PREMIUM HDR: THE IMPACT OF A SINGLE WORD
ON THE QUALITY OF EXPERIENCE OF HDR VIDEO

Peter A. Kara, Maria G. Martini

WMN Research Group, Kingston University
{p.kara, m.martini}@kingston.ac.uk

Aron Cserkaszky

Pazmany Peter Catholic University
cserkaszky.aron@itk.ppke.hu

ABSTRACT

High Dynamic Range (HDR) visualization is converging towards a point where it becomes the de facto standard for multimedia services. Current and future variations in color bit depth, peak brightness, black level and all other characteristics of display and content ultimately lead to a fine-grained market of end-user products with a multitude of different capabilities and also numerous quality standards. Such identifiers and indicators may influence the Quality of Experience (QoE), as they create expectations and preconceptions regarding visual quality. In fact, sometimes even a single word may fundamentally affect user experience. In this paper, we investigate the labeling effect on HDR video QoE. We carried out a series of subjective tests where test participants compared unimpaired HDR videos. The trigger of the labeling effect was that one of the stimuli in the paired comparison was identified as Premium HDR. In order to gain more insight into the biased perception and experience, subjective assessment was separately performed for the different quality aspects of HDR visualization, such as luminance or color.

Index Terms— High Dynamic Range, Quality of Experience, labeling effect, cognitive bias

1. INTRODUCTION

The Human Visual System (HVS) is the axis around which modern visualization technologies revolve. Whether we talk about spatial resolution, frame rate or any parameter of a display or a visual content, the primary goal in research and development is to reduce the gap between digital visualization and the capabilities of the HVS. A common scenario is that the human observer can perceive more than what can be digitally provided, and the long-term scientific objective is to match the limits of such biological sensory system. However, the opposite case is frequent as well, when certain limitations of the HVS can be exploited, i.e., for data compression, thus making perceptual coding possible.

The dynamic range of the HVS is one aspect where conventional capture and visualization technologies – referred to as Low Dynamic Range (LDR) or Standard Dynamic Range (SDR) – are heavily outranked. It is the typical case of an evident, undeniable difference between what can be captured and displayed, and what can be perceived by a person in real life. The recent advances of High Dynamic Range (HDR) motion imaging enable a contrast range in digital visualization that is close to the capabilities of the HVS, and may even exceed them. The high difference between the peak brightness and the black level, combined with the richer range of colors, enables HDR to deliver a much more life-like visual appearance, compared to the technological predecessors.

At the time of this paper, HDR visualization already appears in multiple forms, and related standards emerge continuously. In the commercial sector, the standards of HDR10, its update HDR10 Plus and Dolby Vision are competing, but the ETSI SL-HDR1 standard (originating from Technicolor, STMicroelectronics, and Philips) should be mentioned as well. The related ITU-R standard is Rec. BT.2100[1], which uses wide color gamut according to the Rec. BT.2020 color space[2], supports 10-bit and 12-bit colors, covers both HD and UHD spatial resolution and both Perceptual Quantizer (PQ) and Hybrid Log-Gamma (HLG) transfer function.

Although HDR visualization may greatly enhance the Quality of Experience (QoE), manufacturers boldly labeling their displays as "HDR" could severely damage the perception of the added value and thus result in disappointed users. In order to prevent this from happening, the Ultra HD Alliance defined strict requirements for what can be certified as HDR with the Ultra HD Premium logo. The given display needs to have a peak brightness of at least 1,000 nits, and black level must be lower than 0.05 nits, where the non-SI unit nit is defined as the luminous intensity per unit area of light traveling in a given direction, and 1 nit can be approximated as the light emitted from a single candle. As OLED televisions cannot reach a peak brightness of 1,000 nits, their re-

[1]Rec. BT.2100: Image parameter values for high dynamic range television for use in production and international programme exchange

[2]Rec. BT.2020: Parameter values for ultra-high definition television systems for production and international programme exchange

quirements were specially reduced to 540 nits for peak brightness, and 0.0005 nits for black level, in order to compensate. In comparison, the peak brightness of SDR LED televisions are usually in an interval between 300 and 500 nits.

