Quality and User Experience (2019) 4:2
https://doi.org/10.1007/s41233-019-0026-4

REVIEW ARTICLE

Evaluating experiment design with unrepeated scenes for video quality subjective assessment

Lucjan Janowski[1] · Ludovic Malfait[2] · Margaret H. Pinson[3]

Received: 15 November 2018 / Published online: 25 June 2019

Abstract
The conventional video subjective test design, in which subjects view and rate multiple versions of each source video sequence, was used for decades. New technology, like adaptive streaming, makes it almost impossible to use this design since much longer sequences are needed. In this paper we examine three experiment designs: the conventional design and two alternatives that use each source sequence only once. Based on data collected by three laboratories, we compare the accuracy and scoring behavior of these three designs. We check whether there is a significant difference in scoring behavior between the experiment designs. One of the proposed experiment designs is proposed for immediate use.

Keywords Subjective experiment · Precision · Source video · Experiment design · Comparing tests

Introduction

Subjective video quality experiment design, for testing short sequences, has remained essentially static for decades. Here is the most common scenario. A company needs to optimize video encoder settings or to understand the impact of transmission problems on their service. The experiment specifies five to eight contribution quality video scenes, each $\approx$ 10 s duration, and ten to thirty video processing chains (e.g., codec, encoder, bit-rate, coder settings, network errors, decoder). All scenes are processed through all systems, to form a full matrix of scenes and systems. Subjects view and rate these videos in a carefully controlled environment. This data allows statistically significant comparisons between the codecs, encoding options, and network conditions. An example of a typical test can be found here [19]. A more detailed description is given in [22].

The full matrix design was possible since the typical video quality degradations were possible to study with a sequence duration limited to 10 s. It is even considered to test shorter sequences [15]. This experiment design was chosen to be easy in the era of video tapes and the slide rule; it was not chosen for optimality. Since then, we have seen only small incremental changes. For example, per-subject randomized orderings were adopted when computers started to control subjective tests.

New technologies, which could not be tested with the existing methodology, moved the quality of experience (QoE) community to different experiment designs. This is especially true for adaptive streaming and crowdsourcing. Since adaptive streaming changes the delivered quality depending on the network condition, the test sequences have to be long enough to present such changes [29]. Also crowdsourcing experiments, using so called microtasks, do not use a traditional full matrix. Since a task has to be micro from each user's point of view, a single user can see only few sequences [7].

A radical change in experiment design already happening to make evaluation of the new technologies is possible. Adaptive streaming, virtual reality, and QoE [12, 23] are difficult and perhaps impossible to evaluate when using the conventional experiment design. Repeating sequences

✉ Lucjan Janowski
janowski@kt.agh.edu.pl

Ludovic Malfait
ludovic.malfait@dolby.com

Margaret H. Pinson
mpinson@ntia.gov

[1] AGH University of Science and Technology, al. Adama Mickiewicza 30, 30-059 Kraków, Poland

[2] Dolby Laboratories, Inc., 432 Lakeside Drive, Sunnyvale, CA 94002, USA

[3] National Telecommunications and Information Administration, Institute for Telecommunication Sciences (NTIA/ITS), 325 Broadway, Boulder, CO 80305, USA

makes it difficult to use long sequences or to count on user engagement during the content exploration. The need is for a standard experiment design where each subject views each source sequence only once. The consequences of this change are not obvious.

We will address the needs of two audiences. The first audience is people who cannot reuse scenes and want to understand the impacts of not doing so. The second audience is people who are satisfied with the conventional design and who would need strong proof of the benefits of the methodology to motivate a change.

In this paper, we propose and analyze two subjective experiment designs where each subject views each source sequence only once. We will refer to these as "unrepeated scene experiment designs." Both experiment designs assume that the experimenter must be able to compare different test conditions,[1] which we will refer to as Hypothetical Reference Circuits (HRC) [11]. The first design compares HRCs using source sequences with similar but not identical content (e.g., different time segments from one sporting event). The second design compares HRCs using source sequences with similar coding complexities.

Motivation

Let us begin by looking more closely at the practical and theoretical reasons that motivate these new experiment designs.

Throughout this paper, the following definitions apply:

- *Scores* are the raw data collected from subjects on a subjective rating scale [12]
- *Mean opinion scores (MOS)* the mean of the opinion scores collected for a stimuli in the considered experiments, typically a single experiment
- *Standard deviation of opinion scores (SOS)* is standard deviation of the opinion scores collected for a stimuli in the considered experiments, typically a single experiment
- *Hypothetical reference circuit (HRC)* is one system under test
- *HRC MOS* is the mean of the opinion scores of an HRC, computed by averaging over subjects and scenes. This term is not defined in a recommendation but is useful for our analysis
- *Processed Video Sequence (PVS)* is any video sequence that will be scored by subjects [11].

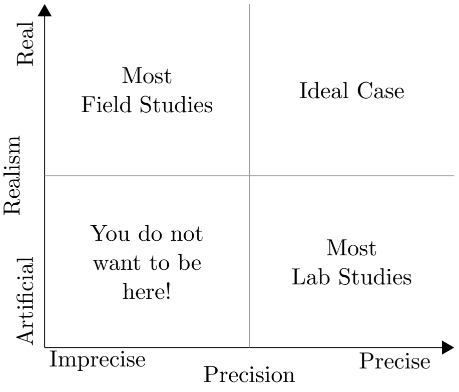

Fig. 1 Experiment design realism versus precision

Practical

Subjective experiments provide an important tool to gather user opinions and improve products. A subjective experiment, like any experiment, has to balance two contradictory constraints (see Fig. 1). On the one hand, the experiment should be precise to obtain statistically significant MOSs with minimum effort. On the other hand, the experiment should be as realistic as possible (called in psychology "Ecological Validity" [27][2]).

The conventional experiment design repeats the same source sequence (SRC) for each HRC. As a result subjects are exposed to the same source material many times during an experiment, which allows them to learn what each source should look like. This provides a direct comparison that increases precision yet decreases realism (Ecological Validity), since a user typically does not see exactly the same content multiple times under different conditions.

Therefore, SRC reuse may impact MOSs. Moreover, for upcoming research areas SRC reuse is impractical or undesirable. Examples include:

- adaptive streaming, due to the requirement for long SRC
- crowdsourcing, where new SRC may improve the odds of subjects paying attention
- quality of experience, where realism is critical
- tests that examine the video quality of cameras.

The conventional experiment design is described by Pinson et al. [22] and used by most published experiments. Variants tend to expand upon the idea of comparing different versions

[1] By test condition we understand the method of implementing some changes to the original video such as specific compression, packet loss, or display technology.

[2] The term "Ecological Validity" is multidimensional and has been discussed for years by the psychology community as described in [27]. More research is needed to fully understand how this could be well applied to video quality assessment.

of a single SRC, e.g., using three different monitors at the same time [5]. These designs may increase precision, yet further decrease realism.

Researchers seldom design experiments to avoid SRC reuse. Crowdsourced experiments can be divided into tasks that show each source only once (e.g., see Ribeiro et al. [24]), yet a subject who performs multiple tasks still views the SRC several times. Frohlich et al. [4] uses content classes instead of SRC to study the impact of content duration on MOSs. This design reduces SRC reuse yet does not eliminate it.

Sullivan et al. [28] observe that ITU-R Rec. BT.500 has its roots in psychophysics. The goal of psychophysics is to find just noticable differences (i.e., quality thresholds). In a nutshell, the conventional experiment design address the needs of video codec developers to fine tune parameters. This method was not designed to help service providers make difficult business decisions, like trade-offs between bit-rate and customer expectations around video quality.

The proposed solution [21, 28] is to more accurately measure the system quality and acceptability by immersing the subject in a more natural viewing experience. This “immersive method” uses distractor questions and longer audiovisual sequences to focus the subject on the intended application. This method avoids SRC reuse yet retains the full matrix of (SRC × HRC) by dividing the subjects into groups and showing each group a different pairing of SRCs to HRCs. The researchers community objected that this particular method is too cumbersome and expensive (Video Quality Experts Group discussions). The stimuli for ten groups would take as much effort to prepare as ten conventional experiments. However, the concept of an immersive method based on human factors is gaining support.

Robitza et al. [25] proposed a more practical immersive design. Their goal was to understand the impact on video quality ratings of network traffic on HTTP adaptive streaming when subjects are engaged by interesting content. Like Sullivan, Robitza used longer sequences (1 min) of entertaining audiovisual content. The distractor questions were eliminated and each HRC was paired with three different SRC. This addresses the cost concerns while retaining the idea of an “immersive” method. The subjects were more entertained and were able to participate in a longer test than is possible with the conventional design. The missing element is a structure to decide how to pair SRCs to HRCs.

