



Comparison of HD and UHD video quality with and without the influence of the labeling effect

Peter A. Kara¹ · Werner Robitza² · Nikolett Pinter³ · Maria G. Martini¹ · Alexander Raake² · Aniko Simon⁴

Received: 9 February 2019
© The Author(s) 2019

Abstract

The past decades have shown a steady rise in the pixel resolution of digital visualization technologies. New TVs and computer displays available on the consumer market are commonly UHD/4K, and 8K recently started appearing in the commercial world. The added value of such higher resolution is not only indicated by the purported level of visual performance, but also by the labels that emphasize the devices' properties, especially the resolution. However, the genuinely perceived difference between two resolutions, such as HD and UHD, may not have the same effect on user experience as the labels attached to the devices or sequences themselves. In this paper we present four subjective studies in which we investigated the influence of the labeling effect in the context of HD and UHD video. Subjects were shown pairs of either HD or UHD video sequences and had to determine whether there was a difference in quality between the sequences. Two of the subjective tests involved labels in these paired comparisons, indicating to the subject which sequence was which, while the other two tests excluded these labels. For both test sets with and without labels, one test used a 3-point comparison scale for the subjective assessment of visual quality, while the other used a more fine-grained 7-point scale. Our findings show that the sole inclusion of quality labels can strongly impact subjective rating behavior and the overall opinion on UHD quality; also, visual differences between HD and UHD video were rarely noticeable by the subjects.

Keywords Quality of experience (QoE) · Video quality · Video resolution · UHD/4K · Labeling effect

The work in this paper is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 643072, Network QoE-Net.

✉ Peter A. Kara
p.kara@kingston.ac.uk
Werner Robitza
werner.robitza@tu-ilmenau.de
Nikolett Pinter
nikolett.pinter@bm.gov.hu
Maria G. Martini
m.martini@kingston.ac.uk
Alexander Raake
alexander.raake@tu-ilmenau.de
Aniko Simon
aniko.simon@sigmatechnology.se

- ¹ WMN Research Group, Kingston University, London, UK
- ² Audiovisual Technology Group, TU Ilmenau, Ilmenau, Germany
- ³ Hungarian Ministry of Interior, Budapest, Hungary
- ⁴ Sigma Technology, Budapest, Hungary

Introduction

Background

The technical term “UHD”—referring to ultra-high-definition displays and contents—has entered our everyday lives in the past decade, and at the time of writing this paper, it is slowly becoming a more common format of multimedia consumption on TVs, tablets, computer screens, cinema, and other technologies. The rise of UHD content is enabled by the fact that more and more UHD-capable displays emerge on the consumer market (according to recent studies, institutes expect “the number of households with a 4K TV to grow from 28 million in 2015 to 335 million in 2020”,¹ and that “222 million homes will own an Ultra HD TV by the end of 2018”²), thus creating a vigorous competition which continuously reduces prices—especially entry-level

¹ <https://technology.ihc.com/609782/4k-tv-and-uhd-the-whole-picture/>.

² <https://www.broadbandtvnews.com/2018/12/09/200mi-homes-have-an-ultra-hd-tv-set/>.

prices—making such displays available to a wider range of consumers. Also, content creation and provision on amateur and professional levels shift towards UHD resolution, including real-time streaming services.

By now, we can state that most multimedia consumers have come across the term UHD in one way or another, even if they have not experienced true UHD content on a UHD display yet. These three letters are found highlighted on stickers and labels on displays in shops; they are emphasized in commercials on TV and the Internet, and content providers promote this attribute whenever they can, particularly when selling UHD on top of existing HD programming. At the same time, the first demos of “8K” entertainment systems are emerging.

UHD in the context of home entertainment is often labeled “4K”. To be more accurate, UHD TV can either refer to “UHD-1” (3840 × 2160 pixels) or “UHD-2” (7680 × 4320 pixels), standardized by ITU-R Rec. BT.2020. The formats named “4K” (4096 × 2160) and “8K” (8192 × 4320) are standardized for cinema, defined by the Digital Cinema System Specification. In practice, UHD/4K commonly refers to the resolution of 3840 × 2160 pixels, as that is the resolution with which UHD TVs are shipped to consumers. In the scope of this paper, we address this specific resolution and compare it with Full HD (1920 × 1080).

Motivation

The research in this paper addresses the perceived differences between HD and UHD video while taking into consideration the labeling effect, that is, the effect a certain label (like “UHD”) may have on users’ opinions or decisions. The core of our work investigates the visual quality achieved by these resolutions, which plays an important role for the overall Quality of Experience (QoE) of a user. As we studied perceived differences, one could assume that the authors intend to question the added value of UHD video compared to HD, or that they doubt the relevance of the presence of UHD-capable displays on the market. This, however, is not the purpose of this research.

Our work is mainly motivated by the appearance and the usage of the term “UHD” on commercial levels. As most of such cases strongly suggest superior visual quality through the higher amount of pixels on the screen, user expectations evidently rise. Expectations can not only influence the overall experience, but they can also affect the actual perception of visual quality [1, 2].

Another motivation is that studies found in the literature are not fully conclusive on whether UHD content can, in general, provide a statistically significant perceived quality difference compared to HD. At the very least, a high level of content dependency was found in many independent tests. Indeed, while electronics shops typically show so-called

“eye candy” video contents on their displays—which are meant to push the limits of the displays’ capabilities to show the potential buyers what visuals such displays can achieve—the average user does not spend the majority of his or her time watching short demo videos.

The inflated expectations combined with a potential lack of major visual differences can lead to persistent forms of cognitive bias, resulting in serious distortions of QoE. Such distortions are present in everyday life, via given cognitive processes. Therefore, understanding these effects is just as important as the efforts to avoid or eliminate them from subjective studies.

Research questions

The paper primarily investigates research questions on the quality of HD and UHD video in the presence (or absence) of the labeling effect, considering different subjective rating scales. We address the phenomenon of the labeling effect by describing a series of subjective tests that we conducted—a total of four studies, each with a different set of participants. In each of these studies, participants had to compare the visual quality of HD and UHD videos on a UHD display and choose the relative quality difference on a rating scale.

In one set of studies consisting of two tests, subjects were made aware of the content resolution: before each video sequence, a label was shown (“HD” or “UHD”). However, some of these labels were intentionally presented in a misleading fashion, providing false information on video resolution, that is, certain paired comparisons involved two identical video sequences, but the labels suggested that they differed. The test paradigm was then repeated with another set of studies again, two tests without labels, in order to obtain quality ratings unbiased by labels, and to check how big the impact of labels would be.

While the two previously described sets of tests differed in label usage, the tests within a set differed in the rating scale. One test used a more coarse 3-point comparison scale (“Worse”, “Same”, “Better”), the other used one with 7 options. This was done for two reasons: first, the overall perceived differences between two sequences—also considering the labeling effect—may not be great enough to be registered on a 3-point scale, but they may be present on a 7-point scale. Second, we hypothesized that a subject may be more willing to claim that two sequences are not the same if they had an option to choose “slight” preference.

With the whole set of four tests (labels/no labels, 3/7-point rating scale) we can generally answer the following questions, in which question 1 and 2 are dependent on one another: (1) Is there a statistically significantly observable difference between uncompressed HD and UHD video in a typical quality rating test? (2) When rating this difference, are users more impacted by the labels than the actual visual

quality of the content? (3) Are users impacted by the granularity of the rating scale when giving their ratings, particularly in the presence of misleading labels, and can the results from those scales be compared?

Paper structure

The remainder of the paper is structured as follows: "[Related research](#)" introduces the related recommendations and QoE studies regarding UHD visualization. "[The labeling effect](#)" details the psychological background of the labeling effect, and shows examples of its presence in QoE-related research. "[Experimental setup](#)" provides an extensive review of the experimental setup used in our subjective study. "[Results](#)" presents the obtained results and their analysis, followed by a discussion in "[Discussion](#)". The paper is concluded in "[Conclusions](#)".

Related research

Standards and recommendations

As with most technologies finding widespread use among the consumers, there are standards that govern how a technology is to be developed, evaluated, and integrated with other technologies. Standards or international recommendations are provided for that interoperability. Among the most relevant international standards on the topic of UHD are documents from the International Telecommunication Union's telecommunication and broadcasting sectors (ITU-T and ITU-R).

While ITU-R Rec. BT.709 addresses HDTV (i.e., TV up to 1080p resolution), Rec. BT.2020 [3] covers UHD and the corresponding specifications of dynamic range, color gamut and primaries, bit depths, frame rates, and pixel resolutions. Additional recommendations include Rec. BT.1769 which specifies parameter values for large screen digital imagery and how to design a system that gives viewers visual experiences of a high-sensation of reality—which UHD was also developed for.

Subjective quality assessment tests are typically carried out in a rigorous fashion: users are placed in a dedicated testing room with specific lighting conditions and a certain viewing distance to the screen. Guidelines in ITU-T Recommendations P.910, P.911, and P.913 as well as ITU-R Rec. BT.500-13 may be applied in those tests. When subjectively evaluating the quality of UHD systems, the aspect of viewing distance plays a crucial role. Typically, for HDTV applications, human testers are seated at a distance of about 3 H to the TV, where H is the height of the display under study. This is specified in ITU-R Rec. BT.500-13 and ITU-T P.910, where the latter says that "the viewing distance should be

defined taking into account not only the screen size, but also the type of screen, the type of application and the goal of the experiment." This preferred viewing distance therefore mainly depends on user preferences and may be determined empirically, but 3 H has emerged as the standard for HD testing. For UHD screens, however, another recommendation, Rec. BT.2022 [4], was developed, which furthermore distinguishes between the preferred viewing distance and the *design* viewing distance. The latter is the most optimal distance at which "two adjacent pixels subtend an angle of 1 arc-min at the viewer's eye". This distance is 1.6 H for UHD-1 resolution. This distance is also employed by the method specified in ITU-R Rec. BT.2095, which is a test protocol for expert viewing.

If we consider a content with HD resolution visualized on an HD display and a screen height of 100 cm, then the aforementioned 3 H viewing distance for a test participant in a subjective study is 300 cm. Concerning the standardized 1.6 H distance, for a 65-inch screen, this distance is around 1.3 m, which may be much too close for many environments—in fact, in a traditional living room scenario, viewers would rarely sit that close to the screen, particularly when watching in a group of people. It should therefore be virtually impossible for users to distinguish HD from UHD at distances any further, which is very likely to happen in real life.

Appearance of UHD in QoE studies

We find the first major wave of literature covering UHD with regard to QoE in the years around 2013–2015, as tools and technical equipment to display UHD content become more readily available. Before that, research on this topic is scarce. Surveying the literature, particularly with a focus on subjectively comparing HD against UHD, it can be observed that this topic has not yet been studied in a conclusive manner.

Bae et al. [5] conducted a subjective study in 2013, in which HEVC-encoded UHD video sequences were presented at different target bitrates, color formats (YUV 4:4:4 and YUV 4:2:0), and viewing distances (0.75 H and 1.5 H). The authors used two source sequences from the year 2010 for their test, so it can be considered one of the first to investigate QoE for UHD video. Choosing a double-stimulus method (DSIS, see ITU-T Rec. P.910), the original source videos were compared against the encoded ones. The main results of the experiment were subjective ratings for the clips, however the authors did not specifically compare HD and UHD.

In 2014, Tanaka et al. [6] investigated the use of a double-stimulus (DSCQS, see ITU-R Rec. BT.500-13) method for subjectively evaluating 4K video quality. They found it to be viable, but did not compare HD and UHD either.

Li et al. [7] compared different upscaling algorithms for use in content preparation for UHD transmissions via a subjective test. The authors took source material at 2160p, 1080p, and 720p resolutions and upscaled the latter two to UHD using several different algorithms. Their choice of viewing distance was motivated as a compromise between 0.75 H suggested in the literature, and 1.6 H based on the preferred viewing distance (as mentioned in the previous section)—they chose 1 H. In a paired-comparison test, which they deemed to be the most reliable for such tasks, they asked subjects to pick the preferred sequence. Notably, users saw HD and UHD video at the same time, rendered in vertical stripes on the display. As expected, original UHD sequences achieved higher preference values, with Lanczos upscaling performing better than other methods. The authors also noted that “visual acuity on high motion content on 4K screen is significantly lower than in the normal HD condition”, and that there is a center bias when viewing content at such low distances. However, the test methodology itself may be questioned: Users may have been focusing more on detecting the stripes than evaluating the overall quality of the stimuli. It may also be hypothesized that by simultaneously showing the sequences, users could more easily discriminate HD and UHD by making quick eye movements between the UHD and HD stripes. This would not be possible when showing sequences one after the other, thereby leading to results more critical than one would achieve in a more realistic evaluation scenario.

In the work of Weerakkody et al. [8], the authors conducted subjective verification tests for the HEVC standard, in which the potential for bitrate savings against H.264 was the main research target. Using five source sequences and H.264 and HEVC encoding, sequences were presented to subjects at two viewing distances (1.5 H and 2 H). An analysis concluded that for subjects sitting at a greater distance, MOS values were higher, but only for HEVC content. This suggests that impairments of the codec could only be visible at closer viewing range.

In 2015, Berger et al. [9] conducted a subjective study in which 15 contents and four sets of bitrates were chosen per content to investigate the impact of lossy video encoding with the HEVC codec. Subjects were asked to rate the visual quality of the processed stimuli. The authors specifically investigated the practical case of transmission chains, considering that in real-life, bottlenecks in networks may require bitrate and/or resolution reduction for video in order to still be viewable. The authors found that downsampling and later upsampling video (i.e., reducing the resolution during transmission) did not yield perceivable visual quality degradations. Quite the contrary, for some contents, perceived quality *improved* due to the contained camera noise or fast motion in the original source videos. They also found that

a UHD transmission chain required only a slightly higher bitrate for the same visual quality when compared to HD.

In a 2016 study, Xie et al. [10] determined the required HEVC encoding bitrates for UHD transmissions through subjective tests. They found that naïve viewers could not distinguish quality above 5.6 MBit/s. They also found a strong content dependency in their results. Further, they mentioned that in a scenario where viewers have no access to the original source material, the quality differences at high bitrates may generally not be perceptible.

Van Wallendael et al. [11] conducted a subjective study in which UHD was compared against HD, choosing a “striped” test method similar to Li (see above). Based on a set of 31 source sequences, they found that in general, UHD content was determined to be sharper than HD, and that the likelihood of properly detecting real UHD sequences ranged from about 40% to 80%, depending on the content. They also note that there are learning effects and biases inherent in the method that may lead to distorted preference ratings, and that the test itself was judged to be difficult.

A dataset for 4K/UHD video was presented in the work of Zhu et al. [12], using twelve different source sequences and encoding conditions with HEVC-compressed video at different target bitrates.

Sotelo et al. [13] gave an overview of different subjective studies related to video compression and UHD video and also conducted their own test. They noted the limited availability of high-quality UHD video content for test purposes. Ten source sequences were encoded with HEVC at different bitrates. Viewers were seated at two distinct viewing distances (1.5 H and 2 H). Their results are inconclusive and reveal a large number of outliers. Furthermore, a comparison with another test conducted in a different country [12], but using the same source material, revealed large differences in MOS.

In a 2018 paper, Mackin et al. [14] presented a video database containing source material from the authors and third-party content. Using different downsampling algorithms, the authors created test sequences that were subjectively evaluated in a single-stimulus method (i.e., no direct comparisons were made). The authors found that, generally, users preferred UHD-1 over all other (lower) resolutions. However, depending on the resampling algorithm used, no significant differences between UHD and HD quality could be detected. Hence, the authors recommend that for transmission chains with limited bandwidth, the use of HD video instead of UHD may be viable.