Needless to say, the difference in visual experience between watching a regular LED TV and an HDR10-certified TV displaying HDR demo (eye candy) content can be staggering. In the upcoming years, numerous further developments in manufacturing and standardization are expected in the field of HDR visualization. It is sufficient to consider the fact that on the level of specifications, Dolby Vision is superior to the current HDR10 and HDR10 Plus standards. As an example, while HDR10 only supports peak brightness up to 4,000 nits, this value is 10,000 in case of Dolby Vision. Also, not every manufacturer is aiming to achieve the HDR certification of Ultra HD Premium for their displays – as it is not mandatory – and some use their own terminology, e.g., "HDR Pro" in case of LG.

The potential future diversity in HDR visualization technology and its corresponding labeling shall provide a vast variety in both visual quality and device information. Let us consider the classic scenario where the end user goes to a shop, observes the displays and their lists of parameters (including their prices), and selects one for purchase. There is an inevitable 2-way effect between the perceived visual quality and the listed parameters: (a) the list of information may affect the way the user perceives the quality, e.g., being aware of a higher bit depth may enhance the experienced difference regarding the actual difference in colors, and (b) the perceived quality may affect the interpretation of the information, e.g., seeing something that is very appealing visually may justify a higher cost or increase the subjective relevance of a technical parameter.

In this paper, we investigate the first case, where the related information affects the perceived quality. More precisely, we carried out an experiment where we focused on the effects of a single, non-technical word on selected QoE aspects of HDR video quality. Test participants were shown identical video stimulus pairs of different source contents in a paired comparison, but one of the stimuli was labeled as a simple "HDR" video, and the other one as a so-called "Premium HDR"; the latter suggested the provision of superior visual quality in comparison to the other. The research question addressed in this paper targeted the cognitive bias evoked by this specific label.

The remainder of the paper is structured as follows: Section 2 introduces the related technical literature on the QoE of HDR visualization, and also a brief description of the labeling effect and its consideration in previous studies. Section 3 details the configuration and parameters of the series of subjective tests carried out, and Section 4 analyzes the obtained results. The paper is concluded in Section 5.

2. RELATED WORK

As HDR visualization is considered to be the next big step in consumer-grade home TV entertainment, its QoE aspects have been and currently are still being extensively investigated. The works of Narwaria *et al.* [1] [2] [3] address HDR QoE, taking into consideration immersion, the natural feeling of the visualized content, visual attention and many more aspects, while also discussing subjective measurement methodologies. The authors particularly investigated tone mapping operators (TMOs) and how they affect the perception of HDR content, and also proposed a novel objective video quality metric for HDR [4].

Trivially, the major added value of HDR visualization from a QoE perspective originates from the *high* dynamic range itself. However, measuring the dynamic range perceived by test participants is quite far from being a trivial task. The work of Hulusic *et al.* [5] introduces a subjective measurement methodology for the perceived dynamic range. The authors carried out a series of subjective tests with 20 test participants, in which HDR images (photographs and video frames) from various sources (e.g., Fairchilds HDR Photographic Survey [6], the Stuttgart HDR Video Database [7], etc.) were assessed on a Full HD (1920×1080) SIM2 HDR display, namely the HDR47ES4MB. All still image stimuli were converted to grayscale, as the research solely focused on the perceived dynamic range. The test participants had to evaluate "the overall impression of the difference between the brightest and the darkest part(s) in the image" using a variation of the Subjective Assessment Methodology for Video Quality (SAMVIQ) [8]. The ratings were collected on a continuous scale (from 0 to 100), which was divided into 5 labeled, uniform intervals ("*very low*", "*low*", "*medium*", "*high*" and "*very high*"). The findings highlight the importance of content characteristics, such as the relative surface of bright areas and the distance, the separation between dark and bright areas.

Although one of the key features of HDR visualization is the higher level of brightness, having a screen that is too bright might not be preferable by the end user. The work of Bist *et al.* [9] proposes a content-based method for brightness control, based on subjective studies of brightness preference. The algorithm operates on a pixel-level; the "bright" pixels of the visualized content are taken into consideration during brightness adjustment, which means that the larger the portion of bright areas on the screen, the lower the level of brightness that shall be set. In their experiment, 16 test participants viewed static images on a SIM2 HDR47ES4MB HDR display, the brightness of which they had to re-adjust in case they found them too bright.

Using physiology in QoE studies is a very well-known approach within the scientific community [10]. Depending on the methodology, subjective tests may provide an immense amount of useful information regarding the personal qual-

ity preferences and the specific perceptual thresholds of the test participant; however, opinion scores do not report anything about the internal physiological levels of the individual. The work of Al-juboori *et al.* [11] uses electroencephalogram (EEG) to analyze the correlation between the perceived quality of HDR images and the different bands of brain activity. The 4 TMOs were applied to the 5 source HDR images and the 20 stimuli were shown to the 28 test participants on an iPhone 6. The results highlight the emotions that were induced by the visualized content, as they correlate with the acquired EEG signals.