Theoretical

There are three theoretical reasons why the conventional design is not optimal.

The first theoretical reason questions the validity of absolute MOSs. In [30] it was shown that comparing two sequences gives statistically the same values as scoring only one sequence. This is surprising, since we know that people are better in comparing than absolute judgement [16]. Truly “absolute” MOSs are probably impossible to obtain since we always compare sequences to our expectations, our memory, the training sequences, etc. Also different sequences which are presented in the experiment are influencing the obtained results [8]. Nevertheless, by repeating sequences we give subjects an easy way to compare different sequences.[3] As a result the Absolute Category Rating (ACR) method does not really produce absolute MOSs, but rather provides an assessment close to a Degradation Category Rating (DCR) method.

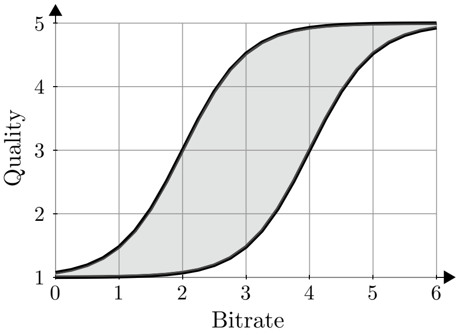

Fig. 2 Depending on the sequence specificity, the transition from MOS 2 to 4 occurs at different bitrates. The gray area denotes unknowns in the correct shape of the curve for specific sequence

The second theoretical reason addresses the goal of the experiment. A typical subjective experiment should bring us closer to the real world; this is the reason why we ask subjects for their opinions. The most realistic scenario is a field study where we observe a user in a typical situation, interacting with a service under investigation. By the user’s actions, we should be able to guess what the service quality is. Such interactions are not focused on testing the system but rather on watching the content. In this sense, repeating exactly the same content makes an experiment less real as shown in Fig. 1.

The third theoretical reason originates from the information needed to characterize a video system. A typical quality transition from bad to excellent can be described by a function having saturation on both ends, like the logit function (see Fig. 2). We need saturation on both ends because a change from 10 kbit/s to 20 kbit/s when compressing 4K video will not change the obtained quality; it is all “Bad.” For the same video, changing from 10 Gbit/s to 20 Gbit/s will not change the quality since it is already the source

[3] An interview with our subjects revealed that even when we asked them to rate each sequence separately, some of them spot a specific place in a sequence to “detect” distortions.

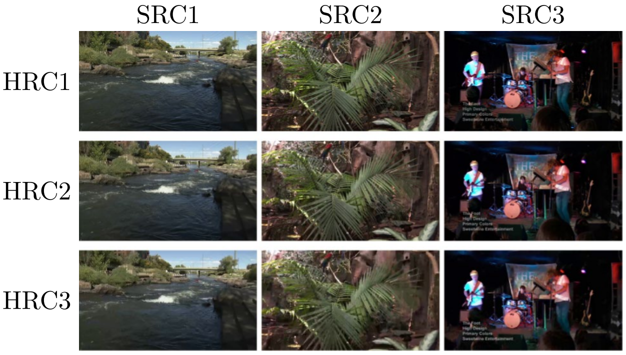

Fig. 3 Conventional design (SRC × HRC)

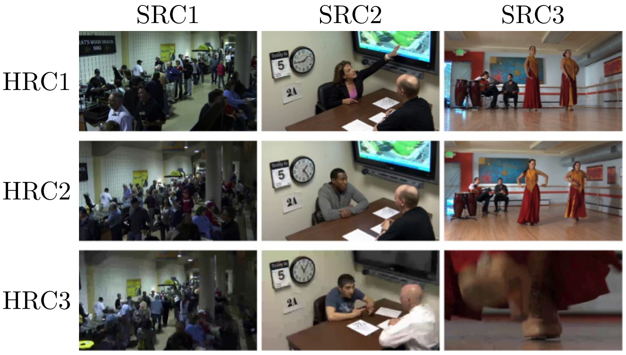

Fig. 4 Related sequences design: (RSRC × HRC)

quality. From the optimization point of view, it is important to distinguish between saturation, where bitrate changes have little impact on quality, and the almost linear transition, where small bitrate changes have significant impact on quality. Bitrate is one dimension that influences the logit function shape. The other dimension is the content characteristics. Differences in content result in not just a single line but a surface, as shown in Fig. 2. In fact, content characteristics are much richer and cannot be described by a single number. So to correctly sample the content characteristic space, we should use as many different contents as possible. Some more details about this problem can be found in Pinson [18].

We recognize that there are some practical advantages in designing experiments with repeated sources. It is more cost effective as the experimenter only has to obtain a small number of SRC. High quality sources can be expensive, especially for new technologies. The conventional design is also less labour-intensive as only a small number of video sequences must be selected and edited.

Experiment designs

We want to explore three experiment designs in this paper: the conventional design, and two variations of the unrepeated scene design. This section describes each experiment design, and more details can be found in Pinson and Janowski [17]. Several other unrepeated scene designs are described but not analyzed.

Conventional design: (SRC × HRC)

Subjective tests are typically designed to include one full-factorial matrix of (SRC × HRC) (see Fig. 3). Fundamentally, the test measures whether or not subjects can perceive a difference between two versions of the same stimuli. The (SRC × HRC) experiment design reflects the real world situation where a store shows the differences among televisions (the various HRCs) by playing the same content (the same SRC) to multiple televisions.

The (SRC × HRC) experiment design is unrealistic because consumers can seldom compare differently impaired versions of a single SRC. Therefore, let us propose two fundamentally different ways to maintain the conventional subjective test design (SRC × HRC), while eliminating SRC re-use. To do this, we will replace each SRC with a set of SRCs, within the experimental design.

Related sequences design: (RSRC × HRC)

Let us define a set of related source sequences (RSRC) to be a set of SRCs that have visually similar content. We replace each SRC in the subjective test design with this set of sequences. Thus, the test design changes from (SRC × HRC) to (RSRC × HRC) (see Fig. 4). During data analysis, the SRCs in each RSRC set are treated as if they were identical. This test design reflects the real world situation where a consumer compares two different television distribution systems by viewing similar subject matter (e.g., football games or news).

This design is similar to the balanced partial block design commonly used in speech quality assessment for codec evaluation and objective model training.

Let us assume a full-factorial design of (RSRC × HRC). In such an experiment we need as many SRCs as we are planning to show PVSs. If there are m RSRC groups and n HRCs, then the test would need $m \times n$ SRCs. This requires sets of n SRCs (so called RSRC) that depict one idea and are produced using similar filming and editing techniques. The obvious choice is to edit different segments from a single production (e.g., a music video, a football game).

There are three advantages to the (RSRC × HRC) experiment design. First, SRC memorization is avoided so the scores should be a more realistic reflection of our subjects' true opinions. Second, boredom is reduced. Third,

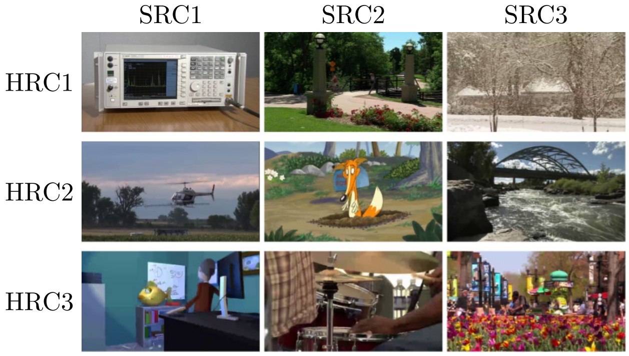

Fig. 5 Coding difficulty design: (CD–SRC × HRC)

the scoring scenario is more realistic (i.e., better matches a user's experience).

There are two disadvantages. First, data analyses will be more difficult, because the SRC and HRC variables are confounded. Second, the SRCs within one RSRC may have very different coding difficulties. This disadvantage inspires our second proposal.

Coding difficulty design: (CD–SRC × HRC)

Given an encoder and a constant bit-rate, we can sort SRCs by quality. The SRCs that look best we will refer to as having low coding difficulty; the SRCs that look worst we will refer to as having high coding difficulty. The quality ordering differs somewhat depending on the codec and bit-rate, so coding difficulty is imprecise.

Let us define a coding difficulty source sequence (CD–SRC) as a set of SRC that have a similar coding difficulty. For example, one CD–SRC set might include sequences with low coding difficulty, while another might include sequences with high coding difficulty. Since the decision was made by an automated algorithm, these SRCs may have very different visual characteristics (e.g., different content types, camerawork, editing, and aesthetics). The intent is that each CD–SRC includes a large variety of subject matter and visual characteristics, so as to disallow comparisons between PVSs. This test design reflects the real world situation where a consumer judges different video distribution systems based on disparate content (e.g., a variety of movies).