Finally, the aspect of using different rating scales has only been addressed in few studies so far, of which [15] is the most thorough. The authors could show that there was no statistically significant difference in typical 5-, 9-, and 11-point discrete and continuous rating scales for typical single-stimulus video quality tests. However, these results

cannot be applied to paired-comparison tests and simpler scales like the 3-point scale employed in our research.

To summarize, it has been shown that UHD video can significantly improve the visual quality compared to HD video. However, these differences strongly depend on viewing distance, the possibility of direct comparison against an original sequence, and—most importantly—the chosen content. In many cases, even within the very strict context of a subjective lab experiment, the differences between HD and UHD are not significant.

QoE models for UHD video

The research shown in the previous section and our studies presented in this paper may be used as data to develop quality models that are suitable for UHD video. In order to correctly develop such models, the validity of the underlying tests has to be ensured as well. Of the many algorithms that exist to predict video quality from input signals, however, only few have been specifically designed for UHD video or verified to be suitable for the tested resolution. There are multiple reasons for this, which relate to *how* video resolution is used in those algorithms, if at all.

One type of metric does not use resolution information for predicting quality. A few of the commonly used image quality estimation metrics such as PSNR or SSIM [16], which are also used for video quality estimation, are of this kind. They can technically be used for images of any resolution, as they do not use resolution as an input factor. In principle, the use of such metrics is valid for a video of any resolution as long as no inference on perceived quality is made without proper empirical data to support this conversion. Hence, for example, translating a PSNR dB score to a Mean Opinion Score (MOS) cannot be done without conducting a subjective study in a pre-determined application context.

There are also metrics that have been trained with subjective rating data on videos of fixed resolutions (or resolution sets), all shown at fixed viewing distances. If these resolutions were smaller than UHD, extending the metrics to support UHD—or any resolution other than the ones they were designed for—would require gathering new subjective data or empirically supported inference about how a higher resolution might change quality ratings. This usually requires a re-training of the entire metric to achieve good prediction accuracy. For example, a video quality metric developed by Netflix, VMAF [17], was recently extended to predict quality for 4K resolution videos using this approach, with the support of new subjective test data.

ITU-T Study Group 12 is currently conducting a follow-up work on the ITU-T Rec. P.1203 family of standards [18] in a joint initiative with the Video Quality Experts Group (VQEG). The standards define an audiovisual quality model for the prediction of HTTP Adaptive Streaming quality. Its

video component has been developed for HD resolution only. The work item is expected to be finished in late 2019, yielding new video quality models for up to 4K resolution that will be internationally standardized. Subjective tests conducted within the scope of this work have made use of an extensive library of pristine UHD content from different sources.

In a 2013 paper, Hanhart et al. [19] evaluated the performance of several common image quality metrics (PSNR, VSNR, SSIM, MS-SSIM, VIF) as well as VQM (a video quality metric) against MOS values obtained from a subjective study, using UHD-only content. The authors concluded that the accuracy and prediction performance of the metrics was mostly content-dependent, and that all metrics performed equally well once fitted to the per-content subjective results, except for VIF, which performed well for all contents. We are not aware of similar, more recent studies on the subject. Also, these studies were only performed on compressed content, thereby leaving open the question how well these metrics perform for uncompressed UHD–HD comparisons.

In general, it has to be stated that the usefulness of any video quality prediction model—particularly when it comes to UHD—strongly depends on the assumed video consumption scenario. As the viewing distance may play a critical role in whether a human can distinguish quality differences between two clips, one has to know for which assumed viewing distance a given model was developed, and whether it takes viewing distance into account at all. To summarize—including the results that we will present in this paper—there are still many open questions when it comes to testing and evaluating UHD quality instrumentally, that is, through models.

The labeling effect

Introduction of the phenomenon

The labeling effect is the result of a process during which the information (a label or multiple labels) regarding an entity alters the way the entity is perceived or experienced. The generalization regarding the senses is important, as it may not only affect the handling of visual data, but it can affect e.g., hearing, smelling, tasting, and the overall sensation of experience as well. Also, it does not need to happen real-time, as the modification of the memory of an experience can have an equivalent gravity. The labeling *effect* is not to be confused with the labeling *theory* of sociology [20, 21], which focuses on the alteration of the individual's self-perception.

A classic example of the labeling effect is when a man walks into a classroom, gives a brief guest lecture, and

leaves. Then the class is asked about the height of the man. If the man at the beginning was introduced as an internationally recognized expert or as a professor of the field, he could be perceived taller, than if he was introduced as a mere assistant or a fellow student [22].

Another example can be a bottle of wine, with a label and an overall presentation suggesting excellence. It may indeed enhance the experience; however, the opposite can occur as well, in case the margin between the expectations and the actual taste is too large, leading to disappointment [23]. In the special case of alcoholic drinks, the alcoholic content itself (given in percentage) can also influence the experience. In the work of Masson et al. [24], the same wine was provided to test participants in different bottles, and the sole variable was the alcoholic content presented on the label. The study found that the ones suggesting higher alcoholic content were more favored, even though the wine itself was the same. In case of such prominent products, the region of origin (even within a country) can play a particularly significant role, especially in case of well-known vineyards [25]. Going beyond written information, the experiment of Lick et al. [26] addressed the connection between the color of the label and the assumed taste. The authors conclude that certain wine tastes are generally associated to very specific colors, e.g., in the study, from the perspective of the test participants, orange labels suggested sweet and fruity flavors.

Pricing is a crucial form of the labeling effect. Price tags fundamentally affect the way the quality of an item is perceived and influence monetary decisions, such as buying the item. The general concept is that the more expensive something is, the better it must be, as there must be a reason why a given item costs more than a different one. This is necessarily present in situations of financial investments and purchases of any level, as the phenomenon builds on our trust in the commercial world. However, a label in such scenarios is not limited to the price tag, but the brand alone can be sufficient to affect perception and experience. Also, the post-purchase experience highly depends on the cost, and the experience itself can become a tool of post-purchase justification. A demonstrative example of consumer price consciousness and the association between price and quality is the work of Sinha et al. [27], particularly dealing with private label brands.

The research of Johansson et al. [28] focused on the “Made in” labels in the aspect of consumer information processing. Such experimental aim can be particularly relevant, as consumers tend to associate different levels of product quality with given countries. The work of Hamzaoui et al. [29] separated the country of origin into country of design and country of manufacture, and thus the authors investigated consumer behavior and perceived quality towards bi-national products. The obtained results indicate that the perceived quality of durable goods is influenced

more by the country of manufacture, and less by the country of design. The country of design can be influential when the overall value of the product is highly dependent on design (e.g., cars), but even in such cases, the country of manufacture still remains dominant. The findings of Ahmed et al. [30] pointed out the correlation between the country of design and product complexity, and Batra et al. [31] highlighted the social signaling value of such products.

In many cases, the labeling effect can surface as a misinformation effect [32], especially in the presence of post-event information. New information and new memories can easily lead to retroactive memory interference, changing the way the perception or experience is remembered. The labeling effect is also connected to the framing effect [33], as the presentation of the information can significantly influence its processing and thus the corresponding decisions. For example, in the work of Gachter et al. [34], the framing of the registration fee of a scientific conference was investigated. In the experiment, the early-registration fee was presented to half of the PhD students as a discount, while to the other half a late-registration penalty was communicated. The results show that in case of a discount frame, only 67% registered before the decisive deadline, but 93% registered when a penalty frame was empathized.

Generally speaking, the labeling effect itself is enabled by cognitive dissonance reduction between conflicting cognitions, the theory of which was introduced by Festinger in 1962 [35]. In the case of this phenomenon, one of the cognitions is perception (genuine experience) or the memory of an experience, and the other one is the collection of thoughts, feelings and memories that can be associated with the label(s). When these cognitions contradict each other to an extent, their dissonant state is reduced or eliminated by changing one of these cognitions, which in case of the labeling effect is the genuine experience or its corresponding memory. It needs to be added that positive reinforcement is possible as well, when cognitions share the same direction (e.g., all associated cognitions agree that a given item has a good quality), but they differ in extent.

Labels in the context of items and services should not be looked as something inherently bad. Although they do affect our perception and experience, but they also help us navigate in the sea of information. A very common type of label is the list of capability parameters, which describes the most important factors of modern electronic devices, such as household utilities or items of entertainment. With such labels, we can directly compare the capabilities of devices, before making a financial decision.

Let us take for instance televisions at a shop. They are usually turned on to show some looping demo content. In case of a UHD/4K-capable TV, high-quality contents are shown in the appropriate resolution, in order to “show off” what the display can achieve. In a way this is actually a

tricky subject, as in a regular use case scenario, the user will not use the display to play such demo content all the time. First of all, many contents may have lower spatial resolutions, and even if a video was shot in UHD/4K, a slight noise or defocus may hinder the potential utilization of this capability.

In a typical scenario, a person wishes to purchase precisely one of the given televisions at the shop. There is a specific time constraint for the decision, as time is not unlimited, especially if having an inconclusive visit to the shop is not an option. The final monetary decision is influenced by the perception of quality and by the available information, in forms of labels (capabilities, brands, prices, etc.). Having a large variety of available televisions naturally creates a greater dependency on the labels, beyond the initial filtering. However, there is a two-way cross-influence between the factors of perception and labels: as discussed earlier, perception is affected by labels, but also, even though labels are not practically changed in any way, the processing of labels can be affected by the visual experience (e.g., “For such a nice picture, this price is not that high after all.” or “I guess I should not be looking at that parameter in the list as it does not make any difference.”). It is important to add that labels can in fact affect each other as well (e.g., “This is not so expensive for this specific brand.”).

In this paper, we address the influence of labels over the perceived quality. More precisely, in our research, we investigated how the perception of HD and UHD contents shown on a UHD/4K-capable TV was affected by labels indicating their spatial resolutions. The concept of dealing with the labeling effect in QoE-related researches is not novel, as there is already a vast scientific literature on the topic.

QoE and other studies

The *E* in QoE stands for “experience”. The definition of QoE narrows down the concept of experience for a given “application or service” [36], yet through the generic nature of this word, the labeling effect can be considered for a broader sense of experience. For example, drinking a glass of cold beer may be an experience, and its properties that are not directly linked to the actual taste may affect the drinker’s satisfaction as well. The work of Jacoby et al. [37] involved price, brand name and the composition of beer as labels. Verbeke et al. [38] investigated the labeling of beef, and Burton et al. [39] focused on nutrition reference information in the context of product evaluation. Generally, the intrinsic and extrinsic properties of different purchasable goods as labels were studied by Szybillo et al. [40] and Richardson et al. [41]. Moreover, brands were also addressed by the work of Delvecchio et al. [42], and Heisey [43] investigated the role of a minimum information environment (involving manipulated information) with regards to the perceived

quality of identical clothings; the same sweater was presented with different information cues, resulting in altered perceptions of quality.

As for the perception of information and its quality (i.e., credibility), the work of Rieh et al. [44] addressed the domain suffix (.org, .gov, .com and .edu) as an influential factor among scholars. The findings indicate that in many cases, credibility was clearly attributed to the given suffix; for example, when a test participant was asked about the credibility of the information presented on a website, the response was “Absolutely [I trust it] because it’s an .org” [45].

In the more conventional sense of QoE, the dissertation of Schöffler [46] investigated listening experience. In the experiments, test participants were asked to take “everything” into consideration when assessing the overall audio quality, and they were explicitly told that the stimuli differed in quality. Emphasizing the difference between the quality levels of the stimuli can lead to a preconception stating that “there should be a difference”, inducing variations in the listening experience and thus in subjective ratings of stimuli which would have none otherwise. Music excerpts of various genres were used as audio stimuli, and test participants also had to self-assess the impact of the songs (and their performers) on their own quality ratings. The song-related information served as a label, especially since test participants were specifically given the task to consider it for the overall subjective assessment. Beyond the presence of the labeling effect, it is noteworthy that certain participants rated lower-quality stimuli higher than undistorted, high-quality stimuli, due to their prior experiences (e.g., the low-quality music excerpts reminded a participant of the pleasant memories of concerts and festivals in the past).

As shown by the previous example, in subjective tests of multimedia quality, the labeling effect may take a foothold, as almost any information can influence quality ratings provided by the test participants. However, certain studies particularly aim at this phenomenon, in order to discover the magnitudes of achievable distortions; how much the labeling effect can distort subjective test results. Many researches with such goal involve mock-up scenarios and stimuli, in which the exact same multimedia quality is provided through a given content, but the associated labels differ.

The experiment of Lamm et al. [47] evaluated simulated search engines biased with—as the authors phrase—“manipulated” user expectations. One group of the test participants were informed prior to quality assessment that the search engine was actually an expensive professional search system (high expectations), while a different group was told that it was only a mere student project (low expectations). This separation was repeated for two distinct levels of objective system performance (low and high), therefore the participants were clustered into four groups; every individual was

only exposed to a single test in the two-factor study design. The two performance levels were simulated via artificially constructed search result lists for the different topics, i.e., the one with the low system performance provided a higher number of irrelevant results. The subjective data collected from the 89 test participants (all female, in order to avoid the gender effect) reported no significant difference based on user expectations, but note that a given test participant was only provided one specific label, and had nothing to compare to. In a follow-up work [48], the authors extended their methodologies with direct service comparison, resulting in eight groups, as all combinations were investigated: a group was first provided a search engine with either an objectively good or bad quality, labeled with one of the previous descriptions, and then another search engine was provided with either good or bad quality—thus half of the test conditions included identical stimuli—but it was given the other label. Again, the labels always differed, so one was labeled expensive and professional, while the other one was the work of a student with unknown quality. Although the results do indicate the significant influence of the labeling effect, it is also shown that expectations maybe be overwritten by performance experience over time. These findings correlate with the conclusions of Szanja et al. [49], stating that certain expectations may fade as time progresses.

Bouchard et al. [50] investigated the sense of presence for virtual reality. Although test participants were immersed in a synthetic environment, they were informed before the test that they would be immersed in a real-time replica of an actual room, containing a real mouse in a cage. In the context of this research, the term “real-time replica” means that test participants were told that what they saw through their viewing equipment was actually happening in the real room at that very moment. This was emphasized via the mouse, as it was awake and therefore was performing certain actions. A different group of the test participants was told the same thing, but without the real-time component of immersion. In reality, every test participant was immersed in the same synthetic environment. The primary goal of the study was to assess whether this aforementioned “real-time” notion contributes to the sense of presence or not; would test participants feel more immersed if they knew that the virtual world around them was not only a copy of something real, but they saw what a person would see while actually standing in the place at the same time? The subjective results indicate a significant difference in the sense of presence of the two groups, favoring the scenario with the real-time component. Furthermore, the study was repeated with the use of simultaneous functional magnetic resonance imaging (fMRI), indicating significant differences in brain areas that are related to immersion and presence, and thus concluding that the misleading information resulted in a genuinely higher sense of presence.

In the contributions of Sackl et al. [51, 52] and Kara et al. [53, 54], the label was the type of connection. In both researches, perceived quality was measured in a mock-up scenario, where the performance of wireless and wireline connection did not differ at all. In fact, in some of these works, there was not even any multimedia transmission, as the stimuli were played from the local storage of the device. By doing so, identical quality was ensured, yet the subjective scores significantly differed. It needs to be noted that the direction of such distortion (whether it enhances or degrades user experience) is not evident; it depends on the test participant. While many test participants had notably lower degrees of QoE in the wireless test cases, others actually perceived the wireless to be much better. Sackl et al. also addressed the Willingness to Pay (WTP) [55, 56], as labels indeed affect the customers’ monetary decisions. User expectations were also in direct focus of a subsequent work [1], due to the socio-psychological reasons mentioned earlier. The authors propose the systematical integration of expectations into QoE-related research. The work addresses two specific expectation types: desired and adequate [57, 58]. While the first one is typically invariant (basic needs), the latter varies more over time and it highly depends on the context (minimum baseline of tolerance). The direct triggers of expectations in the various studies introduced in the publication were different labels, such as technical properties, price and service title (e.g., gold, silver and bronze).

