)$ ($k = 1, 2, \cdots, I, l \neq k$, where k and l represent the index of the PVS and I is the number of PVSs) are small. Here, the $MOS(k)$ and $MOS(l)$ can be replaced with $DSCQS1$, $DSCQS2$, $ACR\text{-}HR1$, or $ACR\text{-}HR2$. The other is that the difference between $MOS(k)$ and $MOS(l)$ is slightly large even if 95% of the CIs of $MOS(k)$ and $MOS(l)$ are not small. This means that subjective values disperse. Therefore, Fig. 15 shows the cumulative distribution of $DSCQS1$ and $ACR\text{-}HR1$, and Fig. 16 shows that of $DSCQS2$ and $ACR\text{-}HR2$. For video set 1, the DSCQS and ACR-HR methods were almost equally stable. As shown in Fig. 15, $DSCQS1$ dispersed more widely than $ACR\text{-}HR1$. Therefore, we consider the DSCQS method to have higher discrimination ability than the ACR-HR method for video set 1. For video set 2, shown in Fig. 16, $ACR\text{-}HR2$ dispersed more widely than $DSCQS2$. Therefore, both methods have almost the same discrimination ability even though the DSCQS method is more stable than the ACR-HR method. As shown in the results for accuracy, the DSCQS and ACR-HR values were marginally affected by range effects, but there was no significant difference between $DSCQS1(j)$ and $DSCQS2(j)$ in 46 of the 48 PVSs or between $ACR\text{-}HR1(j)$ and $ACR\text{-}HR2(j)$ in 30 of the 32 PVSs, where j means the index of common PVS. In addition, as shown in Table 10, the intercept of the ACR-HR method was larger than that of the DSCQS method, considering the ACR-HR method is able to take a value from 1 to 5 and the DSCQS method is able to take a value from 0 to 100. Therefore, it is considered that the values of the ACR-HR method are likely to disperse more widely than those of the DSCQS method. That the values of DSCQS methods do not tend to disperse is consistent with the results of Narita [12]. In his work [12], the scores obtained using this method did not disperse when the impairment range was limited,

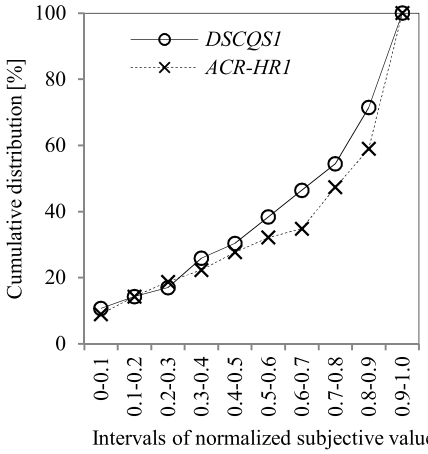

Fig. 15 Comparison of cumulative distribution of $DSCQS1$ and $ACR\text{-}HR1$.

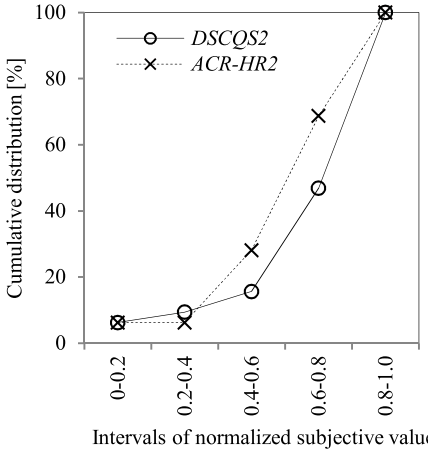

Fig. 16 Comparison of cumulative distribution of $DSCQS2$ and $ACR\text{-}HR2$.

Table 12 Correlation analysis between DSCQS and ACR-HR methods.

	PCC	SRCC
Video set 1	-0.98	-0.95
Video set 2	-0.83	-0.70

whereas the scores obtained using the Modified EBU and EBU methods did. Therefore, the discrimination ability of the Modified EBU is higher than that of the DSCQS method.