Let us assume a full-factorial design of (CD–SRC × HRC) (see Fig. 5). If there are m CD–SRCs and n HRCs, then the test would need $(m\ \mathrm{CD}-\mathrm{SRC}\times\ n\ \mathrm{HRC}) = (m \times n)$ SRCs. This requires a large variety of visually dissimilar content and an automated algorithm to calculate the coding difficulty of an SRC. The Appendix of Fenimore et al. [3] provides the best available algorithm: scene criticality.

There are three advantages to the (CD–SRC × HRC) experiment design. First, SRC memorization is avoided. Second, boredom is minimized. Third, the scoring scenario is more realistic.

There are three disadvantages. First, interactions between SRC, HRC, and subject may have more influence on scores, due to the large variety of SRCs. Second, the number of unrelated SRCs needed for an experiment increases as a factor of the number of HRCs. Third, the scene criticality algorithm is far from ideal. This experiment design would benefit from improved estimates of coding difficulty.

Other designs

The proposed designs are not the only possible solutions. The simplest other solution is to remove the coding difficulty constraint and randomly assign SRCs to HRCs. In that way we only specify HRCs and how many times each is repeated. Since SRCs are not repeated, the same HRC will be matched with different SRCs. This simplistic design is called the "random design" and it was considered in one of our sessions. Nevertheless, it is difficult to reach very strong conclusions about the random design, since the particular sequence of random numbers can influence the obtained results.

An interesting solution proposed in Robitza et al. [25] is to consider a large number of HRCs, where the number of HRCs equals the number of presented sequences. An obvious disadvantage of this solution is the relationship between SRCs and HRCs. We are not able to remove that relationship and so cannot deduce whether one system is better than another. On the other hand, such comparisons are not always of interest and large numbers of HRCs can help us to understand the full complexity of the analyzed HRCs, especially if HRCs are complicated as in Robitza et al. [25]. This experiment design is better suited when HRC comparisons do not appear in the stated goal.

The above cited paper describes the proposed experiment design as "immersive." There is even more immerse method: just show a whole movie [2]. A common problem with continuous quality evaluation is diversity among scoring behaviors [22]. Showing a movie cut to pieces decreases the immersion but still some subjects reported, "I feel like I watched the whole movie." Immersive designs should use this advantage and show sequences that are different parts of a longer sequence. Such experiment designs can be similar to RSRC or CD–SRC, depending on how different the parts are.

None of these other designs are considered in the remainder of this paper. Our goal was to compare different designs in a single subjective experiment. Such an experiment cannot be too large, therefore we were not able to consider all possible designs. Also factors like immersion or randomness

Table 1 The age and gender distribution for AGH and Dolby experiments

Laboratory	Gender	Under 21	21–31	31–40	41–50	51–60	Over 61	Total
Dolby	Female	1	4	2	2	3	1	13
	Male	1	4	3	2	2	0	12
	Total	2	8	5	4	5	1	25
AGH	Female	5	7	4	1	0	0	17
	Male	2	6	3	3	0	0	14
	Total	7	13	7	4	0	0	31

play an important role in the these designs, so they would be more difficult to analyze than the RSRC and CD–SRC designs.

Subjective experiment

This paper uses data from two subjective experiments: AGH/NTIA and AGH/NTIA/Dolby. Both dataset are available on the Consumer Digital Video Library (CDVL, www.cdvl.org).

Dataset AGH/NTIA

We began by conducting a preliminary subjective experiment, AGH/NTIA [17]. Basically, we designed three small experiments, using the first three experiment designs described above (conventional, related sequence and coding difficulty). The resulting PVSs were mixed together and split into three viewing sessions. Some clips were repeated in all three sessions, because an important goal of AGH/NTIA was to understand subject scoring behaviors (see Janowski and Pinson [13]). Subjects answered a short questionnaire between sessions and a longer questionnaire at the end. These questions sought the subjects' opinions of the three experiment designs.

This paper uses the AGH/NTIA questionnaire from [17]. All scoring analyses in this paper use the data from our newer subjective experiment, AGH/NTIA/Dolby.

Dataset AGH/NTIA/Dolby

Dataset AGH/NTIA/Dolby contains six sessions. The basic idea was that each session would answer the same experimental question with a different experiment design. The experimental question is quality comparisons between 10 HRCs: the original video; MPEG-2 with bitrates 7, 4, and 2 Mbit/s; H.264 with bitrates 2, 1, and 0.5 Mbit/s; H.265 with bitrates 1, 0.5, and 0.25 Mbit/s. The experiment adhered to ITU-T Rec. P.913. No demographics were collected for NTIA. Subjects self-reported as having normal vision for both NTIA and Dolby experiments, for AGH experiment all the testers passed a vision test. Age and gender of subjects in AGH and Dolby are presented in Table 1, except one AGH subject whose data are missing. All laboratories recruit subjects by temporary employment agencies trying to get balanced gender and age.

To limit the test duration, each session contained 40 PVSs. Thus, the conventional design (SRC × HRC) required 4 SRC, the related source design required 4 RSRC (i.e., 10 samples from 4 related sequence sets), and the coding difficulty design required 4 CD–SRC (i.e., 40 sequences divided among 4 coding difficulty levels). Each sequence was 8 s in duration. These three scene pools were used for all six sessions.

The first three sessions used the SRC, RSRC, and CD–SRC experiment designs. Those sessions were presented in a random order to each participant, so that session order would not influence the scores. We can consider the first three sessions as a separate experiment, as all three were rated before the other three sessions. These three sessions will be referred to as SRC_1, $RSRC_1$ and CD_1, respectively.

The last three sessions were variations of the first three and were included to test the design stability. The fourth and fifth sessions reused the RSRC and CD–SRC experiment designs, but the sequences were randomly reassigned to HRCs. The sixth session uses the *CD–SRC* design but replaces the scene criticality algorithm with a random number generator. Again, those sessions were presented in random order. These three sessions will be referred to as $RSRC_2$, CD_2 and $Rand_2$. The goal of the sixth session, $Rand_2$, was to provide some insights into the value of the coding difficulty algorithm (or lack thereof).

The experiment started with a short training session. After each session we had a very short questionnaire and then a short break. After all sessions, we had a longer questionnaire that asked more detailed questions. The questions were about liking or disliking particular session. The main goal of the questionnaires was to understand the influence of our experiment designs on how subjects perceived and rated videos.

The experiment was run by three different laboratories: AGH, ITS, and Dolby. In total 81 subjects participated in the study: 32 at AGH, 24 at ITS, and 25 at Dolby. Dolby ran two more subjects, whose data was incomplete; these subjects' data is ignored. The six sessions were not randomly shuffled for ITS subjects, therefore some analyses use only AGH and

Table 2 Summary of the experiments run in order to validate the experiment design and used in the analysis section

Dataset	AGH/NTIA	AGH/ NTIA/ Dolby
Number of SRCs	34	84
Number of HRCs	5	10
Number of SRCs per HRC	Varies	4
Number of PVSs	114	240
Number of PVSs per session	60 or 90	40
Number of Labs	1	3
Total number of subjects	28	81
Scores per PVS	28, 84, or 168	81
Number of sessions	3	6

Table 3 What did you like about this session? What did you not like about this session?

#	Answer description
22	Liked new SRC and/or disliked repeated SRC
16	Liked some content
5	Disliked some content
3	Liked repeated SRC

Table 4 Repeated versus unrepeated SRC

(a) You saw some video sequences many times. How did this impact the way you decided on quality scores?

#	Answer description
8	Compared to memory of prior viewings
8	Easier to decide, more accurate
6	Negative impact on scoring behavior e.g. focusing on small part of sequence
1	Did not influence

(b) You saw some video sequences only once or twice. How did this impact the way you decided on quality scores?

#	Answer
6	No impact
6	More difficult to decide
5	Easier to decide, no rethinking prior scores
4	Unique and interesting
4	Compare to other or the same content

(c) If you saw a new content, how much did your interest in the content impact your scores?

#	Answer
10	No impact
10	Paid more attention
5	Rated them higher

Dolby data. AGH subjects answered a short questionnaire after each session.

A summary of the subjective experiments is given in Table 2.

Questionnaires

We will begin with the questionnaire answers, as these provide subject opinions of different experiment designs.