In a joint work of the authors [59], the label was the brand of the mobile end-user device. Although each device played the same stimuli locally in an unimpaired quality, most of the test participants perceived visual degradations (i.e., playback jitter, tearing, blurred pixel zones, black/missing pixels, etc.) on the unfavored devices. The effect of smartphone brands on user experience was also investigated by the thesis of Ebbing [60].

In a recent work considering the labeling effect, Kara et al. [2] addressed the perception of the different aspects of High Dynamic Range (HDR) visualization. In a paired comparison, video stimuli were compared regarding luminance, frame rate, color and image quality. As this research, similarly to previously detailed works, also used mock-up methodologies, the video sequences for the given sources were identical. The stimuli were differentiated by the labels “HDR” and “Premium HDR”, where the latter suggested superior visual quality compared to the other. The obtained results show that more than 75% of the ratings reported either positive or negative changes in the perceived quality aspects. The majority of these favored the “Premium HDR” videos, yet it needs to be noted that in a considerably high number of cases, test participants experienced degradations in frame rate for the stimuli. In a post-experiment questionnaire, several test participants explained this as a trade-off between frame rate and the other aspects; it was a common

belief among them that the observable improvements in the appearance of the videos came at the price of frame rate.

In the work presented in this paper, we carried out subjective tests with and without labels. When labels were used, certain test conditions were akin to the mock-up methodologies of the prior works, as identical video stimuli were compared with different labels. For example, both stimuli were either identical HD or UHD videos, but the labels stated that one of them was HD while the other one was UHD. These were compared with test conditions where the stimuli genuinely differed, and where the stimuli were identical but the labels reflected this fact rightfully. Also, each and every test condition was subjectively assessed without labels as well. The test conditions and all the important parameters of the experimental setup are detailed in the following section.

Experimental setup

Our main aim was to investigate the impact of the labeling effect and the rating scale on the perceived quality of UHD services. In this section, we present the commonalities between the four studies detailed in the paper, and highlight their differences in the utilized test protocols and in the questionnaires. The four studies are:

- Study 1: Test with labels, 3-point rating scale
- Study 2: Test with labels, 7-point rating scale
- Study 3: Test without labels, 3-point rating scale
- Study 4: Test without labels, 7-point rating scale

All studies used exactly the same video sequences, but were run with a different set of subjects. This way, we could ensure that subjects would not learn about how the test paradigms differed, which could have introduced an avoidable bias. The remainder of this section will detail the experimental setup from a technical and experimental design perspective.

Research environment and the UHD display

All the subjective tests were carried out on a Samsung 55-inch JU6400 6 Series Flat UHD/4K Smart LED TV.³ The display and thus the tests themselves were located in an isolated laboratory environment, in which test participants suffered no audiovisual distractions. Based on the guidelines of the ITU-R Rec. BT.2022, the test participants were seated at a distance of 1.6 H from the display, which in case of the aforementioned 55-inch TV, corresponded to 110 cm. The angle of vision was zero; test participants viewed the display

precisely from the middle. Behind the display was a plain D₆₅ background, as specified by ITU-T Recommendation P.910 [61]. The background fully covered the field of view of the test participants, omitting any visual distraction. The room had lighting conditions set to approximately 25 lux, which did not vary during the subjective tests. This was also ensured by the design of the laboratory, as no external light could enter the room during the experiment.

Test subjects participated individually, separately, as a single position of observation was defined. This is also in alignment with our approach to provide a unique, randomized stimulus order for each and every test participant, making the scenario of multiple observers unavailable.

The series of subjective tests at hand solely focused on visual quality, hence we excluded audio from the research. This means that no stimulus contained audio data, no sound was generated by the speakers of the television during the tests, and no external audio gears were worn by the test participants.

Test conditions

Each condition in our test was a comparison between two video sequences. The comparisons between these sequences were either a) transitions or b) self-comparisons. During transitions, the first video was HD and the second one was UHD, or vice-versa. Self-comparison means that both stimuli in the pair had the same resolution, i.e., the subject saw HD and HD, or UHD and UHD. This means that the videos in self-comparisons not only had the same resolution, but they were in fact exactly the same sequence. Since there were two resolutions, this means that there were a total of four possible comparisons a subject could rate.

Note that in every comparison, test participants were shown sequences from the same original source, in order to make a direct comparison possible. Each test condition was then applied to multiple sources.

For the tests with labels, these conditions were paired with the four possible combinations of labels attached to the sequences. This means that the label could either indicate the correct resolution of the clip to follow, or purposely deceive the user into thinking that another resolution than the one actually shown would be presented. In the latter test conditions, the displayed stimuli in the pair had the same resolution, however, the labels suggested transitions.

Intentional mislabeling is a common practice among such studies, and in a way it is a necessary component of the chosen methodology to directly measure the bias induced by the label. Particularly in the case of mock-up experiments, the sole variable in certain investigated test conditions is the label itself, while the target of assessment does not change (i.e., in our case, video resolution stays the same). The previous section introduced numerous subjective tests which

³ <http://www.samsung.com/uk/tvs/uhd-ju6400/UE55JU6400KXXU/>.

Table 1 Investigated test conditions

ID	Video 1	Video 2	Video 1 label	Video 2 label
1	HD	UHD	HD	UHD
2	UHD	HD	UHD	HD
3	UHD	UHD	HD	UHD
4	HD	HD	HD	UHD
5	UHD	UHD	UHD	HD
6	HD	HD	UHD	HD
7	UHD	UHD	UHD	UHD
8	HD	HD	HD	HD

fundamentally relied on intentional mislabeling [1, 2, 47, 50, 54].

Table 1 introduces the 8 test conditions we investigated. Conditions 1 and 2 were the transitions with correct labels, condition 3, 4, 5 and 6 were the possible combinations of self-comparisons with misleading labels, and 7 and 8 were self-comparisons with correct labels.

In case of the experiment where no labels were present, we simply left out the labels but otherwise used the same test conditions. This means that condition 3, 5 and 7 were practically identical, and 4, 6 and 8 did not differ in any way either.

Source sequences and test stimuli

In the experiments, we used eight different source videos (SRCs). As original content, four UHD/4K movies were chosen: “Big Buck Bunny”, “Sintel” and “Tears of Steel” from Blender, and “El Fuente” by Netflix. From each of those movies, we cut two 10-s sequences to be used as sources. In Fig. 1, one row refers to one content, and shows one representative frame from each of the selected 10-s parts of the videos.

The actual HD and UHD test stimuli were created by using the available uncompressed frame sequences of the original movies, which were merged into video files for playback. We used a YUV 4:2:0 color space, and a frame rate of 24 in every video, created by chroma subsampling and frame rate downsampling using the `ffmpeg pix_fmt` option and `fps` filters, respectively. For two movies (Sintel and Tears of Steel), it was necessary to add letterboxing via black bars to bring them to UHD-1 resolution, since the originals were not available at full UHD-1 height and used a non-16:9 aspect ratio. For the other two (Big Buck Bunny and El Fuente), it was required to drop frames from the original 60 fps and 59.94 fps, respectively, to achieve the target of 24 fps. In a small-scale experiment performed by the authors, this frame rate reduction was deemed noticeable when comparing the original with the processed movie. However, when viewed on their own (i.e., without high-fps reference), the processed movies at 24 fps were not perceived to be “jerky” at all; they

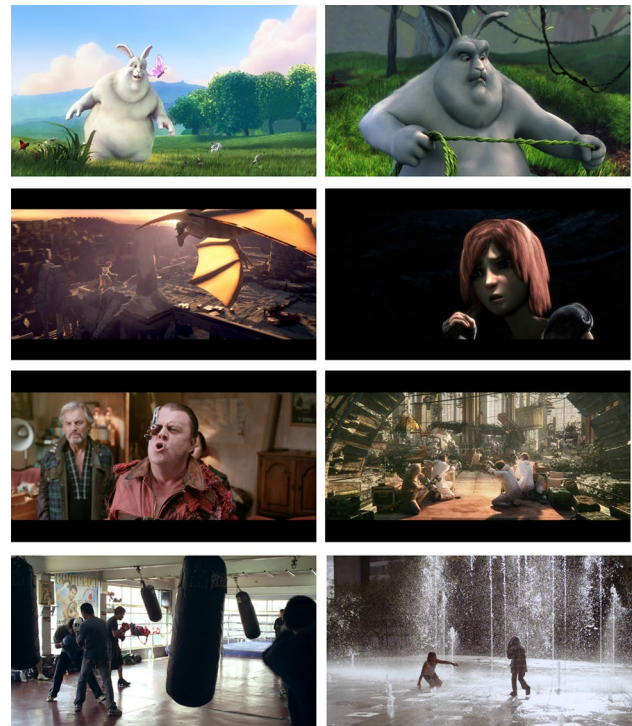


Fig. 1 Source videos used in the experiments. Movies used per row: (1) Big Buck Bunny, (2) Sintel, (3) Tears of Steel, (4) El Fuente

did not stutter in their appearance and therefore they did not noticeably differ from the other two in this aspect.

To generate the test stimuli that showed HD content, we first downsampled the UHD SRCs to HD (1920×1080), and then upsampled them back to UHD. These two sampling procedures were made using the FFmpeg implementation of the Lanczos filter ($\alpha = 3$) using the default configuration in FFmpeg 3.0. This upscaling was performed to ensure that no further scaling would be applied by the software player, the graphics card, or the display itself. Furthermore, the final test stimuli were stored in an uncompressed format. It was ensured that the used software player (`ffplay`) and graphics card (NVIDIA) could handle playback without frame drops.

Spatio-temporal analysis

The primary aim in choosing the original contents and cutting the videos to particular scenes was to achieve diversity in content genre, motion descriptors, saturation, brightness and level of image detail, which was one of the most important parameters. Two contents were CGI animation (“Big Buck Bunny” and “Sintel”), which we found crucial to investigate in this study, as computed graphics can enable a high level of visual detail when rendered at the target resolution. As an example, the hair of the character in the zoomed-in shot of SRC04 from “Sintel” was highly detailed. At the

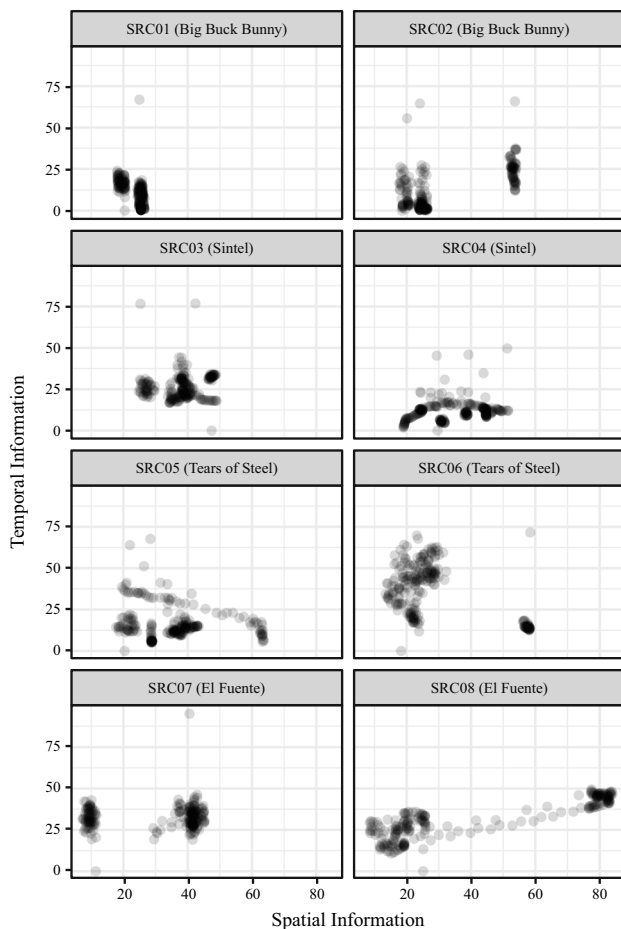


Fig. 2 Spatial and Temporal Information of used SRCs; each point corresponds to one frame. Darker areas indicate higher concentrations

same time, SRC02 from “Big Buck Bunny” mainly had smooth rendered surfaces.

To further investigate the content differences, we conducted an in-depth spatio-temporal analysis. Figure 2 shows the spatial and temporal information (SI/TI) of each frame for each SRC, calculated according to ITU-T Rec. P.910 [61] using a publicly available software implementation.⁴ As each frame corresponds to a single point in the plot, darker patches indicate higher occurrences of SI/TI combinations. It can be seen that SRCs 01 and 02 have relatively little temporal activity with no camera movements; SRC01 is—due to its usage of blurred backgrounds—low in spatial complexity, while SRC02 has areas with higher details. SRCs 03 and 04 show medium-level complexity with camera panning. SRCs 05 and 06 have higher variations in detail; they are higher-paced action movie scenes. Finally, SRCs 07 and 08 are also similar in visual appearance, but particularly SRC08 has the highest spatial complexity of all used SRCs.

⁴ <https://github.com/slhck/siti>.

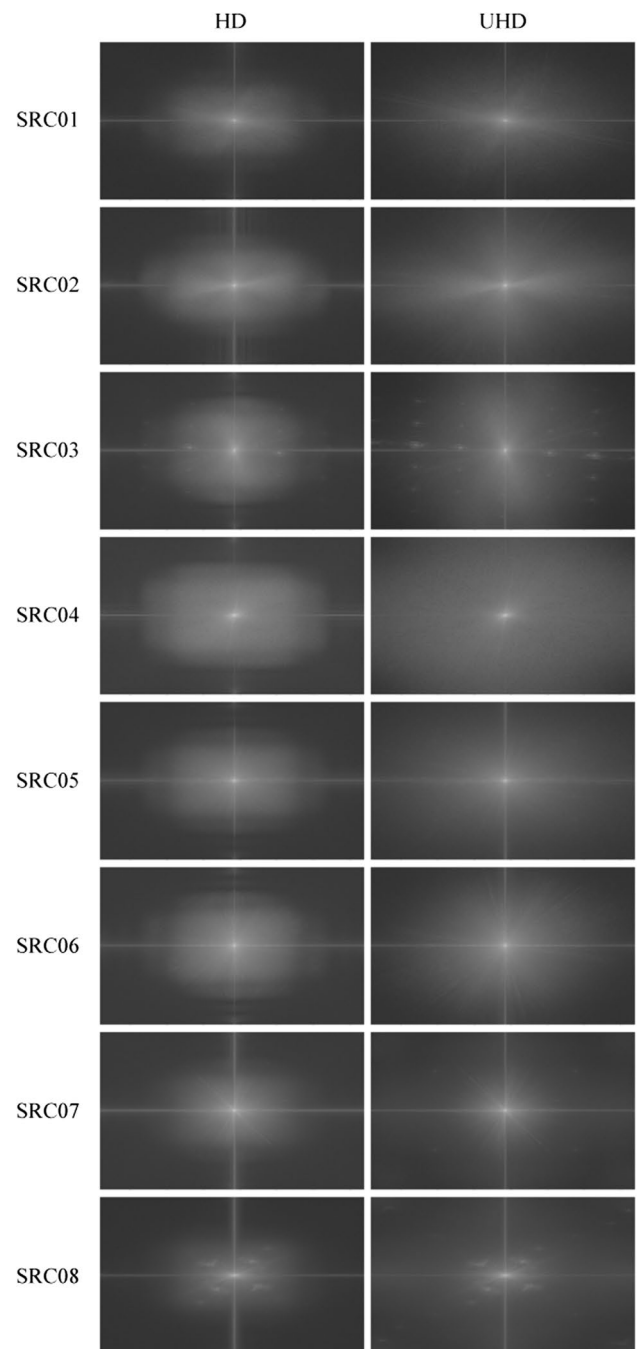


Fig. 3 Average magnitude spectra of 2D-FFT for all sources. From top to bottom: (1) Big Buck Bunny (SRC01, SRC02), (2) Sintel (SRC03, SRC04), (3) Tears of Steel (SRC05, SRC06), (4) El Fuente (SRC07, SRC08). Left column: HD after upscaling to UHD, right column: UHD

To additionally visualize the impact of the scaling operations, Fig. 3 shows the magnitude spectra of the frequencies in each SRC, averaged over all frames. The spectra were calculated using a 2D Fast Fourier Transform (FFT) of the luminance components of the SRCs, additionally shifting the FFT such that the zero frequency

Table 2 Quality metric results for each SRC

Movie	SRC	PSNR	VIF ₀	VMAF
Big Buck Bunny	SRC01	48.70	0.90	97.87
Big Buck Bunny	SRC02	44.79	0.81	96.30
Sintel	SRC03	43.07	0.83	98.42
Sintel	SRC04	40.34	0.71	89.25
Tears of Steel	SRC05	44.75	0.83	98.48
Tears of Steel	SRC06	46.23	0.84	99.56
El Fuente	SRC07	43.35	0.56	99.97
El Fuente	SRC08	42.31	0.67	99.99

component is shown in the center, and logarithmically scaling the resulting values for better visualization [62]. Each original SRC is one row; the upscaled HD variant is shown on the left, while the UHD variant is on the right. It can be clearly seen that the downsampling to HD has reduced the high frequency components present in the UHD original, as identifiable by a smoothed rectangle-like shape in the left column.