Then, we compared our results with those of Huynh-Thu and Ghanbari [13], who compared the DSCQS and ACR-HR methods in an experiment with low-quality assessment videos. They found that the correlation between both methods was high. In our study, we compared both methods in experiments with high-quality assessment videos and low-to-high-quality assessment videos. We calculated the PCC and SRCC between the methods using Eqs. (6) and (7) in Table 12. As shown in Table 12, there are high correlations between both methods in both video sets. This result is consistent with that of Huynh-Thu and Ghanbari [13] who used only low-quality videos. However, the correlation between

Table 13 Comparison of numbers of pairs that show significant difference for each SRC.

SRC	video set 1	video set 2
01	0	2
02	0	2
03	1	1
04	2	2
05	0	1
06	0	0
07	1	1
08	2	2
09	3	2
10	2	2
11	0	1
12	2	2
13	1	0
14	2	3
15	3	2
16	2	1
Total	21	24

Table 14 Comparison of average numbers of pairs that show significant difference for each group.

Group	video set 1	video set 2
1	0.5	1.3
2	2.1	1.8

both methods for video set 1 was higher than that for video set 2. This result is considered to be caused by the fact that the values of the ACR-HR method are likely to disperse more widely than those of the DSCQS method in the high-quality experiment as described above.

Finally, we compared our results using 4k videos with those of Narita [12] using HD videos in the discrimination ability for the DSCQS method. Narita [12] compared the discrimination ability for high- and low-to-high-quality experiments. He shows that there was no significant difference in discrimination ability between these two experiments. In our experiment, to compare the discrimination ability of the DSCQS method for video sets 1 and 2, we performed the Student's t-test with a significance level of 5% between three common pairs on video sets 1 and 2 (i.e., $DSCQSi$ of the uncompressed and $DSCQSi$ of 40 Mbps, $DSCQSi$ of the uncompressed and $DSCQSi$ of 30 Mbps, and $DSCQSi$ of 40 Mbps and $DSCQSi$ of 30 Mbps (i = 1, 2)) for each SRC in the same way as Narita [12]. Table 13 shows the numbers of pairs that show a significant difference with a significance level of 5% for each SRC. The table also shows that the total number of pairs for video set 2 is larger than that for video set 1. In addition, Table 14 summarizes the comparison of discrimination ability in terms of the encoding difficulty. We divided SRCs into two groups, i.e., groups 1 and 2, as described in Sect. 5. The table also compares the discrimination ability between video sets 1 and 2 for each group. The table shows that the average number of pairs that shows a significant difference with a significance level of 5% for video set 2 is larger than that for video set 1 for group 1. On the other hand, the average number of pairs that shows a significant difference with a significance level of 5% for video set 1 is larger than that for video set 2 for group 2. Then, we performed the Student's t-test with a significance level of 5% between video sets 1 and 2 for each group. There was a significant difference between video sets 1 and 2 for group 1 but not for group 2. Therefore, we found that the discrimination ability for the SRCs that are easy to encode in the high-quality range improved when we conducted the high-quality experiment. This indicates that participants could clearly distinguish the small quality difference in 4k videos using the DSCQS method when only the high-quality assessment videos were presented. This result is different from that of Narita [12] who used HD videos. We argue that because 4k videos have higher resolution than HD videos, participants can perceive the details of the presented scene more precisely, even in the high-quality range; thus significant differences between video sets 1 and 2 were observed. However, as the number of SRCs is larger than that of Narita [12], the effect of increasing the number of SRCs is not clear. Therefore, the discrimination abilities of the DSCQS method for 4k assessment videos and HD assessment videos need to be further investigated and compared.