AGH/NTIA Questionnaires

This section summarizes relevant portions of the questionnaires. The free-response answers were categorized by the authors and are presented in three tables. Table 3 summarizes the between session questionnaires. Table 4 summarizes feedback related to the experiment design. Table 5 summarizes feelings of alertness and inattention. Tables 3 and 5 each summarize two questions, and so contain up to 50 responses. Column "#" indicates the frequency of an answer, as categorized by the authors.[4]

The main conclusion we can draw is that most subjects dislike repetitions and like variety. Subjects also had individual preferences for content, regardless of rendering quality (e.g., liking mountain vistas).

When SRC variety is large, subjects are pleased with the experiment and better able to pay attention. When a SRC is repeated, some subjects report a change in their scoring decision process (e.g., focus on a small part of the sequence, pay less attention, or choose lower quality scores).

Some of the CD–SRC 40 video sequences depicted subject matter that was very dissimilar to all other content (e.g., a close-up view of a burning house). Subjects perceived these rare sequences as more difficult to score than repeatedly viewed SRCs (compare Table 4a, b). This might counteract some of the benefits of increasing SRC variety. The RSRC design allows for comparisons among similar content.

Table 4c shows higher quality attributed to new SRCs. This is undesirable but probably only important when the experiment design has some SRCs viewed repeatedly and others only once. Other subjects reported no impact or better focus, which are both positive in terms of how we want subjects to behave.

Table 4a suggests that repeating SRCs creates a test that is closer to a paired comparison than an "absolute" category rating.[5] This might explain why Tominaga et al. [30] found only small differences between Pair Comparison and Absolute Category Rating experiments.

[4] The questionnaire used the term "ratings."

[5] See Pinson et al. [22] for descriptions of these subjective methods.

Table 5 When did you have an easy time staying alert and paying attention? When was it difficult to pay attention?

#	Answer
14	New content helped
14	Repeated content hurt
8	Enjoyable subject matter helped (e.g., cartoons)
6	Disliked content hurt
6	Less alert as the test progressed

Subjects reported increased accuracy when scoring repeatedly viewed SRC. This was subjects' opinion about their precision therefore we will investigate this perception during our data analysis.

AGH/NTIA/Dolby Questionnaires for AGH experiment

After each session, the AGH subjects were asked "What do you like about this session?" and "What do you dislike about this session?" The intention was to provide a simple, almost numerical, way to estimate the probability that a subject liked a particular session. Unfortunately, subjects did not directly answer this question. In many cases they described the process of scoring as being easy or difficult. Therefore, for each session, this section summarizes the opinions in a descriptive way, instead of divided by likes and dislikes.

For the SRC experiment design, we have 64 answers (two per subject). Thirty eight answers were "no comment", mostly because subjects expressed all their opinion in the first answer and left the second empty. A similar situation occurred for the other experiment designs.

Table 6 shows the answers for the SRC experiment design, where the number indicates how many subjects had a similar opinion. Some people focus on their scoring consistency, which makes the voting process much more "thinking by comparison" than "flow of experience." On the other hand, we can see that some distortions, like color change, are probably almost impossible to detect without comparisons.

The RSRC experiment design (see Table 7) received many more comments about the content and the voting process, like blurring or blockiness was worse than clear images. This shows that people paid attention to the flow of the watching "experience" rather than just comparing with previous sequences. On the other hand, some subjects had a more difficult time choosing scores during the RSRC session.

The CD experiment answers (see Table 8) indicate that truly unique content was more interesting to watch but more difficult to score. Subjects recommended the content be divided by similar conditions, such as sequence brightness.

Table 6 What did you like/dislike about this (SRC) session?

#	Answer
5	Repetitions hurt and generate doubts
3	Building scale by comparison
2	This session makes it (scoring?) easier
2	Easier by comparison
1	Specific content was difficult to score
1	I like/dislike high/low quality
1	Repeated sequences helped with color distortion detection
1	Very similar movies
1	The recording quality was difficult to compare
1	Small quality differences

Table 7 What did you like/dislike about this (RSRC) session?

#	Answer
5	Linking to specific voting problem
3	It is more difficult to score than for SRC
3	Interesting (the SRC told a story)
2	Link to specific content
1	Easier to score since I do not compare

Table 8 What did you like/dislike about this (CD–SRC) session?

#	Answer
3	Now interesting before boring
1	The easiest to score is SRC and CD, RSRC most difficult
5	Linking to general quality, more good, more bad etc.
1	More difficult to score—no comparison
1	Linking to recording quality like low lighting condition is different than full sun

We also asked subjects how they defined quality. These answers focused on sharpness, both in terms of picture quality and color. For some subjects, their quality definition changed within the test. Some identified a specific distortion, like blockiness, as especially annoying. These subjects could react to MPEG-2 compression more strongly than others. People said that different sequences required different quality, and they took this into account. Colors appear especially often, which is interesting knowing that most objective video quality metrics are luminance-based. Some subjects mentioned recognition as a quality indicator. No one said that their quality definition depended on the experiment design, but for some experiment designs, some aspects are more obvious, like different movie types had different acceptance thresholds.

We also asked whether repeating the same source influenced scores (see Table 9). The answers indicate that

Table 9 Did repeating the same source influence the score?

#	Answer
12	Help by comparison
8	No
6	More difficult, boring
5	Trying to be consistent is tiring
2	Fitting the scale

Table 10 If you see a scene just once did it influence the score?

#	Answer
17	No influence
5	More difficult
4	Not decided
4	Easier since no comparison
3	More interesting
1	New content increases the score
1	Less precise since new content disturbs me

Table 11 Did content influence the score?

18	No
5	More colorful scenes can get better score
3	Yes
2	Repetition removes the influence of content on my scores
2	Yes, it should since different content needs different quality
2	Interest in content shifts the focus from quality to content

Table 12 When was it easy to focus?

8	At the beginning
8	For better quality
5	No matter
3	Start of sessions
2	If quality was different
2	Easier for worst quality
1	If boring, easier to focus on quality
1	Middle of experiment
1	If interesting

Table 13 When was it difficult to focus?

9	End
5	Bad quality
4	No matter
3	Repeating sequences
2	Specific content
2	Good quality
1	Interesting content
1	First and last session

comparison is the most important aspect of SRC design. It seems that subjects almost perform a pair comparison, by thinking about consistency and comparing sequences.

We asked if truly unique sequences were rated differently (e.g., a topic that was only viewed exactly once, during the CD–SRC session). As shown in Table 10, about the same number of subjects thought that it is easier or more difficult, compared to repeated sources.

We asked subjects whether the content influenced their quality scores (see Table 11). Before the test, the proctor read instructions out loud that asked subjects to disregard the content. Still, some people were honest enough to admit this influence. Two subjects answered that repetitions eliminated the quality influence of content preferences; and two other subjects replied that content should influence the quality score. One used this reasoning: talking heads do not need as good picture quality as a documentary movie showing different landscapes.

The last two questions investigated the ease or difficulty of focusing on the scoring task. The answers, summarized in Tables 12 and 13, indicate that focus is mostly influenced by the time within the experiment. This is obvious. Nevertheless, we learned that it would be harder for subjects to focus on an experiment with low quality sequences.

Subject analysis

Let us begin our data analysis by considering in greater depth the two motivations for choosing the unrepeated scene experiment design:

1. Experiment cannot use repeated scene
2. Unrepeated design is atypical and should be compared with a more traditional design.

In the first case there is no other choice. You would like to know how much your experiment differs from the conventional design. We will refer to such reasoning as "mandatory." The second motivation is curiosity. The conventional design could be used, so the new design must add value, thus improving the experiment. We will refer to such reasoning as "desirable." Those two reasons call for different proof, therefore separate analyses are needed.