Video quality metric results

In addition to conducting a spatio-temporal analysis and visualizing the frequency spectrum, we calculated several well-known video quality metrics available through the `libvmaf` library available in `FFmpeg`. `libvmaf` calculates per-frame estimates of PSNR, SSIM, MS-SSIM, VIF (at four different scales, VIF_{0,1,2,3}), and VMAF itself. As VMAF model, we used the `vmaf_4k_v0.6.1` version that was introduced in June 2018 and is the most suitable for our purposes, as it simulates a viewing distance of 1.5 H and was trained on UHD material. We calculated these metrics by comparing each HD and UHD variant of the same SRC, that is, we wanted to find out if the metrics could discriminate the contents.

In Table 2 we only show the metrics that detected a difference between the HD and UHD SRCs (PSNR, VIF₀, VMAF). Presented are the metric values averaged over all 240 frames. The other metrics (SSIM, MS-SSIM, VIF_{1,2,3}) were all 1.00 on average; they were therefore not able to show differences between the HD and UHD variants. These results are curious insofar as we expected all metrics to indicate *some* change. A detailed analysis of why the metrics perform this way, however, is outside of the scope of this paper and will be part of future work. To summarize, we can see that several published metrics behave unexpectedly on uncompressed UHD content, and that subjective comparison of such sequences is required to quantify quality differences.

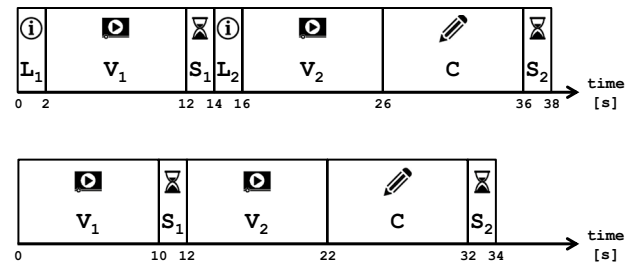


Fig. 4 Visualization order of the experiment with labels (top) and without labels (bottom). A single test included two video stimuli (V), separation screens (S), a subjective comparison period (C), and optionally labels (L)

Rating scales

As explained in the beginning of this section, the subjective tests differed in terms of rating scale used. In order to take the “expressive power” of subjective comparison scales into consideration, we utilized two scales with identical rating concept, but with different level of detail. One was a simple 3-point (“Worse”, “Same”, “Better”) comparison scale, which enabled a basic discrimination of video quality. The other one was the ITU-R Rec. BT-500.13 7-point (“Much worse”, “Worse”, “Slightly worse”, “Same”, “Slightly better”, “Better”, “Much better”) comparison scale, which also let test participants express the magnitude of the experienced difference. It should be noted that a forced-choice rating (i.e., eliminating the “same” option) was not considered for our experiments, since we (1) expected users to not be able to spot a difference in all cases and (2) our test design included a high number of self-comparisons where in fact the stimuli were the same. We therefore wanted to provide a valid option for correctly marking such sequences, thereby keeping cognitive dissonance low during the test.

Test protocols with and without labels

As mentioned before, the four experiments differed in terms of the label: two experiments explicitly identified the spatial resolution of next video stimulus through labels, while the others did not.

Figure 4 shows the chronological structure of the stimuli as they were shown to the subjects; it indicates when and what was shown to every test participant. Every stimulus pair included 2 videos, referred to as Video 1 and Video 2, and the task was to compare Video 2 against Video 1.

In case of labeled videos, first the label of Video 1 (L₁) was displayed for 2 s, followed by the 10-s Video 1 itself (V₁). The same was shown for Video 2 (L₂ and V₂), but before that, a blank separation screen (as recommended



Fig. 5 Labels (top), separation screen (bottom left) and screen during quality assessment (bottom right), as shown on the display

by ITU-T P.913) was on for 2 s (S_1). The stimuli were followed by a 10-s period for the subjective comparison (C), during which a short text was displayed on the screen, asking the test participant to cast the vote on the evaluation sheet (“Please vote!”). The protocol for each comparison ended with another 2-s blank separation screen (S_2), creating a brief pause between the different stimulus pairs.

The aforementioned screens are shown on Fig. 5. In practice, the labels were video segments, showing a single frame for a given duration (2 s), containing either the word “HD” or “UHD”. During the training phase, these screens were used as well, and their meaning was briefly explained to the test participant. Although the labeling was not *verbally* questioned during the experiment, its understanding varied per participant, and so did the compliance with the labels, as it shall be seen later in the paper. Furthermore, it needs to be particularly noted that no information of any sort was provided to the test participants that hinted any possibility of the labels being incorrect.

In the tests where no labels were present, the experimental protocol was the same, but without the 2-s resolution identifiers before the stimuli. This resulted in a minor difference between the total test durations of the experiments. With labels, the test took roughly 40 minutes to complete, while it was approximately 5 minutes less when no labels were present.

In both experiments, the 8 test conditions were applied to all 8 sources. This means that each test participant made 64 comparisons, and thus observed 128 video stimuli.

As it has already been stated earlier, the order of all comparisons was randomized; it was unique for every test participant. Randomization was performed in a way that avoided content repetition, so that adjacent stimuli pairs always had different sources. Since conditions were symmetrical (i.e., both HD → UHD and UHD → HD were tested), order effects could be avoided by the randomization.

Pre- and post-experiment questionnaires

At the beginning of the experiment, before the training phase, the test participants had to fill out a pre-experiment questionnaire, which was the same for both types of experiments (with and without labels). First we gathered basic demographic information, such as age and gender. This was followed by three questions on prior experience and familiarity with UHD/4K, as shown in Table 3. Also, the test participants were subject to screenings based on the Snellen charts and Ishihara plates, in order to ensure the validity of our research.

The post-experiment questionnaire, as shown in Table 4, included five questions that were asked in both experiments. These were to be answered on a quasi-continuous scale ranging from −10 to 10 (without 0). The choice of scale was made based on its usage in the original NASA Task Load

Table 3 The pre-experiment questionnaire

Have you ever heard of “Ultra HD”, “UHD” or “4K”?
– Yes, and I could explain what it means
– Yes, but I could not explain what it means
– No, never
Have you seen a video in UHD / 4K resolution yet?
– Yes
– No
– I do not know
Do you possess a device with UHD / 4K resolution?
– Yes
– No
– I do not know
– I do not wish to answer

Table 4 The post-experiment questionnaire

Common:
– How mentally demanding was the task?
– How physically demanding was the task?
– How hurried or rushed was the pace of the task?
– How successful were you in accomplishing what you were asked to do?
– How insecure, discouraged, irritated, stressed, and annoyed were you?
Tests with labels:
– After having participated in the test, would you say that 4K video is better than HD video?
(Yes. / No. / I don’t know.)
– When comparing HD and 4K, what is the main difference for you?
Tests without labels:
– In case the videos in the pairs differed, what was the main difference for you?

Index (TLX) questionnaire [63], whose questions we partly used. A similar scale is shown in ITU-T Rec. P.910 [61], Figure B3. Positive numbers (right part of the scale) represented high mental and physical demand, rushed test pace, lack of confidence in ratings, irritation, stress and so on and so forth, while negative numbers (left part of the scale) were used to express the opposite. In this context, the opposite of “rushed” is “not rushed”, and not “too slow”.

In all experiments, the test participants were asked whether they considered the UHD stimuli to be generally better than the HD, and more importantly, they were asked about what they thought the source of the difference was. Their answers were collected in written fashion.

Results

In this section, the results from the experiments are presented. In order to enable reproducible research, the obtained experimental results are available in an online repository [64]. The published data includes the raw ratings of each test subject as well as the pre- and post-experiment questionnaire answers. In order to fully protect the anonymity of the participants and to comply with the regulations and policies of the involved institutions, information on age and gender needed to be stripped from the results.

The section begins with the basic demographic information of the test participants and the data collected using the pre-experiment questionnaire. Each study is then statistically analyzed via the comparison of test conditions. This is followed by the investigation of content dependency, in order to observe whether the selected contents influenced the subjective ratings or not. The results achieved for the different scales within a test type are matched to determine the impact of the involved rating scales. The four studies are also analyzed in a per-subject manner; the rating behavior of the individual test participants is addressed. This is then extended by the calculations on rating correctness (i.e., how did the subjective ratings match the actual resolution transitions between test stimuli) and the compliance with labels (i.e., how much did the test participants obey what was suggested by the labels). Finally, the section ends with the detailed analysis of the results of the post-experiment questionnaire.

Panel and pre-experiment questionnaire

Tests with labels

A total of 30 people took part in our experiments with a label shown before each video stimulus. The test participants were from an age range between 18 and 39, and the average age

was 25. The subjective test was completed by 23 males and 7 females.

From the 30 test participants, 8 knew what UHD is, 16 heard about the term and 6 had not heard about UHD prior to the experiment.

The number of participants who had seen UHD videos before the subjective test was 8, while 13 had not, and 9 were unsure about the answer.

At the time of the research, no test participant possessed a UHD-capable device. To be more precise, according to the questionnaire, none of them could state owning such device, as 6 were unsure whether what they had were UHD-capable or not, and 24 were certain that their devices were not UHD-capable.

Tests without labels

Similarly to the tests with labels, a total of 30 people took part in our experiments that did not contain labels regarding the resolution of the video stimuli. The test participants were from an age range between 20 and 40, and the average age was 25. The subjective test was completed by 23 males and 7 females.

From the 30 test participants, 10 knew what UHD is, 16 heard about the term and 4 had not heard about UHD prior to the experiment.

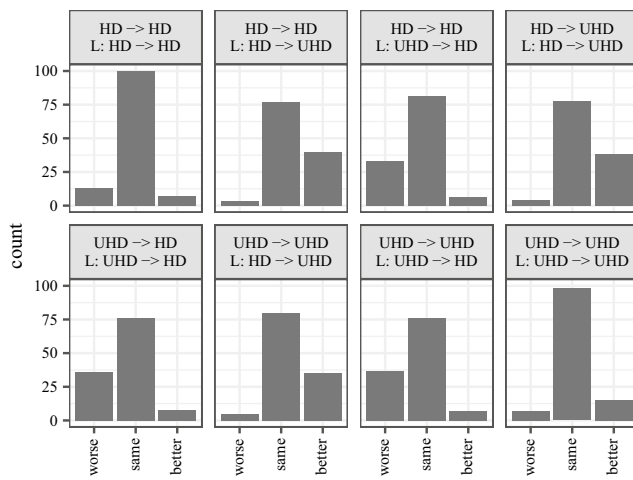
The number of participants who had seen UHD videos before the subjective test was 17, while 10 had not, and 3 were unsure about the answer.

According to the questionnaire, 7 test participants possessed UHD-capable devices, while 20 did not, 2 were unsure whether what they had were UHD-capable or not, and 1 person did not wish to answer the question.

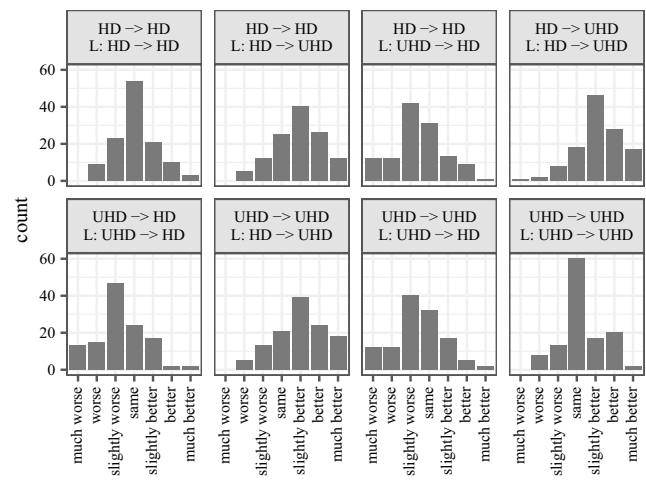
While the experiments with labels were conducted in 2016, the experiments without labels were carried out a year later, in 2017. The second set of studies was performed after publishing the first results at a conference [65] and receiving feedback encouraging us to repeat the experiment without labels. Although there was no notable difference in “UHD awareness” among the test participants of the sets of experiments, there was an apparent raise in prior UHD video experience and in the possession of UHD-capable devices.

Tests with labels

The results of the tests where labels were present during the experiment are shown in Fig. 6a, b, with histograms of the ratings for the 3-point and the 7-point scale, respectively. There are 8 investigated test conditions, as defined in Table 1.



(a) Study 1: Test with labels, 3-point rating scale.



(b) Study 2: Test with labels, 7-point rating scale.

Fig. 6 Histogram of test ratings with labels

Study 1: 3-point scale

When identical video stimuli were shown to the test participants, accompanied by identical labels, in case of both UHD and HD videos (condition 7 and 8, respectively), the provided ratings clearly reflected that users chose to agree with the labels in that there was no difference. However, when misleading labels were introduced for these identical pairs (condition 3, 4, 5 and 6), roughly a third of the given scores indicated that the stimulus with the UHD label was better. On its own, this already implicates a strong presence of the labeling effect in the obtained results. However, when these scoring patterns are compared with the results of *genuinely* different stimuli with correct labels (condition 1 and 2), a peculiar similarity is revealed.

These observations are reinforced by the statistical analysis of the investigated conditions, as shown in Table 5. In order to evaluate whether there were statistically significant differences between the ratings depending on the shown conditions, we first calculated an ANOVA using the condition as independent variable and the ratings of the test participants as dependent variables. The ANOVA ($df = 7, p = 0.00$) indicated a significant impact of the conditions.

To then investigate individual differences between two conditions c_1 and c_2 , the Tukey HSD (T), Bonferroni (B) and Holm (H) multiple comparison tests were conducted. We considered a condition pairing to have a significant influence on the ratings if the Tukey HSD p value was below 0.05.

First of all, there is no significant difference between condition 7 (UHD → UHD, L: UHD → UHD) and 8 (HD → HD, L: HD → HD), as they both show that the test participants found the identical videos with identical labels to be perceptually identical. Second, condition 5 (UHD → UHD, L: UHD

→ HD) received significantly different ratings from condition 7, and the same applies for condition 4 (HD → HD, L: HD → UHD) and 8. These obtained results mean that even though the stimuli did not differ between these conditions, the ratings were still heavily influenced by the labels.