As for the labeling effect, its presence has been investigated in QoE, as labels are pieces of information that characterize the target of observation, and thus may easily occur in real-life scenarios. Such scenarios often involve financial decisions, and therefore labels may affect the user's willingness-to-pay (WTP). Several works of Sackl *et al.* [12] [13] address WTP from the angle of QoE, and also take the labeling effect into consideration [14] [15] [16]. In these works, the label is the type of connection to the access network (e.g., wireless). Similar experiments were also carried out by Kara *et al.* [17] [18], and in a joint research effort, the effect of the mobile device brand was investigated as well [19].

The majority of these studies were performed in so-called "mock-up" experiments, which means that the information provided to the test participant and the real experimental setup differed. Also, the test participant was not made aware of the true nature of the study, not even after the subjective task. In many cases of mock-up experiments – especially when the labeling effect is investigated – the test participant evaluates stimuli that "only differ on paper", but are identical in practice. This means that the label is the only variable, making the phenomenon viable to measure.

In a recent study, Kara *et al.* [20] investigated the labeling effect for HD and UHD/4K video. A paired comparison method was used in the subjective tests, comparing the video stimuli on 3-point and 7-point comparison scales[3]. Before each stimulus, the spatial resolution (HD or UHD) was shown on the screen. However, these labels were designed to be misleading in many test conditions, e.g., the labels suggested one video to be HD and the other one in the pair to be UHD, yet both video stimuli had the same HD resolution and thus were identical. The results indicate statistically significant differences in evaluation scores, demonstrating the effect of the investigated cognitive bias.

In the experiment presented in this paper, we also designed our tests with mock-up methodologies. The following section details the selected parameters of the subjective tests.

3. EXPERIMENTAL SETUP

3.1. Display and Environment

The subjective tests were performed in an isolated, controlled laboratory environment, with dimmed lighting conditions. The ambient luminance was nearly 10 lux and not lower, in order to avoid visual discomfort [9]. The test participants viewed the HDR videos on a SIM2 HDR47ES6MB HDR display[4], with peak brightness over 6000 nits.

The viewing angle was zero degree (center view) during the entire test, and the viewing distance was a fixed 3H according to the recommendation[5] – as Full HD (1920×1080) content was displayed on the full screen of the 47 Full HD display – which corresponded to 1.75 meters.

3.2. Test Protocol

The test itself was a paired comparison, which compared video stimuli on a 7-point comparison scale ("*Much Worse*", "*Worse*", "*Slightly worse*", "*Same*", "*Slightly better*", "*Better*", "*Much better*"). In order to gain a more detailed insight into the cognitive bias created by the labeling effect, instead of comparing the overall QoE, the test participants had to assess 4 aspects of HDR video quality: *luminance*, *frame rate*, *color* and *image quality*.

Before the subjective test, the test participant received training, during which the four aforementioned aspects were interpreted and demonstrated. *Luminance* was described as the perceived difference between the brightest and the darkest portions of the screen; greater difference was to be evaluated better. Although *frame rate* was considerably self-explanatory, it was still explained to every participant in order to avoid confusion and misunderstandings. *Color* was interpreted as the richness, the depth of the colors on the screen. Lastly, *image quality* was approached from the angle of spatial resolution and classic coding artifacts, independently from the three other aspects. Also, before the training phase, test participants were screened for normal vision using the Snellen charts and Ishihara plates.

The double stimulus method was used, with the stimuli in a pair shown after each other. They were separated by a 5-second blank screen, and comparison was performed directly after each pair, in a time window of 10 seconds. The stimulus pairs were also separated by a 5-second blank screen.