Regardless of the reason behind change, we would like to test whether there is a significant difference in scoring behavior between SRC, RSRC, and CD–SRC experiment designs. We are interested in investigating the variance and repeatability of scores. Are there any trends in MOS or SOS? What is the user opinion on different experiment

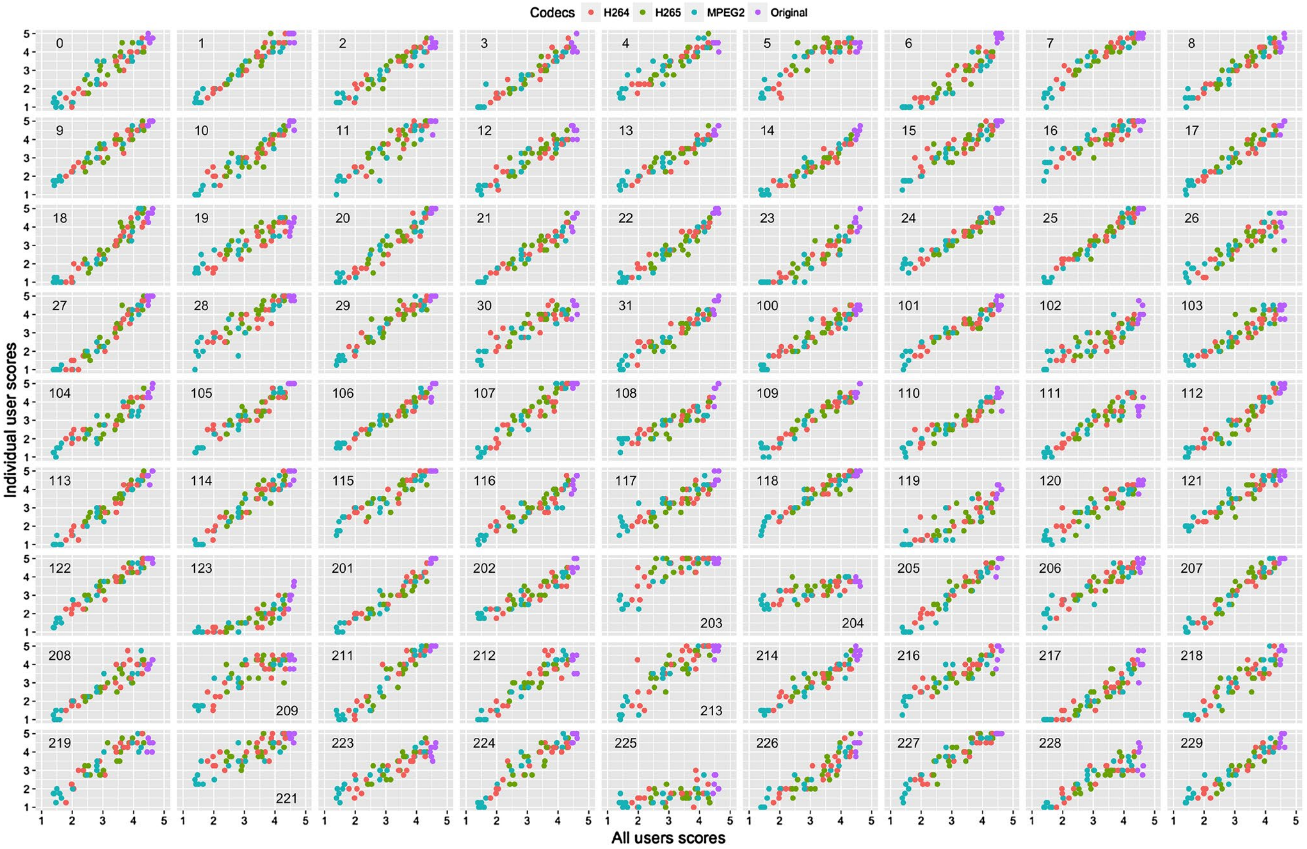

Fig. 6 Relationship between HRC MOS obtained from a single subject and all subjects in the AGH/NTIA/Dolby dataset. The x-axis plots HRC MOS computed across all subjects; the y-axis plots HRC MOS for one subject. Scores are aggregated to create continuous values that emphasize trends

designs? The next two sections investigate these differences. We will begin by validating subjective data.

Behavioral subject screening

According to the theoretical subject scoring model presented in Janowski and Pinson [13], subjects' scoring is a random process. This is expected behavior that must be accepted and not a flaw that can be eliminated. Some subjects' scores contain more random error (β) or a large bias (Δ) compared to other subjects.

Removing bias increases the statistical power of MOSs. Since our goal is to measure small differences between different experiment designs we need stable subjective scores, precise subjective MOSs, and comparable MOSs from all laboratories. Excessive scoring errors and unusual scoring behaviors could hide the differences between the experiment designs.

To analyze each subject's scoring behavior, we generated a scatter plot for each subject versus all subjects' scores in the AGH/NTIA/Dolby dataset (see Fig. 6). We need continuous data for this analysis, so Fig. 6 compares HRC MOS computed from one subject's scores with HRC MOS computed from all subjects' scores. We looked at the scatter plots to identify subjects with unusually large data scattering or atypical scoring trends. From these plots, we see that subjects 5, 123, 203, and 226 did not use the whole scale symmetrically; 204 and 225 scored almost a constant value; and 119, 209, 213, and 221 have strong scattering.

We will generate two sets of subjective data. The first eliminates the above subjects, to form a subset of subjects who are most consistent with the test average. The second is the set of all subjects, regardless of their scoring behavior. Analyses with the full dataset can be used to check the validity of our subject screening.

Subject screening by experiment design

If an experiment design causes an increase in the number of subjects rejected, then it is definitely a drawback of that design. Let us compare the experiment designs based on the number of subjects rejected by Annex A.1 of ITU-T Rec. P.913. Pearson correlation is calculated between each subject and the mean of all subjects. This value was checked against

Table 14 Subject screening using Pearson's linear correlation

subID	CD_1	$RSRC_1$	SRC_1
5	**0.723**	0.867	0.800
103	**0.651**	0.863	0.765
116	**0.637**	0.885	0.786
123	**0.663**	**0.676**	**0.728**
203	0.784	0.883	**0.731**
204	**0.433**	**0.700**	**0.729**
209	**0.715**	**0.666**	0.819
213	0.814	0.923	**0.664**
221	**0.593**	**0.725**	**0.692**
225	**0.475**	**0.364**	**0.653**

Values lower than 0.75 (i.e. the ITU-T Rec. P.913 threshold) are in bold

a threshold of 0.75. The results are presented in Table 14, with each of the first three sessions treated as separate experiments. Table 14 omits subjects whose Pearson correlation values are above 0.75 for all three sessions.

Table 14 shows that a total of 8, 5, and 6 subjects are rejected from sessions CD_1, $RSRC_1$, and SRC_1, respectively. The numbers are close. If we omit subjects who are rejected by all three sessions (123 and 204), then 4, 1, and 2 subjects are rejected. These differences are still too small to reach statistically significant conclusions.

Lab-to-lab comparison

Let us reject the 13 subjects with unusual scoring behaviors ("Behavioral subject screening" section) or low correlation ("Subject screening by experiment design" section) and then compare the MOSs from different laboratories. Standard lab-to-lab comparisons yield very high correlations: 0.98 between Dolby and AGH; 0.98 between Dolby and NTIA; and 0.99 between AGH and NTIA. We conclude that the experiments can be combined to a single set. The larger number of scores per PVS increases the chance of detecting differences.

Subject bias removal

After subject screening, we removed each subject's bias from their scores before calculating MOS and SOS (see Janowski and Pinson [13]). This increases the sensitivity of statistical comparison without impacting MOSs or the cost of the experiment. Our analyses focus on MOS and SOS comparisons, so bias should be removed.

Data precision and stability analysis

SOS analysis

We desire experiment designs that yield more precise data (see Fig. 1), meaning the scores for each PVS are less scattered.[6] We want all subjects to have a similar experience and to be able easily decide on scores. We cannot compare SOSs directly, as the three experiment designs yield different MOSs.

Hossfeld et al. [6] propose a single parameter that characterizes the relationship between MOS and SOS for a particular experiment. We will refer to this parameter as the Hossfeld–Schatz–Egger (HSE) coefficient.[7] The theoretical maximum SOS for each MOS value describes a curve. An experiment's data typically describes a similar curve, lying somewhere below. The HSE coefficient fits this curve to the SOS values of a particular experiment. This condenses an experiment's score distribution into a single value. The curve is characterized by equation:

$$SOS(x)^2 = a(-x^2 + 6x - 5) \tag{1}$$

where x is MOS, $SOS(x)^2$ is the variance of scores (SOS^2) for particular MOS, and a is the HSE coefficient that characterizes the experiment.

The HSE coefficient can be used to describe the difficulty of the scoring task for many different experiments, as shown in [6]. It provides an elegant and effective way to measure the spread of scores in an experiment independently from the MOSs obtained within the experiment.

Figure 7 plots the relationship between MOS and SOS expressed in (1) for the SRC_1, $RSRC_1$, and CD_1. The HSE values obtained by least-square fitting are 0.225, 0.221, 0.216 for SRC_1, $RSRC_1$, and CD_1 respectively. These HSE differences are not statistically significant [31]. This indicates that changing from the SRC to RSRC or CD–SRC design does not increase HSE.

Note that the HSE data contradicts the subject questionnaire feedback, in which subjects reported increased accuracy when scoring repeatedly viewed SRC.