Although the differences between condition 3 (UHD → UHD, L: HD → UHD) and 7, and condition 6 (HD → HD, L: UHD → HD) and 8 are measurable and also seem apparent from the histogram, they are not statistically significant. Third, there is no statistical difference between condition 1 (HD → UHD, L: HD → UHD), 3 and 4, and between condition 2 (UHD → HD, L: UHD → HD), 5 and 6. Therefore, it can be stated that the test participants perceived the identical videos the same way as the ones with actual visual differences. This indicates that the influence of the labeling effect on the subjective scores was evidently greater than what the test participants were able to perceive.

Study 2: 7-point scale

The rating tendencies for the 7-point scale were similar compared to the results obtained for the 3-point scale. However, the main difference here was that as test participants were given a greater freedom in the expression of quality comparison, which resulted in more scoring deviation. Furthermore, the usage of “slight” ratings consistently dominated the quality assessment for the test conditions where resolution change was indicated through the labels.

The statistical analysis for the 7-point scale is provided in Table 6. Here as well, we first conducted an ANOVA to check the general impact of conditions, which turned out significant ($df = 7, p = 0.0$). The Tukey, Bonferroni and

Table 5 Statistical analysis of the investigated test conditions (c_1 and c_2) in Study 1 (with labels, 3-point rating scale). The p values (p) of Tukey HSD multiple comparisons are given, along with significance (s)

c_1	Video	Label	c_2	Video	Label	p	s
1	HD → UHD	HD → UHD	2	UHD → HD	UHD → HD	0.000	*
1	HD → UHD	HD → UHD	3	UHD → UHD	HD → UHD	1.000	
1	HD → UHD	HD → UHD	4	HD → HD	HD → UHD	1.000	
1	HD → UHD	HD → UHD	5	UHD → UHD	UHD → HD	0.000	*
1	HD → UHD	HD → UHD	6	HD → HD	UHD → HD	0.000	*
1	HD → UHD	HD → UHD	7	UHD → UHD	UHD → UHD	0.022	*
1	HD → UHD	HD → UHD	8	HD → HD	HD → HD	0.000	*
2	UHD → HD	UHD → HD	3	UHD → UHD	HD → UHD	0.000	*
2	UHD → HD	UHD → HD	4	HD → HD	HD → UHD	0.000	*
2	UHD → HD	UHD → HD	5	UHD → UHD	UHD → HD	1.000	
2	UHD → HD	UHD → HD	6	HD → HD	UHD → HD	1.000	
2	UHD → HD	UHD → HD	7	UHD → UHD	UHD → UHD	0.000	*
2	UHD → HD	UHD → HD	8	HD → HD	HD → HD	0.096	
3	UHD → UHD	HD → UHD	4	HD → HD	HD → UHD	0.987	
3	UHD → UHD	HD → UHD	5	UHD → UHD	UHD → HD	0.000	*
3	UHD → UHD	HD → UHD	6	HD → HD	UHD → HD	0.000	*
3	UHD → UHD	HD → UHD	7	UHD → UHD	UHD → UHD	0.096	
3	UHD → UHD	HD → UHD	8	HD → HD	HD → HD	0.000	*
4	HD → HD	HD → UHD	5	UHD → UHD	UHD → HD	0.000	*
4	HD → HD	HD → UHD	6	HD → HD	UHD → HD	0.000	*
4	HD → HD	HD → UHD	7	UHD → UHD	UHD → UHD	0.006	*
4	HD → HD	HD → UHD	8	HD → HD	HD → HD	0.000	*
5	UHD → UHD	UHD → HD	6	HD → HD	UHD → HD	1.000	
5	UHD → UHD	UHD → HD	7	UHD → UHD	UHD → UHD	0.000	*
5	UHD → UHD	UHD → HD	8	HD → HD	HD → HD	0.047	*
6	HD → HD	UHD → HD	7	UHD → UHD	UHD → UHD	0.000	*
6	HD → HD	UHD → HD	8	HD → HD	HD → HD	0.132	
7	UHD → UHD	UHD → UHD	8	HD → HD	HD → HD	0.632	

Holm tests were conducted in the same fashion as with the 3-point scale.

Similarly to the 3-point scale, there is no significant difference between condition 7 and 8 (identical pairs with correct labels). There are statistically significant differences between condition 3, 5 and 7 (UHD → UHD with different labels), and between condition 4, 6 and 8 (HD → HD with different labels), resulting in stronger corresponding conclusions compared to the 3-point scale. Finally, there is no statistical difference between condition 1, 3 and 4 (L: HD → UHD with different video resolutions), and between condition 2, 5 and 6 (L: UHD → HD with different video resolutions), just as in the case of the 3-point scale.

Tests without labels

The results of the tests where labels were not present during the experiment are shown in Fig. 7a, b, with histograms of the ratings for the 3-point and the 7-point scale, respectively. There are 4 distinct test conditions, since labels were not

shown. All test conditions given in Table 1 were separately assessed, but they are clustered in the analysis (3, 5 and 7; 4, 6 and 8), as they were identical not only in content, but from the perspective of the test participants as well. Again, test conditions 3, 5 and 7 had the same video stimuli (UHD → UHD), just as 4, 6 and 8 did (HD → HD), but there were no labels to differentiate them.

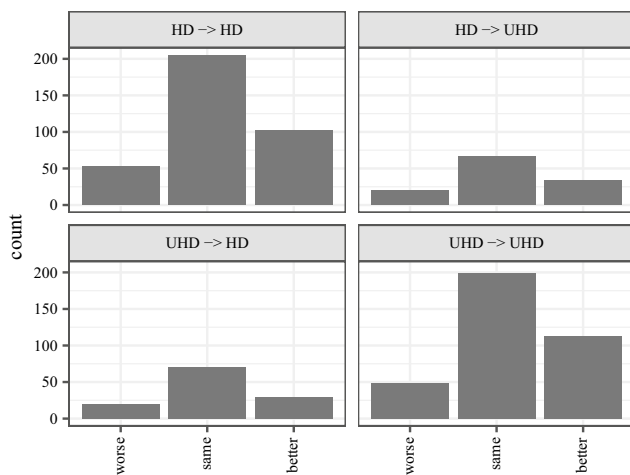
Study 3: 3-point scale

For all four investigated test conditions, the histograms of the ratings indicate similar tendencies: users primarily identify sequences as the “same” quality, followed by rating the second stimulus better (with roughly half as many ratings). Preference of the first stimulus received the fewest scores (again with roughly half as many ratings as the previous option).

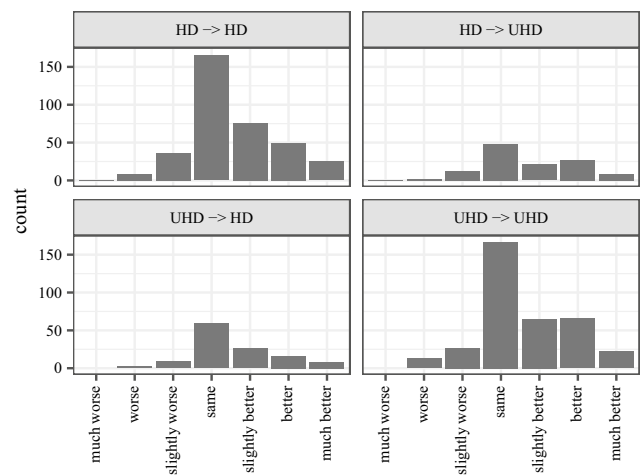
Based on an ANOVA conducted between conditions and ratings, we could see no significant impact for the 3-point scale ($df = 7, p = 0.829$). All Tukey, Bonferroni and Holm

Table 6 Statistical analysis of the investigated test conditions (c_1 and c_2) in Study 2 (with labels, 7-point rating scale). The p values (p) of Tukey HSD multiple comparisons are given, along with significance (s)

c_1	Video	Label	c_2	Video	Label	p	s
1	HD → UHD	HD → UHD	2	UHD → HD	UHD → HD	0.000	*
1	HD → UHD	HD → UHD	3	UHD → UHD	HD → UHD	0.971	
1	HD → UHD	HD → UHD	4	HD → HD	HD → UHD	0.726	
1	HD → UHD	HD → UHD	5	UHD → UHD	UHD → HD	0.000	*
1	HD → UHD	HD → UHD	6	HD → HD	UHD → HD	0.000	*
1	HD → UHD	HD → UHD	7	UHD → UHD	UHD → UHD	0.000	*
1	HD → UHD	HD → UHD	8	HD → HD	HD → HD	0.000	*
2	UHD → HD	UHD → HD	3	UHD → UHD	HD → UHD	0.000	*
2	UHD → HD	UHD → HD	4	HD → HD	HD → UHD	0.000	*
2	UHD → HD	UHD → HD	5	UHD → UHD	UHD → HD	0.951	
2	UHD → HD	UHD → HD	6	HD → HD	UHD → HD	0.962	
2	UHD → HD	UHD → HD	7	UHD → UHD	UHD → UHD	0.000	*
2	UHD → HD	UHD → HD	8	HD → HD	HD → HD	0.000	*
3	UHD → UHD	HD → UHD	4	HD → HD	HD → UHD	0.999	
3	UHD → UHD	HD → UHD	5	UHD → UHD	UHD → HD	0.000	*
3	UHD → UHD	HD → UHD	6	HD → HD	UHD → HD	0.000	*
3	UHD → UHD	HD → UHD	7	UHD → UHD	UHD → UHD	0.000	*
3	UHD → UHD	HD → UHD	8	HD → HD	HD → HD	0.000	*
4	HD → HD	HD → UHD	5	UHD → UHD	UHD → HD	0.000	*
4	HD → HD	HD → UHD	6	HD → HD	UHD → HD	0.000	*
4	HD → HD	HD → UHD	7	UHD → UHD	UHD → UHD	0.006	*
4	HD → HD	HD → UHD	8	HD → HD	HD → HD	0.000	*
5	UHD → UHD	UHD → HD	6	HD → HD	UHD → HD	1.000	
5	UHD → UHD	UHD → HD	7	UHD → UHD	UHD → UHD	0.000	*
5	UHD → UHD	UHD → HD	8	HD → HD	HD → HD	0.003	*
6	HD → HD	UHD → HD	7	UHD → UHD	UHD → UHD	0.000	*
6	HD → HD	UHD → HD	8	HD → HD	HD → HD	0.002	*
7	UHD → UHD	UHD → UHD	8	HD → HD	HD → HD	0.906	



(a) Study 3: Test without labels, 3-point rating scale.



(b) Study 4: Test without labels, 7-point rating scale.

Fig. 7 Histogram of test ratings without labels

values consequently indicate the lack of any statistically significant difference between any given two test conditions, hence the detailed results table is omitted. This means that there was no evident visual difference between the UHD video stimuli and the upscaled HD videos. This conclusion is reinforced by the similarity in scoring between condition 1 and 2, which are technically the opposites of each other.

As for the repeated scoring distribution, it can be linked to a simple assessment bias due to the lack of clear visual differences. Although it was not emphasized during the training phase that the stimuli *will* visually differ, with 64 paired comparisons, it is not difficult for a test participant to get the feeling that there *should* be a difference. Furthermore, as the visualization on the large UHD TV was generally pleasing and there were no additional impairments implemented, it was easier to rate the second stimulus to be the better one, via memory bias targeting the first one.

Study 4: 7-point scale

The quality assessment of the test conditions using the 7-point scale resulted in similar but more deviating tendencies compared to the usage of the 3-point scale. Here, the ANOVA also shows no significant impact of the condition on the ratings ($df = 7$, $p = 0.693$). The Tukey, Bonferroni and Holm statistical analysis also conclude that lack of significant differences in the subjective scores and hence are not shown in detail.

At a first glance of the results of study 3 and 4, one may question the validity of the subjective tests, since the UHD → HD comparisons did not conclude a worsening visual experience. In fact, the distribution in their histograms are very much like the results of HD → UHD comparisons. However, there are a couple of facts that must be taken into consideration: (1) By using Lanczos upscaling—which is known to perform better than other methods in this context [7]—and by avoiding lossy video compression, the HD stimuli were of high quality without any additional visual degradation. (2) Although the source videos were not “eye-candy” materials, even HD videos with software/GPU/device upscaling can appear generally impressive to individuals on such a large, flat, modern display. (3) Apart from test conditions 1 and 2, the pair contained identical videos sequences. (4) With many comparisons to assess, an individual in such scenario may easily feel that there *should* be a difference, even if there is genuinely none. (5) The video stimuli in a pair were shown after each other and not simultaneously, and therefore the memory bias could degrade the first while the second one was watched.

Considering all the above, such results are feasible, valid, and not surprising. Again, while such ratings do question the ability of individuals to distinguish UHD and upscaled HD

videos based on professional but not “eye-candy” contents, they most certainly do not question the added value of UHD visualization in general. The discussion on the properties of the experiment and the test design choices are further detailed later in the paper.

Content dependency

We further checked whether the content itself could have impacted the ratings, as previous research has often shown. In our statistical analysis, similar to the tests performed in the previous sections, multiple comparisons using Tukey HSD were carried out, using the SRC as independent variable and the ratings of the test participants as dependent variables. The analysis was first run on the data grouped by experiment type and scale type (112 comparisons), and then also grouped by condition (896 comparisons).

The results generally indicate that the content did *not* play a significant role in the subjective assessment. In the first analysis, only 2 out of 112 comparisons concluded significant differences, and for the second one, this was 1 out of 896. These differences were both measured for the experiment without labels, using the 3-point scale.

Without grouping by conditions, the ratings of SRC05 (first clip of Tears of Steel) significantly differed from SRC06 (second clip of Tears of Steel) and SRC07 (first clip of El Fuente). When the sources were separately analyzed for each test condition, SRC05 and SRC07 showed a statistically significant difference, too. We could, however, not identify any content characteristics that could have caused this difference, particularly looking at the spatio-temporal information (see Fig. 2) and frequency spectra (Fig. 3).

To summarize, it appears that the content itself had only a minor influence on the quality ratings. While the difference between the frequency components of the different movies is evident from Fig. 3, these differences are not strong enough to cause measurable differences in rating behavior.

Rating scale correspondence

One of our research questions asked about whether the used rating scale would impact the results, and whether the two scales could be compared. For the analysis presented in this subsection, the ratings obtained via the 7-point scale were therefore collapsed and mapped onto the options of the 3-point scale, in order to demonstrate the differences in scale usage. This means that the three positive and three negative comparisons of the 7-point scale were converted into one option each (e.g., both “Slightly better”, “Better” and “Much better” become “Better”). The resulting scoring distribution is presented on Fig. 8. It is clearly shown that there are strong differences that we have observed previously on Fig. 6a, b; during the experiment

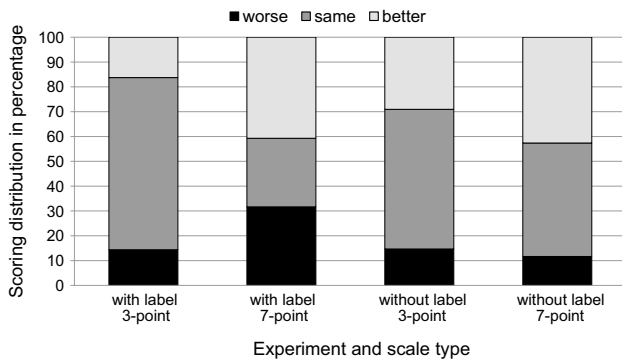


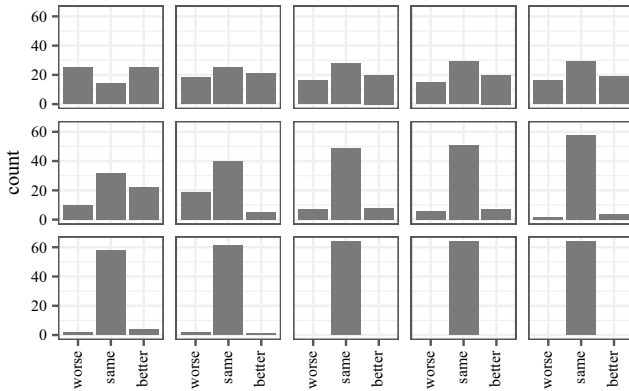
Fig. 8 Rating distributions mapped onto the 3-point scale

with labels present, while the 3-point scale produced 69.37% of the ratings to indicate the lack of visual difference between the stimuli, the corresponding value with

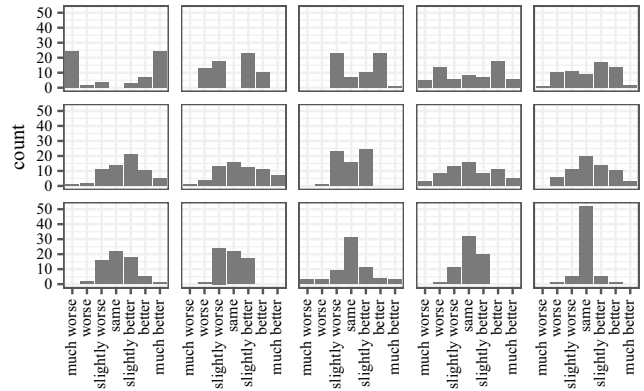
the 7-point scale was only 27.6%. A similar tendency is present for the results of the experiment without labels, but the extent is less strong. Here, only the positive scores increased, due to the aforementioned assessment bias.