Similarly, for the ACR-HR method, we performed the Student's t-test with a significance level of 5% between $ACR\text{-}HRi$ of 30 Mbps and $ACR\text{-}HRi$ of 40 Mbps (i = 1, 2) for each SRC. This pair is the only common pair in video sets 1 and 2 for the ACR-HR method. For video set 1, there was a significant difference between video sets 1 and 2 for SRCs 02, 10, and 15. For video set 2, there was no significant difference for any of the 16 SRCs. Because there is only one common pair in video sets 1 and 2, discrimination abilities of the ACR-HR method for the high-quality assessment videos and low-to-high-quality assessment videos must be compared.

7. Conclusion

To investigate the impact of range effects on the double stimulus continuous quality-scale (DSCQS) and absolute category rating with hidden reference (ACR-HR) methods in high-quality assessment videos and to clarify the subjective quality assessment method, which can assess quality of 4k video services that are expected to provide high-quality video, we conducted experiments using high-quality assessment videos and low-to-high-quality assessment videos. We compared both methods in terms of accuracy, stability, and discrimination ability. First, with regard to accuracy, we found that the DSCQS and ACR-HR values were marginally affected by range effects. However, there was no significant difference between DSCQS values using high-quality assessment videos and DSCQS values using low-to-high-quality assessment videos in 46 of 48 processed video sequences (PVSs). Similarly, there was no significant difference between ACR-HR values using high-quality assessment videos and ACR-HR values using low-to-high-quality assessment videos in 30 of 32 PVSs. Second, the DSCQS method was as stable as the ACR-HR method in the experiment using

low-to-high-quality assessment videos but more stable in the experiment using high-quality assessment videos. Finally, the DSCQS method had higher discrimination ability than the ACR-HR method in the experiment using low-to-high-quality assessment videos, whereas both methods had almost the same discrimination ability in the experiment using high-quality assessment videos. We thus determined that the DSCQS method is better at minimizing the range effects than the ACR-HR method in the high-quality range. In addition, we found that participants could distinguish the small quality difference in 4k videos using the DSCQS method when only the high-quality assessment videos were presented even though there was no significant difference between experiments using high-quality assessment videos and low-to-high-quality assessment videos in the work of Narita [12] using HD videos.

In the future, higher quality video services will be developed. For example, Japan is considering "Super Hi-Vision" or 8k content, with transmissions anticipated as early as 2016 (via satellite), with the goal of launching services by 2020 for the Tokyo Olympics [2]. In addition, some electronics manufacturers have recently released High Dynamic Range (HDR) displays. Therefore, whether range effects for the subjective assessment values of such higher quality services are similar to those for the subjective assessment values of 4k video services needs to be investigated.