Error analysis

Janowski and Pinson [13] propose a model for scoring behavior based on subject bias and subject error:

$$o_{ij} = \psi_j + \Delta_i + e_{ij} \tag{2}$$

[6] This and other analyses ignore the special cases where the goal of the experiment is to detect differences among subjects.

[7] Hossfeld et al. [6] use the term "SOS parameter a." We believe that term can be confusing when used outside of the context of their paper.

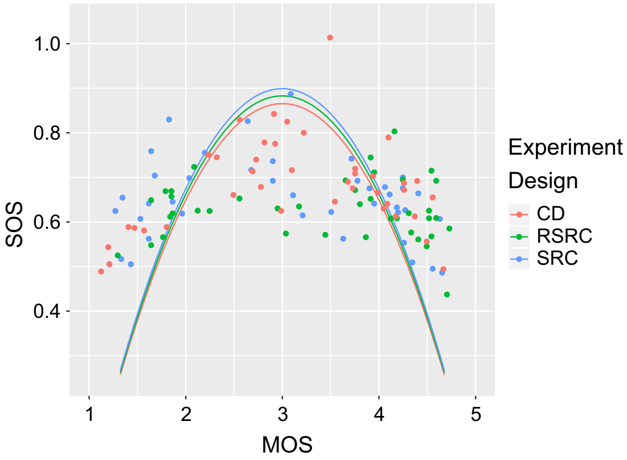

Fig. 7 HSE coefficient curves characterizing the three experiment designs show (SRC × HRC) and (RSRC × HRC) have similar scatterings of scores

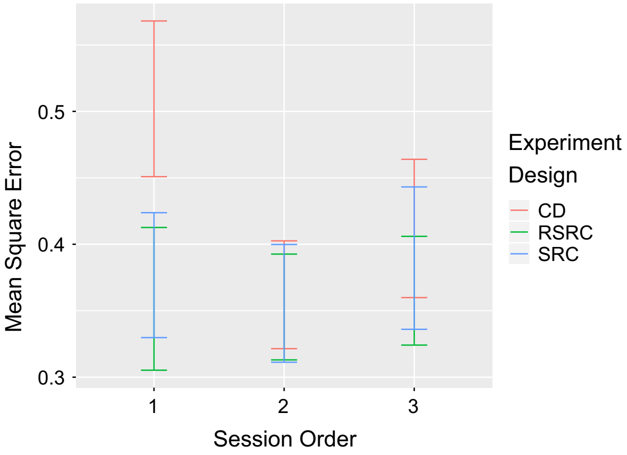

Fig. 8 Error with the confidence interval for each session and experiment design

where

- o_{ij} is the score given by subject i for PVS j
- ψ_j is the PVS MOS
- Δ_i is subject bias
- e_{ij} is the error

Variable e_{ij} includes multiple factors (e.g., subject i's scores are imprecise, PVS j is difficult to score).

By solving for e_{ij}, we can analyze the errors of the subjects' individual scores:

$$e_{ij} = o_{ij} - \Delta_i - \psi_j \tag{3}$$

By "error" we mean the deviation of observed values from the mean, not any mistake on the part of the subject.

Since the error can be either positive or negative, and we are interested in error itself, we calculate the square. In general, e_{ij} should be as small as possible but the five point scale limitation makes it impossible for e_{ij} to be lower than a certain level. Also, when comparing different PVSs, differences in e_{ij} can be caused by ψ_j being closer or farther from a discrete value of the scale (e.g., if ψ_j is 3, the minimum possible SOS is zero; if ψ_j is 3.5, the minimum possible SOS is 0.5).

Figure 8 shows the spread of e_{ij} for each experiment design, organized by session order (i.e., whether that experiment design was viewed first, second, or third). We want to know whether session order and experiment design influence the obtained error. There are no differences except for the surprisingly high error obtained when the first session has the CD experiment design. This could be caused by a different scoring behavior when the CD session appeared first. After the first session, subjects' scores are influenced by all prior sessions' PVSs and MOSs. It is difficult to say if we should consider this to be a positive or negative feature of the CD experiment design.

Interestingly, Fig. 8 shows an expected trend of lowest error for the second session, where subjects know the experiment well and are not yet tired.

Distribution of MOSs

Let us examine the distribution of MOSs within an HRC, to gain insights into which experiment designs do a better job of representing the "big picture" of all subject matter. We will conduct this analysis two ways.

First, we used the Student's t test to compare whether the MOSs associated with the original video in SRC_1 and $RSRC_1$ were independent samples from the same normal distribution at the 5% confidence level. This analysis was repeated for each HRC (we have ten different HRCs, see "Dataset AGH/NTIA/Dolby" section) and all possible session pairs (we have six different sessions so there are 15 different pairs). Of the 150 comparisons, only five (3.3%) were from different distributions. This is within the expected response at the 5% level.

Second, we combined all data into a single distribution, to enable figures that visually portray the data. Each of the ten HRCs were normalized for zero mean and unit variance. This aggregated data indicates an approximate distribution of MOSs for a generic HRC. This analysis aggregates MOSs instead of scores, because we consider each PVS to be one realization of the HRC. Considering each HRC in each session separately, we used the two-sample Student's t test to test whether that HRC's four MOS values were independent samples from the normal distribution described by the other 236 normalized MOSs. In all 60 cases (i.e., 6 sessions × 10 HRCs), the Student's t test supported this hypothesis at the 5% level.

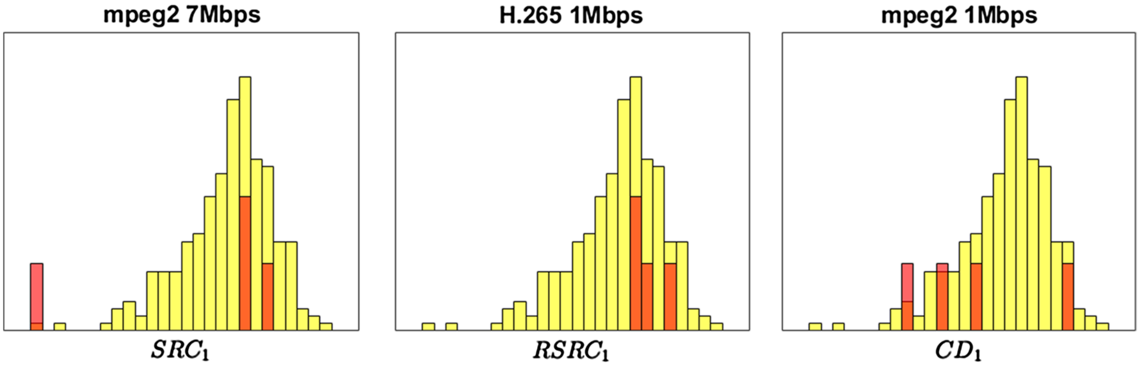

Fig. 9 Histogram of an emulated "typical" HRC (yellow) overlaid by an undesirable distribution from each experiment design (red). Each HRC's MOSs have been normalized for zero mean and unit variance

Basically, the conventional and unrepeated scene designs all produced MOSs that characterize the same set of HRCs. However, a visual examination of the normalized data indicates that none of the experiment designs did a good job of representing the "big picture." Many of these distributions are obviously biased. Figure 9 shows three of the worst cases. Part of the problem is simply that four sequences cannot represent "all video content," which we already know. Figure 3 of Pinson et al. [20] provides evidence and an explanation.

Impact of experiment design on conclusions

We want to compare the three designs based on the conclusions reached by the experiment. The problem is that we do not have ground truth data. All subjective experiments are influenced by design decisions (e.g., monitor, subject matter, range of quality) and severely limited in scope (e.g., number of codecs, bitrates, scene content, subjects). The conventional design has a symmetry and 50 year history that appeals to engineers. This does not prove validity or optimality.

We will assume as truth data the authors' a posteriori estimate of coding impairments and the MPEG committee's claim that each generation of codec yields equivalent quality at one third to one half the bitrate. This separates the AGH/NTIA/Dolby dataset into four quality levels:

- Original video
- 7 Mbit/s MPEG-2, 2 Mbit/s H.264, 1 Mbit/s H.265
- 4 Mbit/s MPEG-2, 1 Mbit/s H.264, 0.5 Mbit/s H.265
- 2 Mbit/s MPEG-2, 0.5 Mbit/s H.264, 0.25 Mbit/s H.265.

We will ignore pairs of HRCs where equivalent quality is expected (e.g., 2 Mbit/s H.264 and 1 Mbit/s H.265). An ideal experiment should be able to detect quality differences for all other HRC pairs (e.g., H.264 at 1 Mbits/s has higher quality than MPEG-2 at 2 Mbits/s).