Per-subject rating behavior

Let us now have a look at the ratings of the test subjects individually. The histograms for the tests with and without labels are shown on Figs. 9a, b, 10a, b. A histogram represents the rating behavior of a single individual, and as there were 15 test participants per study, there is a total of 60 histograms in these figures. The histograms are ordered by the usage of the middle option (“Same”) in order to enhance the visualization of the differences between rating behaviors.

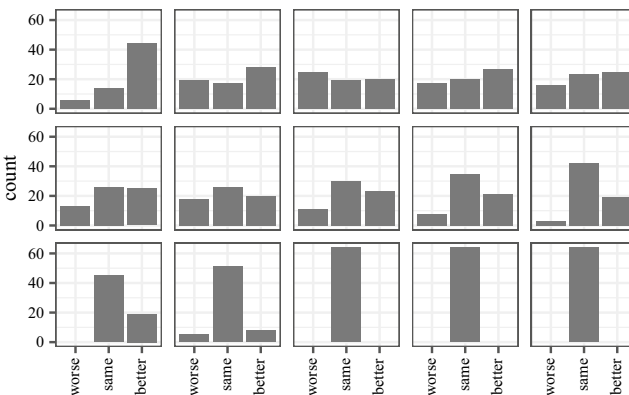


(a) Study 1: 3-point rating scale.

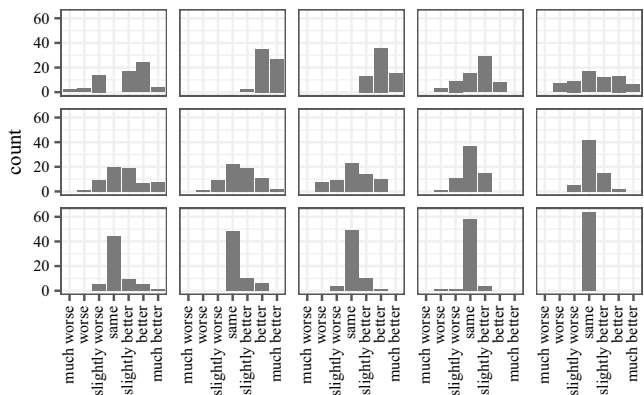


(b) Study 2: 7-point rating scale.

Fig. 9 Histogram of per subject test ratings with labels



(a) Study 3: 3-point rating scale.



(b) Study 4: 7-point rating scale.

Fig. 10 Histogram of per subject test ratings without labels

Certain rating behavior extremes stand out at first glance, such as the scores of test participants who did not distinguish any stimuli in the pairs and therefore provided 64 identical ratings. This applied to seven individuals from the entire pool of test subjects (more than 10%). The opposite is worth mentioning as well, where test participants avoided this specific rating option (no stimuli pair was assessed as the same). This was only present for the 7-point scale, which provided three times as many options to rate visual differences.

These individual results can be matched with the test stimuli and with the labels. Matching subjective ratings with the stimuli tells us the achieved rating correctness, that is, the correlation between what resolution was used and how it was reflected in the scores. Matching subjective ratings with the labels indicates a sort of obedience to the labels, as it shows how much the test participants agreed with that the labels suggested. The criteria of rating correctness and compliance with labels are shown on Table 7. For example, if the first video (V_1) was HD, and the second one (V_2) was UHD, then the options “Slightly better”, “Better” and “Much better” were correct in this sense.

Rating correctness

The results of the rating correctness analysis are presented on Fig. 11. One bar in the figure represents the rating correctness of a single individual. When labels were present during the experiment, the 3-point and the 7-point scale produced average values of 61.04% and 40.52%, respectively. When labels were not included in the experiment, the corresponding 3-point and 7-point scale averages were 47.71% and 41.56%. In both experiments, the 3-point scales achieved higher percentages of rating correctness. This is partially due to the fact that it did not enable the rating freedom of the 7-point scale, and thus more subjective assessments deemed the stimuli to be the same, as shown on Fig. 8. As 75% of the stimulus pairs contained identical videos, those test participants who used only the corresponding option in the scales evidently achieved a rating correctness of 75%. Furthermore, this was in fact the highest level of measured rating correctness for both experiments and both scale types.

Table 7 Criteria of rating correctness (based on the resolution of V) and compliance with labels (based on label L)

V_1/L_1	V_2/L_2	Rating
HD	HD	Same
HD	UHD	Slightly better, Better or Much better
UHD	HD	Slightly worse, Worse or Much worse
UHD	UHD	Same

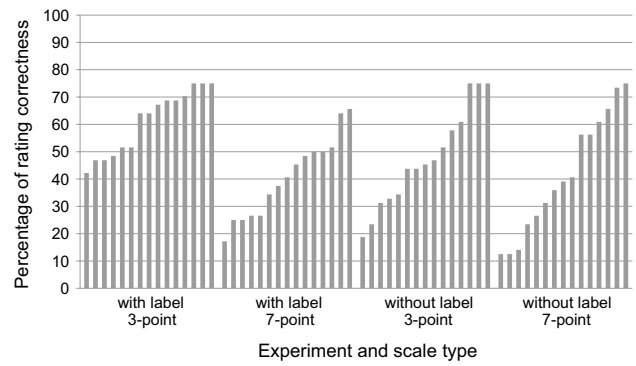


Fig. 11 Percentage of rating correctness

Compliance with labels

The results of the analysis on the compliance with labels are presented on Fig. 12. One bar in the figure represents the compliance rate of a single individual. Only half of the test participants were involved in this analysis, as the other half participated in tests without labels. While the average compliance for the 3-point scale was 43.43%, the corresponding value for the 7-point scale was 59.27%. Again, the 7-point scale made it possible for the test participants to indicate smaller differences via the options “Slightly worse” and “Slightly better”. This fact is quite relevant to this analysis, since it supports the marking of the perceived differences evoked by cognitive bias. Furthermore, two participants achieved 100% compliance, which means they never disagreed with the labels. As for those participants who only used the middle option in the test, their compliance value was 25%, since only 25% of the presented labels suggest the lack of difference.

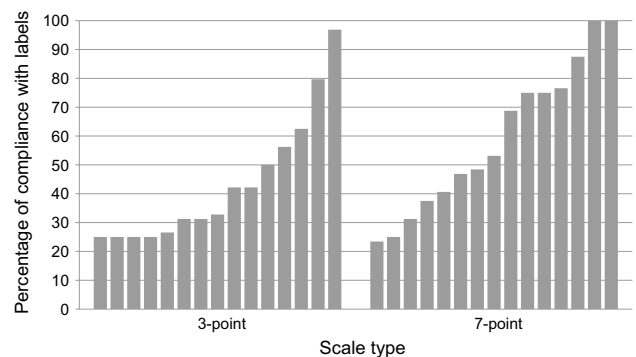


Fig. 12 Percentage of compliance with labels

Post-experiment questionnaire

Common questions

In this analysis, the data collected by the five post-experiment questionnaire items is addressed. As a reminder, a 20-option quasi-continuous scale ranging from -10 to 10 was used, with the exclusion of the middle option 0. Negative and positive scores are represented by dark and light columns, respectively, and the results are separately shown for the two experiment types (with and without labeling). As the results are presented on histograms, the height of a single bar reports the number of votes on a given option. Extreme scores are towards the edge of the figures; for example, the number of ratings indicating the lowest level of mental demand is represented by the dark column on the left end of the figure, while the corresponding highest level can be found on the right.

Both experiments were assessed similarly regarding mental demand (see Fig. 13). One could expect that labels mentally support subjective quality evaluation as their presence may guide the observer, yet it did not result in any significant difference. When labels were shown, the average rating was -0.33, with 11 negative and 19 positive scores, and without labels, the corresponding values were -0.16, 12 and 18. The distribution of scores was similar as well, with many test participants indicating either very low or slight mental demand.

According to the results, the experiments were less demanding physically than mentally (see Fig. 14). The ratio of positive and negative values were roughly the opposite, with 10 positive and 20 negative scores for both experiments. The average values were -2.96 and -2.4, with and without labels, respectively. Approximately a third of the test participants expressed very low physical demand, and cases of high demand were rarely registered.

The pace of the experiment was not deemed to be hurried or rushed. Again, the obtained results did not differ in a statistically significant extent (see Fig. 15). With labels, the

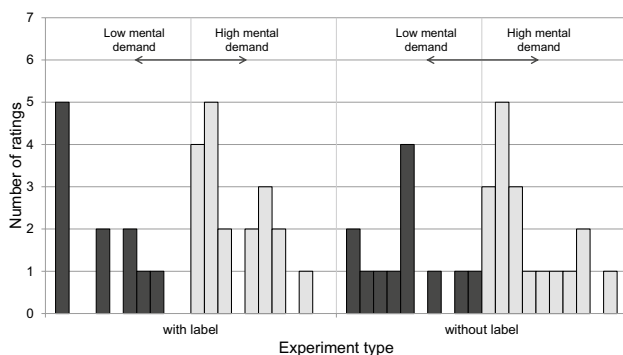


Fig. 13 Histogram of the results on mental demand

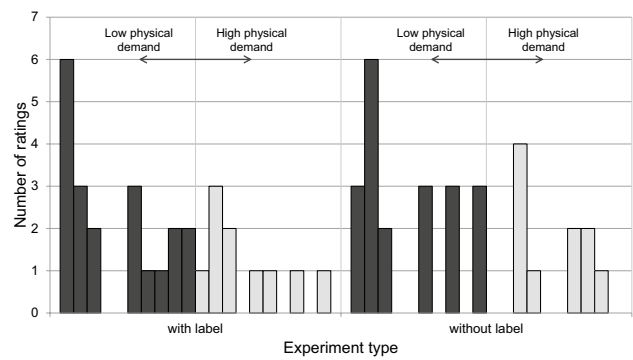


Fig. 14 Histogram of the results on physical demand

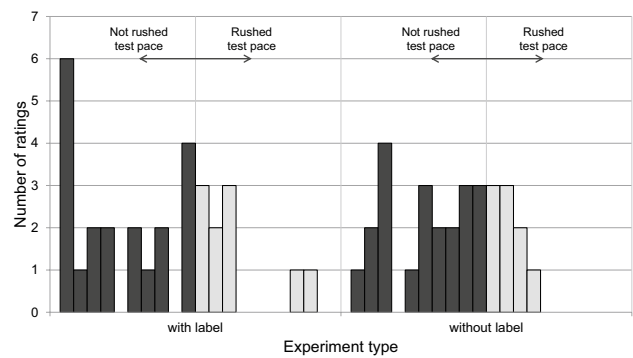


Fig. 15 Histogram of the results on test pace

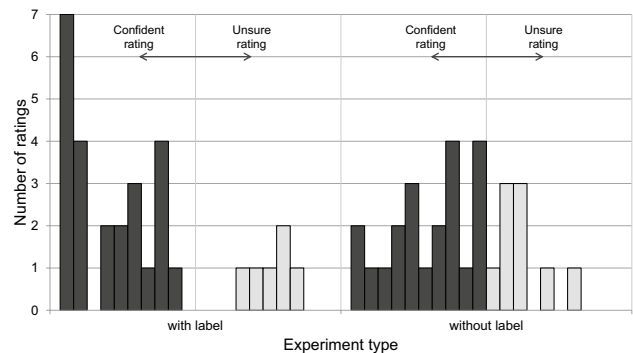


Fig. 16 Histogram of the results on task success

average was -3, with 10 positive and 20 negative scores. Without labels, these were -2.83, 21 and 9. For the two experiments combined, there were only a total of 2 scores between 5 and 10 on the scale, indicating that only 2 out of 60 test participants considered the test structure (see Fig. 4) to be too rushed.

Compared to the previous components of the questionnaire, the results on the self-assessment of scoring task success show much greater differences (see Fig. 16). Yet the differences even for these ratings are not statistically

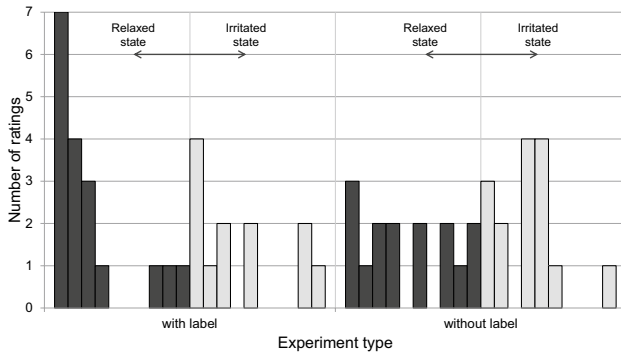


Fig. 17 Histogram of the results on irritation

significant, due to the high deviation of scores. The averages were -4.26 and -2.4 , with 24 and 21 positive, and 6 and 9 negative scores, for the experiments with and without labels, respectively. Such results can point out that the presence of labels can improve the overall rating confidence of the test participants. Therefore, ironically, it needs to be noted that distorted, biased ratings are submitted with more confidence.

The greatest scoring difference in the post-experiment questionnaire was achieved for the last common component, regarding the irritation of the test participant (see Fig. 17). Although no statistically significant difference was found, the averages were -3.1 and -1 , with 18 and 15 positive, and 12 and 15 negative scores, for the experiments with and without labels, respectively. With labels, 14 test participants gave a score of -8 or lower, indicating the lack of annoyance, while without labels, only 6 did. Apparently, the guidance provided by labels reduced the overall level of irritation during the experiment.

Preference statement with labels

When labels were presented, test participants were asked whether they found UHD/4K to be better than HD video or not. The results were balanced, as 11 test participants claimed 4K to be better, 10 stated the opposite, and 9 could not come to a conclusion based on the observed video stimuli and their labels.