As detailed in Figure 1, for a given content i – where i is a content identifier between 1 and 10, corresponding to the source order randomized for each participant – the first instance of the content (VA_i) is played, followed by the stimulus separation (S_i), and then the identical second instance (VB_i) is shown. After this, VB_i is compared to VA_i in the

---

[3]Rec. BT.500: Methodology for the subjective assessment of the quality of television pictures

[4]SIM2 HDR47ES6MB display: http://hdr.sim2.it/hdrproducts/hdr47es6mb
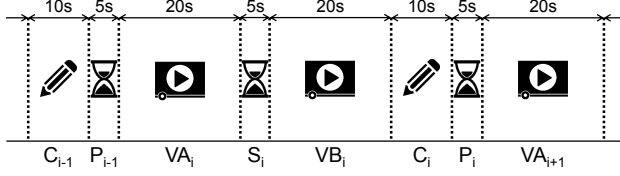[5]Rec. BT.710: Subjective assessment methods for image quality in high-definition television

Fig. 1. Temporal structure of the subjective test. The structure of the test for content i is demonstrated, where i is at least 2 and at most 9, so the denotation of $i-1$ and $i+1$ are valid.

comparison period (C_i), and finally the separation screen between the stimuli pairs is displayed (P_i). As this given structure is repeated over the duration of the subjective test, if i is at least 2 but at most 9 (i.e., neither the first not the last pair), then VA_i occurs directly after the comparison period and the separation screen of the prior content $i-1$ (C_{i-1} and P_{i-1}, respectively), and P_i is followed by the first instance of the subsequent content $i+1$ (VA_{i+1}).

The assignment of the so-called "Premium HDR" video varied between test participants. For half of the participants, the "Premium" video was always the first one in the pair (VA), and for the other half, it was the second one (VB). Again, this means that for each and every test participant, the assignment of the label was consistent and did not change during the test. As the labeling effect can influence both perception and the memory of perception, we decided to include this given division between the test participants in order to investigate the role of label order. However, the detailed analysis of the measured effect of the label order is out of the scope of this paper.

3.3. Source Contents and Test Stimuli

The experiment involved 10 source video contents from the Stuttgart HDR Video Database [7]. The sources in alphabetical order were *Beerfest Lightshow*, *Bistro*, *Carousel Fireworks*, *Cars Longshot*, *Fireplace*, *Fishing Longshot*, *Poker Fullshot*, *Poker Travelling Slowmotion*, *Showgirl 1* and *Smith Welding*.

From each source sequence, an unimpaired 20-second clip was cut. As the duration of the videos were 20 seconds, each segment of the test structure was 60 seconds long (see Figure 1). The stimuli were 10-bit videos, with $Y'C_BC_R$ color space and 4:2:0 chroma subsampling.

Each and every video had a constant frame rate of 25 fps. In a stimulus pair, technically the exact same video was shown twice (see Figure 1). As there were 10 source video contents and each stimulus was shown twice, a total of 20 videos were presented to the test participants. As stated earlier, the order of the contents was randomized for each test participant. The total duration of the subjective test was approximately 10 minutes, so a single session was held per participant, without any breaks.

As the experiment solely focused on visual quality, the option of audio was excluded from the tests. The video stimuli inherently contained no audio, and the display itself emitted no sound from its speakers during the test.

3.4. Test Participants

A total of 40 individuals participated in the tests; 30 males and 10 females. The age interval was from 20 to 56, and the average age was 30. 10 participants had prior HDR video experience, and the rest had never seen any HDR video before the experiment.

4. RESULTS

In this section, the obtained subjective scores are represented by their numerical counterparts, ranging from -3 to $+3$. During the subjective tests, the test participants were presented a combination of the available qualitative tags for stimulus comparison – defined in the previous section – and these values, emphasizing a uniform distance between the values of the scale. In this analysis, positive values favor the "Premium HDR" stimulus, while negative values indicate that it was deemed to be worse in the given aspects.

Each of the 40 test participants compared 4 quality aspects of 10 stimulus pairs, thus we collected 1600 subjective scores in our experiment. In an ideal scenario without the existence of cognitive bias through the labeling effect, all these 1600 scores would have reported the given aspects to be the "*Same*". However, according to the scoring distribution, only 356 (22.25%) of them were zero, and 1244 (77.75%) assessed a certain level of either positive (1089 scores) or negative (155 scores) difference (see Figure 2).

The most frequent quality comparison score was "*Slightly better*", followed by "*Better*", "*Same*", "*Much better*", "*Slightly worse*", "*Worse*" and "*Much worse*". This order took all of the investigated aspects into consideration. If we separate them, we can observe rather similar mean values for *luminance*, *color* and *image quality* (see Figure 3). In fact, the aforementioned order in score frequency applied to all three of them (see Figure 4), and there was no statistically significant difference between them.