This yields a set of HRC comparisons, which we will evaluate using the Student's *t* test. We do not correct the significant level [1], report specific *p* values or use more advanced statistical methods. Our goal is to validate if one experiment shows different conclusions than other if we use the same statistical method for both experiments. We believe that keeping this method simple makes the comparison easier to understand. The obtained data are made available, therefore it is possible to test more advanced data analysis methods.

Considering the HRC in our study, an ideal experiment would detect differences among 100% of these HRCs. Table 15a reports the ability of the Student's *t* test to discriminate between pairs of HRCs, computed and reported separately for each session.

Table 15b compares conclusions reached by different sessions. Columns **A** and **B** are the sessions to be compared. Column = lists the percent of differences not statistically significant or the same conclusions reached by both sessions. Column **A+** lists the percent of comparisons where session **A** is more sensitive (i.e., **A** detected a significant difference between paired HRCs but **B** did not). Column **B+** lists the percent of comparisons where session **B** is more sensitive. Column **error** lists the percent of comparisons where sessions **A** and **B** reach opposite conclusions, which would indicate a grievous error. None of the sessions reach opposite conclusions, despite the inadequate sampling of four scenes per HRC.

Likewise, none of the HRC comparisons reached the opposite conclusion to our truth data. That is, there were no cases where SRC, RSRC, or CD–SRC sessions showed differences between HRCs that we expected to have equivalent quality. If any of the sessions detected a significant difference in quality, those results agreed with our expectations.

Table 15 Ability to distinguish HRC quality differences

SRC_1	$RSRC_1$	CD_1	$RSRC_2$	CD_2	$Rand_2$
(a) HRC pairs differentiated with student's t test					
39%	64%	58%	83%	64%	67%
A	**B**	**=**	**A+**	**B+**	**Error**
(b) Impact of experiment design on the ability of student's t test to differentiate HRCs					
SRC_1	$RSRC_1$	75%	0%	25%	0%
SRC_1	CD_1	69%	6%	25%	0%
SRC_1	$RSRC_2$	56%	0%	44%	0%
SRC_1	CD_2	69%	3%	28%	0%
SRC_1	$Rand_2$	61%	6%	33%	0%
$RSRC_1$	$RSRC_2$	75%	3%	22%	0%
CD_1	CD_2	78%	8%	14%	0%

The results of Table 15a and the first row of Table 15b mean that 39% of the comparisons are statistically significant for session SRC_1 and they are also statistically significant in session $RSRC_1$. In addition 36% of the comparisons are not statistically significant for both sessions; hence 75% in column =. 25% of comparisons are statistically significant in session $RSRC_1$ and they are not statistically significant for session SRC_1; hence 25% in column **B+**. A single comparison is Student-*t* test of two different HRCs. For example, we compare results obtained for all sequences compressed with MPEG-2 7Mbit/s and H.265 0.25Mbit/s, for a specific session. For that comparison we expect that the MPEG-2 sequences have the statistically better quality.

From Table 15a, we see that the unrepeated scene designs show roughly a 50% improvement in ability to discriminate between HRCs versus the conventional design. This is encouraging but not conclusive. The identical SRC per HRC aspect of the conventional experiment design is easy to trust. This eliminates a degree of freedom (SRC variability) and simplifies the comparison of codec behavior. Unrepeated scene experiment designs do not have that quality.

A follow-on experiment is needed to compare and contrast the discrimination power of the conventional design and the unrepeated source design. A follow-on experiment would also help us prove whether the results in Table 15 reflect a more accurate estimation of the true HRC quality or random variations in the source material (not of interest). We also need a way to quantify when SRC are similar enough to be considered equivalent (for the purposes of measuring quality) and how many unrepeated SRC are needed to robustly characterize an HRC.

We must also consider that scene reuse may alter how subjects score videos and hide differences among HRCs. Recall the questionnaire responses, where some subjects described the scoring task as "easier" and "more accurate" with the conventional design. We know from the HSE analysis that this behavioral reporting of "accuracy" does not agree with our statistical measurement of precision (i.e., the scattering of scores around a MOS).

Perhaps the phenomenon perceived as "easier and more accurate scoring" instead reflects a change in how subjects think about quality and choose scores. Kahneman [14] explains that, when faced with difficult decisions or complex questions, people often substitute an easier question. This is so intrinsic to how we think, that people are not aware of the substitution. Scene reuse allows subjects to replace the complex judgment task ("What is the quality of this video?") with a simpler memory task ("Have I seen this video coded like that before? How did I score it?"). This substitute question would feel "accurate" because the scores are internally consistent.

In Table 15a, *Rand*2 is tied for second place. This indicates that the coding difficult algorithm is unnecessary, which is good news for people who must use an unrepeated scene experiment design. Randomly assigning scenes to HRCs appears to be as accurate as a carefully thought out heuristic.

Example experiment designs

We will now provide applied examples of experiment designs for common industry problems.

Let us first consider a service provider who wants to compare the quality delivered by their system with the quality delivered by their competitor's system. The unrepeated scene experiment design is mandatory: the competitor's processing chain cannot be accessed. The experimenter wants to limit the comparison to football games, because this content is important to customers, has high coding difficulty, and places real-time challenges upon the video production team.

The RSRC experiment design would be appropriate. The experiment design would identify exact scenes that typically appear in a football game (e.g., a close shot following fast action, a wide shot that shows most of the field, and a person talking with an overlay of game statistics). The experimenter would record one or more football games from each system, and find video segments with these characteristics. The experiment would enable a comparison between the two systems for an important demographic (football fans). The video sequences would have short durations (8–12 s) so that the experimenter could gain some insights into situations where each company's service is superior.

If instead the service provider wanted an overall comparison of the two systems, then the CD–SRC experiment design might be more appropriate. The experiment design might specify that the video sequences will be drawn at random from the ten most popular shows to play during a particular week, which will be different for the two systems. The random element would ensure unrelated SRCs and prevent the experimenter's opinions of the video content from biasing the experiment.

Let us now suppose the service provider is considering making a major change to their distribution chain. The company is considering seven options. The management team wants the system comparisons to be as realistic as possible. They do not want to choose a more expensive option if a less expensive option will supply acceptable video quality when customers consume their actual service. Unlike engineers, the management team places no value on direct comparisons of one scene for multiple coding options—they don't even want to see such data. Put simply, the management team wants an executive summary.

In this case, the unrepeated scene experiment design is desirable. The RSRC design would be preferable due to the lower cost of choosing and editing the videos. The list of RSRCs might include a particular music video, evening news commentators, a popular serialized show, a football game, and etc. The experiment design would include immersive elements, such as longer video sequences (e.g., 30 sec) and audio compressed according to their current distribution chain. This will help subjects remain entertained and engaged throughout the test. If the experiment design specified 20 types of scene content to be paired with their seven HRCs, the total experiment would contain 140 PVSs. Assuming a self-paced ACR test, each of the 24+ subjects would complete the test in less than two hours, and the large scene pool (20 RSRC) will robustly characterize the video provider's content. The immersive design will reduce the chances of an erroneous business decision.

Conclusion

In this paper, we compare three experiment designs:

- Conventional full matrix design (SRC × HRC)
- Related sequence design (RSRC × HRC)
- Coding difficulty design (CD–SRC × HRC.)

We conducted two subjective experiments that include all three experiment designs. We analyzed the scores for significant changes in scoring behaviors. Our goal is to understand the consequence of unrepeated scene experiment designs (i.e., where each subject views each SRC only once).

The conventional experiment design is a full factorial matrix of (SRC × HRC). Based on our analyses, it is plausible that some subjects change their scoring criteria over the course of a subjective test in response to viewing the same SRC multiple times. This demonstrates the drawback of the conventional (SRC × HRC) design.

The RSRC and CD–SRC experiment designs avoid repeated viewing of SRC. The RSRC design replaces each SRC with a set of visually similar content. The CD–SRC design replaces each SRC with a set of content with similar coding difficulty. This eliminates the option of performing comparisons between an individual SRC for different HRCs. When compared to the conventional design, our analysis indicates that unrepeated scene experiment designs are superior based on subjective feedback and equivalent based on score distributions (expected SOS).

We prefer the RSRC experiment design over the CD–SRC experiment design. The CD–SRC design is harder to implement, due to the high cost of obtaining a large variety of subject matter.

The unrepeated scene experiment designs find distinctions among HRCs that are not found by the conventional design. Our preliminary analysis indicates that the unrepeated scene experiment designs may be superior in ability to distinguish among HRCs. However, more research is needed to characterize the impact on MOS and HRC MOS when an experiment is designed around an (RSRC × HRC) matrix instead of an (SRC × HRC) matrix.