References

[1] M. Inouye and S. Rosen, UltraHD TV, 4K STB and HEVC STB adoption, ABI Research, 2013.
[2] Y. Sugaya, H. Fujii, A. Sato, H. Matsuda, S. Chaki, and H. Inagaki, "Research and development policies for ultra-high-presence video technology toward 4K/8K services," NTT Technical Review, vol.12, no.5, pp.1–4, May 2014.
[3] T. Hayashi, "QoE-centric operation for optimizing user quality of experience," NTT Technical Review, vol.13, no.9, pp.1–4, Sept. 2015.
[4] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," April 2008.
[5] ITU-R Recommendation BT.500, "Methodology for the subjective assessment of the quality of television pictures," Jan. 2012.
[6] R. Pauliks, K. Tretjaks, K. Belahs, and R. Pauliks, "A survey on some measurement methods for subjective video quality assessment," WCCIT, Sousse, Tunisia, pp.1–6, June, 2013.
[7] Q. Huynh-Thu, M.N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," IEEE Trans. Broadcast., vol.57, no.1, pp.1–4, March 2011.
[8] C. Lee, H. Choi, E. Lee, S. Lee, and J. Choe, "Comparison of various subjective video quality assessment methods," SPIE-IS T, San Jose, USA, vol.6059, pp.605906-1–605906-7, Jan. 2006.
[9] L. Polak, M. Slanina, T. Kratochvil, and P. Llopis Pardo, "A comparison of single and double stimulus subjective assessment of full HD video sequences," Radioelektronika, pp.1–4, Bratislava, Slovakia, April 2014.
[10] T. Tominaga, M. Masuda, J. Okamoto, A. Takahashi, and T. Hayashi, "Performance comparisons of subjective quality assessment methods for video," IEICE Trans. Commun., vol.E97-B, no.1, pp.66–75, Jan. 2014.
[11] T. Kawano, K. Yamagishi, and T. Hayashi, "Performance comparison of subjective assessment methods for stereoscopic 3D video quality," IEICE Trans. Commun., vol.E97-B, no.4, pp.738–745, April 2014.
[12] N. Narita, "Effect of impairment ranges on reliability of the modified EBU method," IEICE Trans. Fundamentals, vol.E78-A, no.11, pp.1553–1555, Nov. 1995.
[13] Q. Huynh-Thu and M. Ghanbari, "A comparison of subjective video quality assessment methods for low-bit rate and low-resolution video," IASTED Signal Image Process, pp.70–76, Honolulu, USA, Aug. 2005.
[14] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," Proc. SPIE, vol.5150, pp.573–582, San Jose, USA, July 2003.
[15] J. Lee, F. Simone, and T. Ebrahimi, "Subjective quality evaluation via paired comparison: Application to scalable video coding," IEEE Trans. Multimedia, vol.13, no.5, pp.882–893, Oct. 2011.
[16] L. Janowski and M. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," IEEE Trans. Multimedia, vol.17, no.12, pp.2210–2224, Dec. 2015.
[17] J. Lee, "On designing paired comparison experiments for subjective multimedia quality assessment," IEEE Trans. Multimedia, vol.16, no.2, pp.564–571, Feb. 2014.
[18] F. Speranza, T. Martin, and R. Renaud, "Subjective quality assessment and the effect of context in expert and non-expert viewers," SPIE-IS T, vol.5294, pp.201–210, San Jose, USA, Jan. 2004.
[19] K. Kawashima, J. Okamoto, and T. Hayashi, "Verification on stability and reproducibility of DSCQS method for assessing 4K Ultra-HD video quality," QoMEX, pp.214–219, Singapore, Sept. 2014.
[20] ITU-T Recommendation H.265, "High efficiency video coding," April 2013.
[21] IPTVFJ STD-004, "IP broadcasting specifications," Jan. 2016 (in Japanese).
[22] ARIB STD-B32, "Video coding, audio coding and multiplexing specifications for digital broadcasting," April 2017 (in Japanese).
[23] Report ITU-R BT.2246, "The present state of ultra high definition television," July 2015.

Kimiko Kawashima received her B.E. and M.E. degrees in engineering from Keio University in 2008 and 2010. She joined NTT laboratories in 2010 and has been engaged in researching the quality assessment of visual communication services. She is currently working on the quality assessment of 3D and 4k video services. She is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan. She received the Young Researchers' Award (IEICE) in Japan in 2013.

Kazuhisa Yamagishi received his B.E. degree in electrical engineering from the Tokyo University of Science in 2001 and his M.E. and Ph.D. degrees in electronics, information, and communication engineering from Waseda University in Japan in 2003 and 2013. In 2003, he joined NTT Laboratories, where he has been engaged in the development of objective quality-estimation models for multimedia telecommunications. From 2010 to 2011, he was a visiting researcher at Arizona State University. He received the Young Investigators' Award (IEICE) in Japan in 2007 and the Telecommunication Advancement Foundation Award in Japan in 2008.

Takanori Hayashi received his B.E., M.E., and Ph.D. degrees in engineering from the University of Tsukuba, Ibaraki in 1988, 1990, and 2007. He joined NTT Laboratories in 1990 and has been engaged in the quality assessment of multimedia telecommunication and network performance measurement methods. Currently, he is the manager of the Communication Quality Group in NTT Laboratories. He received the Telecommunication Advancement Foundation Award in Japan in 2008 and the Telecommunication Technology Committee Award in Japan in 2012.