However, *frame rate* was assessed differently. The mean score was significantly lower compared to the other aspects, as the number of positive scores was the lowest, while it received the most zero and negative scores. Moreover, the number of negative scores *frame rate* received was near to what the other three aspects had combined.

It needs to be noted that more than half (201 out of 400) of the scores for *frame rate* were indeed positive, meaning that the test participants providing those scores experienced an improvement in this aspect for the stimuli with "Premium HDR" quality. Yet there were many who either did not perceive a change in *frame rate* or experienced degradation.
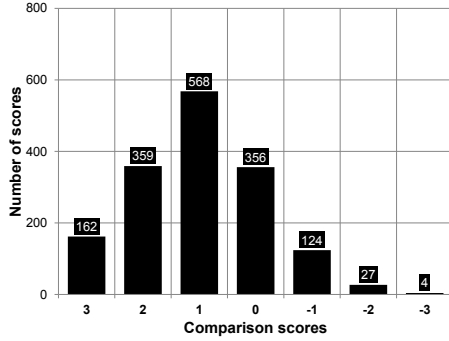
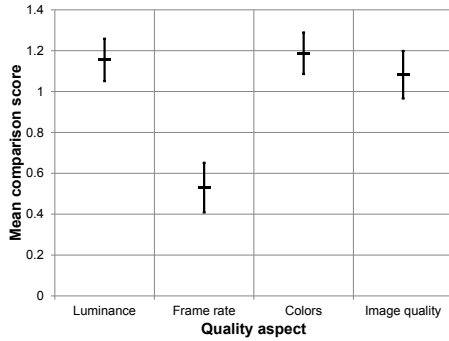**Fig. 2**. Scoring distribution of the subjective tests.

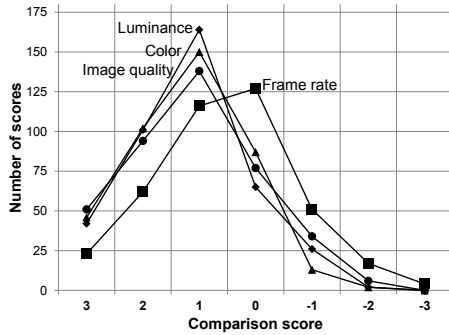**Fig. 3**. Mean comparison scores with 0.95 CI.

Fig. 4. Scoring distribution of the quality aspects.

Although the experimental setup did not define any feedback beyond the comparison scores, some test participants provided us valuable insights to their visual experience. One of the test participants, who works in the movie industry, claimed that

> "The first version (Premium HDR) is always more pushed to the limits; it's actually more magical, but less controlled. The second one (HDR) feels more controlled, less magic. Personally I would go for a middle path. The frame rate doesn't seem to improve significantly."

We also had test participants who consistently experienced frame rate drops in the "Premium HDR" videos, while perceiving improvements in the other aspects. Their comparison patterns can be summarized by the following feedback:

> "It is such a pity that these incredible visuals come at the expense of frame rate. Yet to be fair, it is most certainly worth it."

These cognitions originated from the concept of compensation, the idea of balance; if certain aspects become better, then their improvements negatively affect the performance of others. One could suggest that such bias might be limited to test participants with educational backgrounds of engineering or computer science, but these patterns appeared randomly within the observer population. The impressive visuals of HDR compared to regular SDR or LDR TV experience are easier to connect with a "premium" quality when it comes to *luminance*, *color* and even *image quality*, compared to a *frame rate* of 25 fps, when 60 fps is spreading in the everyday use case scenarios. Also, from the three highlighted aspects, *image quality* received the least positive and the most negative scores, even though it was not statistically different from the other two. Having the same experiment in UHD resolution is expected to boost this aspect in the positive direction.

Regarding the effect of the label order, no significant difference was found between the mean scores of the two groups, and the general findings applied to this scoring separation as well.

5. CONCLUSIONS

In this paper, we presented how the labeling effect influences the quality aspects of HDR videos. The obtained subjective test results indicate that more than 75% of the comparison scores were affected by the presence of the label "Premium HDR". The perceived frame rate of the video stimuli was the least affected, yet we also conclude that several test participants experienced frame rate degradations, while improvements were observed for the other aspects. This can be explained with cognitions related to the concept of overall balance; the enhancement of specific features must come at the expense of other features. In future works, we aim to address waiting times for HDR video playback and their susceptibility, their exposure to the effects of cognitive bias.