Claimed source of perceived difference

At the end of both post-experiment questionnaires, test participants were asked about the “source of difference” between the video stimuli in the pairs. The indicated reasons were diverse, but we categorized them into a list of keywords (see Fig. 18). The last category “combinations of the above” refers to cases where the test participants included at least two difference types in their answers; e.g., if a test participant claimed that the video stimuli differed in colors and frame rate, then in this analysis, they are added to

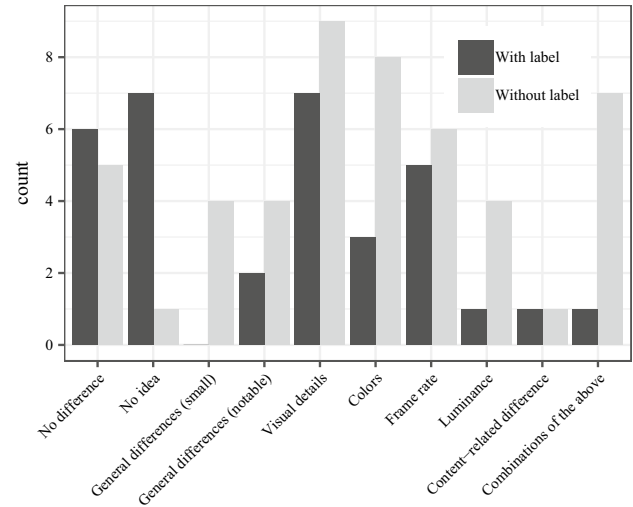


Fig. 18 Histogram of the “source of difference” in the post-experiment questionnaire

categories “colors”, “frame rate” and also “combinations of the above”. A notable category is “no difference”, where the test participants stated that no visual difference could be perceived between the stimuli. These are also well-reflected in the per-subject rating analysis (see Figs. 9a, b, 10a, b). When labels were present, 7 out of 30 test participant could not explain what the difference was or whether there was a difference at all. This number was only 1 when no labels were provided. From these 7 test participants in the experiment with the labels, 4 of them could not explain what UHD was and 3 had never heard about UHD before; 3 had never seen UHD before and 4 did not know whether they had seen UHD before or not; 4 did not possess a UHD-capable device and 3 were unsure about the resolution of their device. Summa summarum, none of those who could not determine the source of the perceived visual difference knew what UHD was, had seen UHD videos prior to the experiment and possessed a UHD-capable device, or at least was not aware of it.

The relevance of this information is that people generally refrain from providing an answer than providing a wrong, incorrect answer. These seven test participants perceived differences between the video stimuli—as reflected by their scores—and were aware of the suggested reason via the labels, but they did not have prior experience with the given resolution and thus avoided answering this question of the post-experiment questionnaire in order to prevent a technically false statement. Although each and every test participant was precisely informed that data was handled confidentially and that no registered rating or answer could be linked to any individual, many of them still were afraid of being judged based on their lack of knowledge on the subject. As an illustrative example, one of these test participants particularly commented after the test,

“I saw that there was some sort of a difference, but I just didn’t want to write something stupid.”

Again, when no labels were present, this applied to only one test participant. There were, however, 24 test participants who experienced differences and provided feedback on the matter. Generally all keywords appeared more frequently in this experiment, compared to the test with labels. It is important to note that when no labels guided the test participants, 7 out of the 24 identified multiple types of visual differences—e.g., frame rate and colors – while with labels, it only occurred once.

General differences varied a lot in ways of phrasing, formulating. Small differences indicate that the test participants managed to detect visual differences, but they were either difficult to perceive and/or did not have a significant impact on the experience. Notable differences include cases when test participants made a general remark about the visual quality, such as “4K is nicer” or stating that “everything” is different. Answers regarding visual details were the most relevant to the actual perceivable differences, yet in numerous cases, test participants noticed differences in colors, frame rate and luminance (brightness).

Frame rate is a particularly interesting aspect, as multiple test participants reported in the test with labels that HD had better frame rate. This is analogous to the findings regarding HDR QoE [2]. For each and every stimulus, the frame rate was constant, unvarying, yet differences between the stimuli in the pairs were experienced, due to the concept of the trade-off between frame rate and other quality aspects. For example, one of the aforementioned participants stated the UHD had better visual details, but HD had better frame rate. In the experiment without labels, changes in frame rate were indeed indicated as well, but not in such manner.

We must clearly state that the video stimuli did *not* differ in colors, frame rate and luminance at all; the stimuli only varied between the two resolutions, according to the investigated test conditions (see Table 1). Yet the test participants experienced differences in these aspects. This phenomenon can be explained through the process of cognitive dissonance

reduction [35]. The test participants were presented 128 short videos (64 pairs), where the stimuli in the pairs did not differ significantly, if they even differed at all. In fact, 3 out of 4 pairs showed identical videos in both experiments. Among many other factors, the sole number of video stimuli can evoke a cognition that suggests that “there should be a difference”. When labels were presented, the theoretical difference was indicated, but there was no information on how that would manifest in the perceived quality. Without labels, the only hint a participant could have extracted was from the pre-experiment questionnaire, asking three questions about UHD. In many of the cases, the cognition “there should be a difference” was matched with the perception “there is no difference”, and the latter was overruled in order to eliminate this dissonant cognitive state.

Correlation between questionnaire results and quality ratings

Beyond providing a general insight for the pool of test participants, the questionnaire results can also be used to enhance the understanding of the obtained quality ratings via correlation analysis. The quality ratings were clustered by the subjects’ answers to the questionnaire, and the average ratings were compared separately for experiment (with or without labels), rating scale and test condition, resulting in 96 statistical tests. The results are shown in Table 8. Only 9 out of 96 multiple comparisons indicated statistically significant differences in ratings.

As an example of how knowing about the subjects improves our understanding of the results, the first question in the pre-experiment questionnaire (see Table 3) asked about familiarity with UHD/4K. The first line in Table 8 tells us that when labels were involved and the 3-point scale was used, and both video stimuli were UHD, but the first one was labeled as HD, those most familiar with the terms “UHD/4K” favored the second stimulus (with the UHD label) significantly more compared to those who only heard about the terms.

Table 8 Comparison between options (o_1 and o_2) for the first question of the pre-experiment questionnaire (“Have you ever heard of “Ultra HD”, “UHD” or “4K”?”). The presence of label, scale type (s) and conditions (c) are indicated, the p values of Tukey HSD (p) are given, along with the option with higher quality ratings (o). For the meaning of answers, see Table 3

Label	s	c	o_1	o_2	p	o
With	3	3	Yes (1)	Yes (2)	0.029	Yes (1)
With	3	8	Yes (1)	Yes (2)	0.007	Yes (1)
With	3	8	Yes (1)	No	0.013	No
With	7	3	Yes (1)	No	0.047	Yes (1)
With	7	6	Yes (2)	No	0.004	No
With	7	8	Yes (1)	No	0.020	No
With	7	8	Yes (2)	No	0.045	No
Without	3	5	Yes (1)	No	0.029	No
Without	3	8	Yes (1)	Yes (2)	0.043	Yes (1)

The results for the second and the third item of the pre-experiment questionnaire are shown in Tables 9 and 10, respectively. In case of the third question, the fourth answer type—which is technically a lack of information—was not included in this analysis, as in practice, it may be any of the first three answers, and thus its inclusion does not provide any meaningful conclusion.

For the first question, the majority of significant difference was found in the experiment with labels, for both scale types. The second question was similarly balanced for scale types, but also for experiment types. Significant differences in the third one were more for the experiment without labels, dominantly for the 7-point scale. One of the most notable phenomena in the analysis is that for the third question, the scores of those who confirmed owning a UHD-capable device in the experiment without labels, using the 7-point scale, were always significantly higher for the second

stimulus in every test condition than the scores of those who did not have a such a display.

The results for the first question of the post-experiment questionnaire in the study with labels are shown in Table 11. With two types of scales, eight conditions and three answer types, there were a total of 48 statistical comparisons, from which 28 resulted in significant differences. The most important finding here is that 25 out of 32 comparisons involving the first answer—stating that UHD is better than HD—show statistically significant differences, commonly preferring the stimuli with the UHD label. Furthermore, 6 out of these 7 comparisons where no significant difference was found either belonged to condition 7 and 8, where the labels suggested no difference. Note that the ratings linked to the first answer differ from those associated not only with the second answer—stating that UHD is not better than HD—but with the third one—stating the lack of a confident answer—as well. Wherever there were significant differences between

Table 9 Comparison between options (o_1 and o_2) for the second question of the pre-experiment questionnaire (“Have you seen a video in UHD / 4K resolution yet?”). The presence of label, scale type (s) and conditions (c) are indicated, the p values of Tukey HSD (T) are given, along with the option with higher quality ratings (O). IDK = “I don’t know”

Label	s	c	o_1	o_2	T	O
With	3	1	Yes	No	0.048	Yes
With	7	3	Yes	No	0.013	Yes
With	7	3	Yes	IDK	0.022	Yes
With	7	6	Yes	IDK	0.031	IDK
With	7	8	Yes	IDK	0.034	IDK
Without	3	5	Yes	No	0.023	No
Without	3	5	Yes	IDK	0.009	IDK
Without	7	1	Yes	No	0.023	Yes
Without	7	5	Yes	No	0.048	Yes
Without	7	8	Yes	No	0.001	Yes

Table 10 Comparison between options (o_1 and o_2) for the third question of the pre-experiment questionnaire (“Do you possess a device with UHD / 4K resolution?”). The presence of label, scale type (s) and conditions (c) are indicated, the p values of Tukey HSD (p) are given, along with the option with higher quality ratings (o). IDK = “I don’t know”

Label	s	c	o_1	o_2	p	o
With	7	1	No	IDK	0.017	IDK
With	7	2	No	IDK	0.032	No
With	7	4	No	IDK	0.000	IDK
With	7	5	No	IDK	0.048	No
With	7	7	No	IDK	0.028	IDK
Without	3	5	Yes	No	0.028	No
Without	3	5	Yes	IDK	0.005	IDK
Without	3	7	Yes	No	0.031	No
Without	7	1	Yes	No	0.006	Yes
Without	7	2	Yes	No	0.003	Yes
Without	7	3	Yes	No	0.021	Yes
Without	7	4	Yes	No	0.003	Yes
Without	7	5	Yes	No	0.001	Yes
Without	7	6	Yes	No	0.010	Yes
Without	7	7	Yes	No	0.014	Yes
Without	7	8	Yes	No	0.000	Yes

Table 11 Comparison between options (o_1 and o_2) for the question of the post-experiment questionnaire “Would you say that 4K video is better than HD video?”, in the study with labels. The scale types (s) and conditions (c) are indicated, the p values of Tukey HSD (p) are given, along with the option with higher quality ratings (o). IDK = “I don’t know”

s	c	o_1	o_2	p	o
3	1	Yes	No	0.000	Yes
3	1	Yes	IDK	0.000	Yes
3	2	Yes	No	0.000	No
3	2	Yes	IDK	0.000	IDK
3	3	Yes	No	0.000	Yes
3	3	Yes	IDK	0.000	Yes
3	4	Yes	No	0.000	Yes
3	4	Yes	IDK	0.000	Yes
3	5	Yes	No	0.000	No
3	5	Yes	IDK	0.000	IDK
3	6	Yes	No	0.000	No
3	6	Yes	IDK	0.000	IDK
3	8	Yes	IDK	0.036	IDK
7	1	Yes	No	0.000	Yes
7	1	No	IDK	0.035	IDK
7	2	Yes	No	0.000	No
7	2	Yes	IDK	0.000	IDK
7	3	Yes	No	0.002	Yes
7	3	Yes	IDK	0.031	Yes
7	4	Yes	No	0.000	Yes
7	4	Yes	IDK	0.001	Yes
7	4	No	IDK	0.003	IDK
7	5	Yes	No	0.000	No
7	5	Yes	IDK	0.001	IDK
7	5	No	IDK	0.039	No
7	6	Yes	No	0.000	No
7	6	Yes	IDK	0.002	IDK
7	7	Yes	No	0.008	Yes

the second and the third answer or all three answers, the first one favored the stimulus with the UHD label the most, then the third one, and finally the second.

Discussion

The results presented in the previous section highlight the potential impact of the labeling effect in the context of perceived quality. As detailed in the beginning of the paper, the labeling effect is unavoidable—simply inevitable—in real life. In fact, labels are desired by manufacturers and content providers; they serve commercial purposes and drive user decisions. Therefore, should QoE studies make use of labels more often, in order to reduce the gap between the experience measured in the lab and what is actually experienced in real life? As we would like to discuss in this section, they may induce unnecessary cognitive bias for many types of subjective tests, where the inclusion of labels would do more “harm” to the collected results than the realism it could

bring. The lessons learned from the results we achieved (labeling may have a stronger influence than actual visual properties) can also be employed for discussing the general usefulness of labels in QoE studies.

Labels in QoE studies

The topic of the labeling effect in QoE studies cannot simply be dealt with by saying “avoid labeling in the experimental design”. Information that may bias the test participants should be categorized based on the type of information and how the test participants encounter it. Both can be explicit and implicit. While *explicit information* requires no additional effort for processing, *implicit information* needs to be derived from the experimental environment, that is, test participants are left to discover it themselves. The tests with labels shown in this paper involved *explicit* information, which was provided to the test participants in an obvious manner. Labels can also be supplied implicitly. For example, if the label is the brand of the display, but the test calls no direct attention to it, then it depends on the test participants whether they notice it or not. Such implicit handling should be avoided in general, unless the research question particularly demands it (e.g., if the study aims to find out whether the test participants notice certain information or not), or the goal of the test is to deceive participants and steer their attention to other seemingly more important aspects of the study. An experimenter cannot reliably steer how much attention subjects will pay to seemingly irrelevant details of the test, and hence the results *without* explicitly hinting subjects at labels may be—somewhat counterintuitively—more biased than if the test design called for subjects to take into account the labels.

How often these labels are shown is also an important decision to make. The experiment we conducted showed the labels very frequently, i.e., the test participants were notified of the (purported) video resolution before every single stimulus. This study could be repeated with a single notification in the beginning—e.g., analogous to condition 1, 3 and 4 (see Table 1)—which applies to all stimulus pairs, but the test is conducted without the repetition of this information. In this case, the influence of the labeling effect may diminish over time.

Contrasting the results of our experiment with the ones presented in the related literature, we can see that our tests have been using more design elements from a Quality of Experience and User Experience viewpoint than standardized video quality tests. Typically, such tests in the area of UHD have often been on “visual quality only”, with ratings solely focusing on minuscule details of compression performance or scaling quality. Once we move away from this paradigm and attach labels to the stimuli, our interpretation

of the results may change significantly. Hence, a careful shift towards more realism in testing may reveal factors that go beyond low-level techniques in the high-resolution domain.

How would we and how should we design such studies? What experimental methodology, which combination of the aforementioned techniques and parameters would be closest to real-life scenarios? Let us examine two cases. The first situation is a person buying a UHD-capable display in a shop. In a QoE study, this could correspond to an experiment on willingness to pay, as the financial decision at hand predominantly depends on the quality. The labels would be explicit and emphasized, covering technical capabilities and price. The labels—especially the price—would either be frequently (explicitly) or constantly (explicitly or implicitly) shown to the test participants. The test could be performed with either a single apparatus or multiple; although having multiple displays is more realistic, a single one is sufficient as well, if the different stimuli (with different quality parameters) represent different displays, and the one used in the experiment has the sufficiently high capabilities to properly accommodate all stimuli.

The second situation would be a person watching a UHD-capable display in a home scenario. When modeling such a situation in a QoE study, it is important to distinguish whether the display has been recently purchased or it has been used for a while. The reason of its relevance is post-decision dissonance [66, 67]. It is a process of decision justification (e.g., “buying this TV was a good choice”), in which—similarly to other forms of cognitive bias—the perception of quality may be affected. If this effect is not excluded from the experimental configuration, then the subjective test needs to include a user decision prior to quality assessment, and the labels play a more significant role as well. If a study does not take the post-decision dissonance into consideration, it is sufficient to explicitly present the labels at the beginning of the test. The relevance of labels in the modeling of such scenario is that a person is usually aware of fundamental information regarding his or her TV. While some people may know their displays better than others, and such knowledge may easily fade, one tends to remember at least the property that convinced him or her towards the decision of purchase. Furthermore, in many parts of the world, people change their displays more frequently (handhelds and larger screens alike), and therefore more decisions are made, such decisions reinforce each other (e.g., brand loyalty), and knowledge on display properties has less time to fade.