Studies of new technologies sometimes force researchers to use an unrepeated scene experiment design using a pool of diverse content. The CD–SRC design is suitable for such experiments and may have unproven advantages (see "Error analysis" section), but our coding difficult algorithm seems unnecessary. In this case we recommend a Random design, where a large set of SRC are randomly apportioned to HRCs (see "Impact of experiment design on conclusions" section).

This paper examines precision and stability, which are relatively easy to characterize. However, the goal of the unrepeated experiment design is to introduce a more realistic measure of HRC quality, potentially at the cost of decreased precision. This paper does not examine this more complex issue of whether unrepeated experiment designs do a better job of estimating the quality of a system, as it will be perceived by a large and diverse population of end users.

The philosophical question is how do we validate a method; and the need for an answer increases as new video services are introduced. The critical problem is not to propose modified methods, but to objectively determine which methods we can trust.

The approach of unrepeated signals was adopted quite some time ago by the speech quality assessment community. For example ITU-T Rec P.800 stipulates that a source sample should be presented only once to the subject, especially for the assessment of Listening Effort [10]. Experiments designed for speech quality assessment are also very similar to the related source design (RSRC) as the sources are typically sentences spoken by a limited set of talkers, typically 4 to 8. Each talker can be viewed as a "scene" as each spoken sentence is different for each HRC but the voice characteristics remain consistent. Subjects become familiar with the voice of each talker as the test progress. Speech quality assessment experiments also present analogies to the coding difficulty design (CD–SRC) as the sentences spoken by each talker are typically taken from the list of Harvard sentences [26].

Similarly to video coding, speech coding may deliver variable quality depending on the complexity of the input. Harvard sentences provide phonetically balanced sets of sentences which are used to expose systems under test to a controlled and balanced set of sounds.

The speech quality assessment community also uses the balanced block experiment design which groups the subjects into different panels where each panel assesses the same set of HRCs but different stimuli [9]. This approach leads to fewer scores per stimulus but it enables the evaluation of each HRC with more sources, providing a more holistic assessment of the systems under test. With this type of design, the analysis is usually performed per HRC rather than per PVS. The potential application of the balanced block design to video quality assessment is a subject for further study.

Open data

This paper uses data from two subjective experiments: AGH/NTIA and AGH/NTIA/Dolby. These dataset are now available on the Consumer Digital Video Library (CDVL, www.cdvl.org).

Acknowledgements The Lucjan Janowski's work was supported by the Polish Ministry of Science and Higher Education with the subvention funds of the Faculty of Computer Science, Electronics and Telecommunications of AGH University (Grant No. 15.11.230.075).

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Brunnström K, Barkowsky M (2018) Statistical quality of experience analysis: on planning the sample size and statistical significance testing. J Electron Imaging 27:11–27
2. den Broeck WV, Jacobs A, Staelens N (2012) Integrating the everyday-life context in subjective video quality experiments. In: 2012 fourth international workshop on quality of multimedia experience, pp 19–24
3. Fenimore C, Libert J, Wolf S (1998) Perceptual effects of noise in digital video compression. In: 140th SMPTE technical conference and exhibit, pp 1–17
4. Frohlich P, Egger S, Schatz R, Muhlegger M, Masuch K, Gardlo B (2012) Qoe in 10 seconds: are short video clip lengths sufficient for quality of experience assessment? In: 2012 fourth international workshop on quality of multimedia experience (QoMEX), pp 242–247
5. Hoffmann H, Itagaki T, Wood D, Hinz T, Wiegand T (2008) A novel method for subjective picture quality assessment and further studies of HDTV formats. IEEE Trans Broadcast 54(1):1–13
6. Hossfeld T, Schatz R, Egger S (2011) Sos: the MOS is not enough! In: 2011 third international workshop on quality of multimedia experience, pp 131–136
7. Hossfeld T, Hirth M, Redi J, Mazza F, Korshunov P, Naderi B, Seufert M, Gardlo B, Egger S, Keimel C (2014) Best practices and recommendations for crowdsourced QoE—lessons learned from the qualinet task force crowdsourcing. Technical report, QUALINET
8. Hoßfeld T, Biedermann S, Schatz R, Platzer A, Egger S, Fiedler M (2011) The memory effect and its implications on web QoE modeling. In: 2011 23rd international teletraffic congress (ITC), pp 103–110
9. International Telecommunication Union (2011) Practical procedures for subjective testing. ITU-T handbook
10. ITU-T Recommendation (1996) P.800: methods for subjective determination of transmission quality. Technical report P.800. International Telecommunication Union, Geneva
11. ITU-T Recommendation (2015) ITU-T P.913: methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment. Technical report P.913. International Telecommunication Union, Geneva
12. ITU-T Recommendation (2017) ITU-T P.10: vocabulary for performance, quality of service and quality of experience. Technical report P.10/G.100. International Telecommunication Union, Geneva

13. Janowski L, Pinson M (2015) The accuracy of subjects in a quality experiment: a theoretical subject model. IEEE Trans Multimed 17(12):2210–2224
14. Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux, New York
15. Mercer Moss F, Wang K, Zhang F, Baddeley R, Bull DR (2016) On the optimal presentation duration for subjective video quality assessment. IEEE Trans Circuits Syst Video Technol 26(11):1977–1987. https://doi.org/10.1109/TCSVT.2015.2461971
16. Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychol Rev 63:81–97
17. Pinson M, Janowski L (2014) AGH/NTIA: a video quality subjective test with repeated sequences. Technical report. NTIA Technical Memorandum TM-14-505
18. Pinson MH (2016) Why is no-reference model development failing? VQEG meeting in London Oct 2016, ftp://vqeg.its.bldrdoc.gov/Documents/VQEG_London_Oct16/MeetingFiles/VQEG_VIME_2016_117_NR%20model%20development.pptx. Accessed 21 June 2019
19. Pinson MH, Wolf S, Cermak G (2010) HDTV subjective quality of h.264 versus mpeg-2, with and without packet loss. IEEE Trans Broadcast 56(1):86–91. https://doi.org/10.1109/TBC.2009.2034511
20. Pinson MH, Barkowsky M, Le Callet P (2013) Selecting scenes for 2d and 3d subjective video quality tests. EURASIP J Image Video Process 2013(1):50
21. Pinson MH, Sullivan M, Catellier AA (2014) A new method for immersive audiovisual subjective testing. In: Eighth international workshop on video processing and quality metrics for consumer electronics (VPQM 2014)
22. Pinson MH, Janowski L, Papir Z (2015) Video quality assessment: subjective testing of entertainment scenes. IEEE Signal Process Mag 32(1):101–114
23. Raake A, Egger S (2014) Quality and quality of experience. In: Möller S, Raake A (eds) Quality of experience: advanced concepts, applications and methods. Springer, Cham, pp 11–33
24. Ribeiro F, Florencio D, Zhang C, Seltzer M (2011) Crowdmos: an approach for crowdsourcing mean opinion score studies. In: 2011 IEEE international conference on, acoustics, speech and signal processing (ICASSP), pp 2416–2419
25. Robitza W, Garcia MN, Raake A (2015) At home in the lab: assessing audiovisual quality of http-based adaptive streaming with an immersive test paradigm. In: 2015 seventh international workshop on quality of multimedia experience (QoMEX), pp 1–6
26. Rothauser HE (1969) IEEE recommended practice for speech quality measurements. IEEE Trans Audio Electroacoust 17:225–246
27. Schmuckler M (2001) What is ecological validity? A dimensional analysis. Infancy 2(4):419–436
28. Sullivan M, Pratt J, Kortum P (2008) Practical issues in subjective video quality evaluation: human factors versus psychophysical image quality evaluation. In: Proceedings of the 1st international conference on designing interactive user experiences for TV and Video, ACM, New York, NY, USA, UXTV ’08, pp 1–4. https://doi.org/10.1145/1453805.1453807
29. Tavakoli S, Egger S, Seufert M, Schatz R, Brunnstrom K, Garcia N (2016) Perceptual quality of HTTP adaptive streaming strategies: cross-experimental analysis of multi-laboratory and crowdsourced subjective studies. IEEE J Sel Areas Commun 34(8):2141–2153
30. Tominaga T, Hayashi T, Okamoto J, Takahashi A (2010) Performance comparisons of subjective quality assessment methods for mobile video. In: 2010 second international workshop on quality of multimedia experience (QoMEX), pp 82–87
31. Wuensch KL (2019) Comparing correlation coefficients, slopes, and intercepts. Technical report, East Carolina University. http://core.ecu.edu/psyc/wuenschk/docs30/CompareCorrCoeff.pdf. Accessed 21 June 2019

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.