6. REFERENCES

[1] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet, "High dynamic range visual quality of experience measurement: Challenges and perspectives," in *Visual Signal Quality Assessment*, pp. 129–155. Springer, 2015.

[2] Manish Narwaria, Matthieu Perreira Da Silva, Patrick Le Callet, Giuseppe Valenzise, Francesca De Simone, and Frédéric Dufaux, "Quality of experience and HDR: concepts and how to measure it," in *High Dynamic Range Video*, pp. 431–454. Elsevier, 2016.

[3] Manish Narwaria, Matthieu Perreira Da Silva, Patrick Le Callet, and Romuald Pepion, "Tone mapping based HDR compression: Does it affect visual experience?," *Signal Processing: Image Communication*, vol. 29, no. 2, pp. 257–273, 2014.

[4] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet, "HDR-VQM: An objective quality measure for high dynamic range video," *Signal Processing: Image Communication*, vol. 35, pp. 46–60, 2015.

[5] Vedad Hulusic, Giuseppe Valenzise, Edoardo Provenzi, Kurt Debattista, and Frederic Dufaux, "Perceived dynamic range of HDR images," in *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2016, pp. 1–6.

[6] Mark D Fairchild, "The HDR photographic survey," in *Color and Imaging Conference*. Society for Imaging Science and Technology, 2007, vol. 2007, pp. 233–238.

[7] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel, "Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays," in *Digital Photography*, 2014, p. 90230X.

[8] Jean-Louis Blin, "SAMVIQ – Subjective assessment methodology for video quality," *Rapport technique BPN*, vol. 56, pp. 24, 2003.

[9] Cambodge Bist, Rémi Cozot, Gérard Madec, and Xavier Ducloux, "QoE-based brightness control for HDR displays," in *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017, pp. 1–6.

[10] Ulrich Engelke, Daniel P Darcy, Grant H Mulliken, Sebastian Bosse, Maria G Martini, Sebastian Arndt, Jan-Niklas Antons, Kit Yan Chan, Naeem Ramzan, and Kjell Brunnström, "Psychophysiology-based QoE assessment: a survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 6–21, 2017.

[11] Shaymaa Al-Juboori, Is-Haka Mkwawa, Lingfen Sun, and Emmanuel Ifeachor, "Investigation of relationships between changes in EEG features and subjective quality of HDR images," in *International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 91–96.

[12] Andreas Sackl, Sebastian Egger, Patrick Zwickl, and Peter Reichl, "The QoE alchemy: Turning quality into money. Experiences with a refined methodology for the evaluation of willingness-to-pay for service quality," in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2012, pp. 170–175.

[13] Andreas Sackl, Patrick Zwickl, and Peter Reichl, "From Quality of Experience to Willingness to Pay for Interconnection Service Quality.," in *Networking Workshops*. Springer, 2012, pp. 89–96.

[14] Andreas Sackl, Raimund Schatz, and Alexander Raake, "More than I ever wanted or just good enough? User expectations and subjective quality perception in the context of networked multimedia services," *Quality and User Experience*, vol. 2, no. 1, pp. 3, 2017.

[15] Andreas Sackl, Kathrin Masuch, Sebastian Egger, and Raimund Schatz, "Wireless vs. wireline shootout: How user expectations influence Quality of Experience," in *Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2012, pp. 148–149.

[16] Andreas Sackl, Patrick Zwickl, Sebastian Egger, and Peter Reichl, "The role of cognitive dissonance for QoE evaluation of multimedia services," in *Globecom Workshops*. IEEE, 2012, pp. 1352–1356.

[17] Peter A. Kara, László Bokor, and Sándor Imre, "Distortions in QoE assessment of 3D multimedia services on multi-access mobile devices," in *Ninth International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2013, pp. 311–318.

[18] Peter A. Kara, László Bokor, and Sándor Imre, "Seeing is believing and vice versa: Investigation of the altered perception during subjective assessment of streaming multimedia," in *Tenth International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*. IEEE, 2014, pp. 539–545.

[19] Peter A. Kara, László Bokor, Andreas Sackl, and Mariana Mourão, "What your phone makes you see: Investigation of the effect of end-user devices on the assessment of perceived multimedia quality," in *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2015, pp. 1–6.

[20] Peter A. Kara, Werner Robitza, Alexander Raake, and Maria G. Martini, "The Label Knows Better: The Impact of Labeling Effects on Perceived Quality of HD and UHD Video Streaming," in *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, 2017.