To summarize, we see a potential in using experimental techniques focused both on visual properties as well as the underlying users’ or buyers’ psychology—the (ab)use of labels being one element of such studies.

Experimental design and source contents

One could argue that the experiments presented in this paper did not make use of the full potential of UHD, as the source contents were not as visually appealing as typical UHD demo videos. It is true that the test participants were not shown, for example, slow-motion close-up macro shots of the human eye or the wings of an exotic butterfly. However, as it has already been discussed earlier in the paper, it is not typical or even realistic user behavior to use a UHD display solely for such demo materials in a home scenario of multimedia consumption.

The source videos of this experiment were diverse in the sense that the contents included CGI animation, live-action clips and CGI-enhanced live-action scenes as well. When we designed the experiment, there was a reasonable thought that the choice of the source videos would affect the results, i.e., that there would be statistically significant impact of the content, or that the content could lead to no visual differences being apparent. As it has been presented in the analysis, when grouped by the test conditions (and therefore any other factor was ruled out), only 0.11% of all the comparisons differed significantly.

Professional contents usually intend to artistically exploit the technology they use. For example, if a movie is shot in stereoscopic 3D, it is expected to have at least one scene where the added value of 3D is justified through the visuals. The same idea can be applied to UHD contents as well. However, it needs to be noted that not all professional content shot in UHD considers the resolution during production, as the term “professional content” is not limited to high-budget feature films, and in fact, it is not even limited to movies. Furthermore, as a quality feature integrates into the use case scenarios and becomes the de facto standard, the will of the content creator(s) to emphasize it diminishes.

Unlike the professional content of the movie industry, where the cameras are handled by individuals with the necessary expertise and the scenes are adjusted in a way to provide the desired visual quality, user contents have no such criteria. The handling of the camera is highly emphasized here, as the captured noise and blur can easily degrade the added value of a higher resolution, as other studies have found.

Test subject behavior and the “ground truth”

In the analysis of the results, test subject behavior was addressed in the form of correctness and compliance with labels, which reflected our main intention of the studies: to see how much labels could distort the rating behavior of the test participants. The metric of rating correctness could also describe how much subjective scores are in alignment with the *objective* quality of the stimuli.

In other words, it would be possible to utilize objective metrics and then measure how much the ratings agree with their results, which would allow us to also quantify the impact of additional factors like the labels. In this case, there would be no need for a second set of experiments without labels, which we had to perform in order to know what the “true” impact of the labels on subject behavior would be. However, as we have observed, current video quality metrics are inadequate for comparing uncompressed UHD sequences. Hence, knowing the “ground truth” becomes a complex task in technological areas for which not enough validated tools are available, particularly if that ground truth is strongly influenced by other, non-technical parameters. This calls for properly conducted between-subject tests rather than relying on purely instrumental measures of quality.

Conclusions

Summary of findings

The paper has presented a series of experiments addressing the influence of the labeling effect on the perception of HD and UHD video. The obtained results indicate that the labeling effect had a significant impact on the subjective scores, regardless of test condition and source content. The corresponding study without labels concludes the lack of statistical difference between the two video resolutions, regardless of rating scale. However, the choice of rating scale greatly affected the test with the labels, as the more fine-grained 7-point comparison scale enabled the expression of the slighter perceived differences, in contrast to the 3-point scale.

Potential continuation of research

Future work on the topic has the potential to further study the labeling effect in the context of UHD video. The study using explicit labels could be repeated with implicit labels, and then of course a post-experiment questionnaire that records whether the test participants considered the label or not. Longer sequences could be used, not only to investigate QoE over time, but also to address the fading of the labeling effect in studies with labels that are either frequently presented or only once in the beginning of the test. The investigation of QoE over time could track voting times, to consider the potential fatigue of test participants in such exhaustive studies. The role of source content could also be addressed in more detail, using videos ranging from low-quality user-generated content to exceptional demo materials. Furthermore, user decisions could be involved in order to investigate

the effect of post-decision dissonance on the perceived quality. Finally, the experiments presented in this paper and all their potential continuations could be performed using 4K and 8K instead of HD and UHD.

Acknowledgements The authors would like to thank all the test participants and everyone who made the presented research possible.

Compliance with ethical standards

Conflict of interest All authors state that there is no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Sackl A, Schatz R, Raake A (2017) More than I ever wanted or just good enough? User expectations and subjective quality perception in the context of networked multimedia services. *Qual User Exp* 2(1):3
2. Kara PA, Cserkaszkzy A, Martini MG (2018) Premium HDR: the impact of a single word on the Quality of Experience of HDR video. In: International conference on multimedia and expo (ICME), emerging multimedia systems and applications (EMSA). IEEE
3. ITU-R (2014) Rec. BT.2020, Parameter values for ultra-high definition television systems for production and international programme exchange
4. ITU-R (2012) Rec. BT.2022, General viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays
5. Bae SH, Kim J, Kim M, Cho S, Choi JS (2013) Assessments of subjective video quality on HEVC-encoded 4K-UHD video for beyond-HDTV broadcasting services. *IEEE Trans Broadcast* 59(2):209–222
6. Tanaka Y, Ochi D (2014) QoE assessment methodologies for 4K video services. *NTT Tech Rev* 12(5):2014
7. Li J, Koudota Y, Barkowsky M, Primon H, Le Callet P (2014) Comparing upscaling algorithms from HD to Ultra HD by evaluating preference of experience. In: Sixth international workshop on quality of multimedia experience (QoMEX), pp 208–213
8. Weerakkody R, Mrak M, Baroncini V, Ohm J-R, Tan TK, Sullivan GJ (2014) Verification testing of HEVC compression performance for UHD video. In: 2014 IEEE global conference on signal and information processing (GlobalSIP). IEEE, pp 1083–1087
9. Berger K, Koudota Y, Barkowsky M, Le Callet P (2015) Subjective quality assessment comparing UHD and HD resolution in HEVC transmission chains. In: Seventh international workshop on quality of multimedia experience (QoMEX), pp 1–6
10. Xie F, Pourazad MT, Nasiopoulos P, Slevinsky J (2016) Determining bitrate requirement for UHD video content delivery. In: IEEE international conference on consumer electronics (ICCE). IEEE, pp 241–242
11. Van Wallendael G, Coppens P, Paridaens T, Van Kets N, Van den Broeck W, Lambert P (2016) Perceptual quality of 4K-resolution

- video content compared to HD. In: 2016 eighth international conference on quality of multimedia experience (QoMEX). IEEE, pp 1–6
12. Zhu Y, Song L, Xie R, Zhang W (2016) SJTU 4K video subjective quality dataset for content adaptive bit rate estimation without encoding. In: 2016 IEEE international symposium on broadband multimedia systems and broadcasting (BMSB). vol 2016-July. IEEE, pp 1–4
 13. Sotelo R, Joskowicz J, Anedda M, Murrioni M, Giusto DD (2017) Subjective video quality assessments for 4k uhdtv. In: 2017 IEEE international symposium on broadband multimedia systems and broadcasting (BMSB). IEEE, pp 1–6
 14. Mackin A, Afonso M, Zhang F, Bull D (2018) A study of subjective video quality at various spatial resolutions. In: 2018 25th IEEE international conference on image processing (ICIP). IEEE, pp 2830–2834
 15. Huynh-Thu Q, Garcia M-N, Speranza F, Corriveau P, Raake A (2011) Study of rating scales for subjective quality assessment of high-definition video. *IEEE Trans Broadcast* 57(1):1–14
 16. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 13(4):600–612
 17. Netflix (2017) Toward a practical perceptual video quality metric. <https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652>
 18. ITU-T (2017) P.1203: Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport
 19. Hanhart P, Korshunov P, Ebrahimi T (jul 2013) Benchmarking of quality metrics on ultra-high definition video sequences. In: 2013 18th international conference on digital signal processing (DSP). IEEE, pp 1–8
 20. Mead GH (1934) *Mind, self and society*, vol 111. University of Chicago Press, Chicago
 21. Tannenbaum F (1938) *Crime and the community*. Ginn and Co., Boston
 22. Wilson PR (1968) Perceptual distortion of height as a function of ascribed academic status. *J Soc Psychol* 74(1):97–102
 23. Zhao W, Zhou X (2011) Status inconsistency and product valuation in the california wine market. *Organ Sci* 22(6):1435–1448
 24. Masson J, Aurier P, d’hauteville F (2008) Effects of non-sensory cues on perceived quality: the case of low-alcohol wine. *Int J Wine Bus Res* 20(3):215–229
 25. d’Hauteville F, Fornerino M, Philippe Perrouy J (2007) Disconfirmation of taste as a measure of region of origin equity: an experimental study on five french wine regions. *Int J Wine Bus Res* 19(1):33–48
 26. Lick E, König B, Kpossa MR, Buller V (2017) Sensory expectations generated by colours of red wine labels. *J Retail Consum Serv* 37:146–158
 27. Sinha I, Batra R (1999) The effect of consumer price consciousness on private label purchase. *Int J Res Mark* 16(3):237–251
 28. Johansson JK (1989) Determinants and effects of the use of “made in” labels. *Int Mark Rev* 6(1):47–58
 29. Hamzaoui L, Merunka D (2006) The impact of country of design and country of manufacture on consumer perceptions of binational products’ quality: an empirical model based on the concept of fit. *J Consum Mark* 23(3):145–155
 30. Ahmed SA, d’Astous A (1995) Comparison of country of origin effects on household and organizational buyers’ product perceptions. *Eur J Mark* 29(3):35–51
 31. Batra R, Ramaswamy V, Alden DL, Steenkamp J-BE, Ramachander S (2000) Effects of brand local and nonlocal origin on consumer attitudes in developing countries. *J Consum Psychol* 9(2):83–95
 32. Loftus EF, Hoffman HG (1989) Misinformation and memory: the creation of new memories. *J Exp Psychol General* 118(1):100
 33. Tversky A, Kahneman D (1985) The framing of decisions and the psychology of choice. In: *Environmental Impact assessment, technology assessment, and risk analysis*. Springer, pp 107–129
 34. Gächter S, Orzen H, Renner E, Starmer C (2009) Are experimental economists prone to framing effects? A natural field experiment. *J Econ Behav Organ* 70(3):443–446
 35. Festinger L (1962) *A theory of cognitive dissonance*, vol 2. Stanford University Press, Stanford
 36. Brunnström K, Beker SA, De Moor K, Dooms A, Egger S, Garcia M-N, Hossfeld T, Jumisko-Pyykkö S, Keimel C, Larabi M-C et al (2013) Qualinet white paper on definitions of quality of experience
 37. Jacoby J, Olson JC, Haddock RA (1971) Price, brand name, and product composition characteristics as determinants of perceived quality. *J Appl Psychol* 55(6):570
 38. Verbeke W, Viaene J (1999) Consumer attitude to beef quality labeling and associations with beef quality labels. *J Int Food Agribus Mark* 10(3):45–65
 39. Burton S, Biswas A, Netemeyer R (1994) Effects of alternative nutrition label formats and nutrition reference information on consumer perceptions, comprehension, and product evaluations. *J Public Policy Mark* 13:36–47
 40. Szybillo GJ, Jacoby J (1974) Intrinsic versus extrinsic cues as determinants of perceived product quality. *J Appl Psychol* 59(1):74
 41. Richardson PS, Dick AS, Jain AK (1994) Extrinsic and intrinsic cue effects on perceptions of store brand quality. *J Mark* 58:28–36
 42. DeVecchio D (2001) Consumer perceptions of private label quality: the role of product category characteristics and consumer use of heuristics. *J Retail Consum Serv* 8(5):239–249
 43. Heisey FL (1990) Perceived quality and predicted price: use of the minimum information environment in evaluating apparel. *Cloth Text Res J* 8(4):22–28
 44. Rieh SY, Belkin NJ (1998) Understanding judgment of information quality and cognitive authority in the www. In: *Proceedings of the 61st annual meeting of the American society for information science*, vol 35, pp 279–289
 45. Rieh SY, Belkin N (2000) Interaction on the web: Scholars’ judgment of information quality and cognitive authority. In: *Proceedings of the annual meeting-American society for information science*, vol 37. *Information Today*; 1998, 2000, pp 25–38
 46. Schöffler M (2017) Overall listening experience—a new approach to subjective evaluation of audio. Doctoral thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)
 47. Lamm K, Mandl T, Womser-Hacker C, Greve W (2010) The influence of expectation and system performance on user satisfaction with retrieval systems. In: *EVIA@ NTCIR*, pp 60–68
 48. Lamm K, Mandl T, Womser-Hacker C, Greve W (2010) User experiments with search services: methodological challenges for measuring the perceived quality. In: *Proceedings of the 3rd workshop on perceptual quality of systems, PQS*
 49. Szajna B, Scamell RW (1993) The effects of information system user expectations on their performance and perceptions. *Mis Q* 17:493–516
 50. Bouchard S, Dumoulin S, Talbot J, Ledoux A-A, Phillips J, Monthuy-Blanc J, Labonté-Chartrand G, Robillard G, Cantamessa M, Renaud P (2012) Manipulating subjective realism and its impact on presence: preliminary results on feasibility and neuroanatomical correlates. *Interact Comput* 24(4):227–236
 51. Sackl A, Masuch K, Egger S, Schatz R (2012) Wireless vs. wireline shootout: how user expectations influence quality of experience. In: *Fourth international workshop on quality of multimedia experience (QoMEX)*. IEEE, pp 148–149

52. Sackl A, Zwickl P, Egger S, Reichl P (2012) The role of cognitive dissonance for QoE evaluation of multimedia services. In: IEEE globecom workshops. IEEE, pp 1352–1356
53. Kara PA, Bokor L, Imre S (2013) Distortions in QoE assessment of 3D multimedia services on multi-access mobile devices. In: IEEE 9th international conference on wireless and mobile computing, networking and communications (WiMob). IEEE, pp 311–318
54. Kara PA, Bokor L, Imre S (2014) 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, pp 539–545
55. Sackl A, Egger S, Zwickl P, Reichl P (2012) The QoE alchemy: turning quality into money. experiences with a refined methodology for the evaluation of willingness-to-pay for service quality. In: 2012 fourth international workshop on quality of multimedia experience (QoMEX). IEEE, pp 170–175
56. Sackl A, Zwickl P, Reichl P (2012) From Quality of Experience to willingness to pay for interconnection service quality. In: Networking workshops. Springer, pp 89–96
57. Higgs B, Polonsky MJ, Hollick M (2005) Measuring expectations: forecast vs. ideal expectations. Does it really matter? *J Retail Consum Serv* 12(1):49–64
58. Zeithaml VA, Berry LL, Parasuraman A (1993) The nature and determinants of customer expectations of service. *J Acad Mark Sci* 21(1):1–12
59. Kara PA, Bokor L, Sackl A, Mourão M (2015) 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, pp 1–6
60. Ebbing J (2017) The halo effect of smartphone brands: smartphone-s brand equity influence on the user experience of third-party smartphone applications. Master's thesis, University of Twente
61. ITU-T (2008) Rec. P.910, Subjective video quality assessment methods for multimedia applications
62. Katsavounidis I, Aaron A, Ronca D (2015) Native resolution detection of video sequences. In: SMPTE 2015 annual technical conference and exhibition, vol 1100. IEEE, pp 1–20
63. Hart SG, Staveland LE (1988) Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. In: *Advances in psychology*, vol 52. Elsevier, pp 139–183
64. Kara PA, Robitza W, Pinter N, Martini MG, Raake A, Simon A (2019) HD/UHD comparison test data. <https://zenodo.org/record/2561049>
65. Kara PA, Robitza W, Raake A, Martini MG (2017) 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
66. Festinger L (1964) Conflict, decision, and dissonance. Stanford University Press
67. Knox RE, Inkster JA (1968) Postdecision dissonance at post time. *J Personal Soc Psychol* 8(4p1):319

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