Quality and User Experience (2019) 4:3
https://doi.org/10.1007/s41233-019-0025-5

RESEARCH ARTICLE

A common framework for the evaluation of psychophysiological visual quality assessment

Sebastian Bosse[1] · Kjell Brunnström[2,3] · Sebastian Arndt[4] · Maria G. Martini[5] · Naeem Ramzan[6] · Ulrich Engelke[7]

Received: 14 July 2018

Abstract
The assessment of perceived quality based on psychophysiological methods recently gained attraction as it potentially overcomes certain flaws of psychophysical approaches. Although studies report promising results, it is not possible to arrive at decisive and comparable conclusions that recommend the use of one or another method for a specific application or research question. The video quality expert group started a project on psychophysiological quality assessment to study these novel approaches and to develop a test plan that enables more systematic research. This test plan comprises of a specifically designed set of quality annotated video sequences, suggestions for psychophysiological methods to be studied in quality assessment, and recommendations for the documentation and publications of test results. The test plan is presented in this article.

Keywords Video quality · Psychophysiology · Quality assessment · Subjective tests · Electroencephalography · Video quality expert group · VQEG

Introduction

In multimedia systems the quality of the received signal is ultimately evaluated by humans. Due to the lack of understanding of human perception and general processes underlying quality formation, the reliable assessment of perceptual quality builds on psychophysical judgment tests in which a human observer gives an overt response on the quality of the presented signal. Traditionally, subjective quality assessment is performed using questionnaires, either open-ended or based on psychometric scales, such as n-point Likert scales. As valuable as these studies are, they are based on conscious responses by the participants and often do not provide sufficiently deep insight into underlying perceptual and cognitive processes. Moreover, the explicit task of giving a judgement response interferes with natural viewing behavior, and consequently, with natural viewing experience ('Schrödinger's cat of quality assessment'). Psychophysical quality assessment is furthermore restricted to supra-threshold stimuli and as such to consciously detectable distortions.

In psychophysical quality assessment, responses of individual subjects are condition-wise averaged in order to arrive at the mean opinion score (MOS), the de-facto metric for the quantification of perceptual quality. Unfortunately, responses of categorical scales, such as Likert-scales, should

✉ Naeem Ramzan
naeem.ramzan@uws.ac.uk

Sebastian Bosse
sebastian.bosse@hhi.fraunhofer.de

Kjell Brunnström
kjell.brunnstrom@ri.se

Sebastian Arndt
sebastian.arndt@ntnu.no

Maria G. Martini
M.Martini@kingston.ac.uk

Ulrich Engelke
ulrich.engelke@data61.csiro.au

1 Fraunhofer Institute for Telecommunications – Heinrich Hertz Institute, Einsteinufer 37, 10587 Berlin, Germany

2 RISE Research Institutes of Sweden AB (Acreo), Box 1070, 164 25 Kista, Sweden

3 Mid Sweden University, Sundsvall, Sweden

4 Norwegian University of Science and Technology, O.S. Bragstadsplass 2B, 7491 Trondheim, Norway

5 Kingston University, London, UK

6 University of the West of Scotland, Hamilton, UK

7 CSIRO Data61, Kensington, WA 6152, Australia

Published online: 16 July 2019

be considered as ordinal data, rather than interval data, for which summary statistics such as mean or standard deviation are not appropriate representations [45].

As a potential remedy for these flaws of psychophysical quality assessment and in order to shed light on the cognitive processes underlying quality formations, researchers recently started to study psychophysiological approaches to quality assessment [8, 16, 22]. These methods aim at bypassing or complementing overt responses of subjects by the measurement of physiological responses that are related to perceived quality. Over the last years a lot of progress has been made: many psychophysiological correlates of perceived quality have been identified, experimental paradigms have been proposed and data analysis methods have been studied and results are promising. However, each experimental setup is based on a multitude of design decisions, comprising the psychophysiological signal to be studied, the device used to record the signal, the way how stimuli are presented, the (potentially machine learning-based) methods to analyze the data—and the stimuli used itself. Most studies presented in the literature vary in all of these design decisions, which makes it almost impossible to arrive at a conclusive comparison of different psychophysiological quality assessment approaches. At the same time, however, it is not widely understood and agreed upon how such experiments can and should be performed. There are basically no best practice guides and data sets available that allow to perform such research effectively and in a reproducible manner.

To study these novel approaches to quality assessment, the Video Quality Expert Group (VQEG) started the Psycho-Physiological Quality Assessment (PsyPhyQA) project. As a step forward to more systematic and comparable research, PsyPhyQA developed a test plan for the investigation of psychophysiological methods for visual quality assessment. One of the main contributions of the test plan is a set of distorted and undistorted video sequences that was specifically designed to be used in the research of psychophysiological quality assessment. This dataset available at https://www.cdvl.org/ as *PsyPhyQA Video Dataset*. With this article we follow the current trend towards stronger open science practices. By revealing details regarding planned experimental designs and data analyses and suggesting systematic procedures, transparency and replicability of basic QoE research will be increased and should ultimately benefit its practical application in the field.

This article presents the work of PsyPhyQA and summarizes the test plan. We invite researchers (also outside of VQEG) working on psychophysiological visual quality assessment to make use of the test plan, i.e. the proposed test sequences and suggested evaluations and analyses. Specifically, we believe that the field and community can strongly benefit from a commonly used dataset to make results of experiments more comparable.

The remainder of the article is structured as follows. "The video quality experts group" section starts with a brief presentation of VQEG and a description of its modus operandi. In "State-of-the-art of physiological measurements in quality assessment" section we present a brief state of the art in physiology for quality assessment. In "Data Set of Video Sequences" section we describe a data set that is the core of the test plan and was specifically designed for psychophysiological quality assessment. Psychophysical test procedures that we propose should accompany psychophysiological assessment studies at the current state of research are described in "Psychophysical tests" section. Physiological measurement" section describes experimental parameters that VQEG considers to study and sketches the experimental plans of PsyPhyQA. Experimental and methodological aspects and evaluations that should be documented for the sake of reproducible and comparable research are summarized in "Documentation of test results" section. Challenges and limitations are discussed in Challenges and limitations" section and the article is concluded in Conclusion" section.

The video quality experts group

The Video Quality Experts Group [50] was established in 1997 as a forum of international experts working in the field of perceptual video quality. VQEG is an international and independent group that is open to all interested organisations and individuals and does not require any membership or fees (see also [29] and [19]).

A very important tool for VQEG is the VQEG test plan, which defines exact procedures for performing scientific validation of subjective test procedures and objective models. These test plans describe the scope of the validation project, characteristics of source content, the scope and nature of video quality degradations, the subjective rating method and the subjective test environment and evaluation metrics. Importantly, the test plans are worked out and approved by consensus in advance amongst VQEG participants usually at the face-to-face meetings according to the voting rules [50].

State-of-the-art of physiological measurements in quality assessment

Measuring the reaction towards external stimulation directly without asking test participants explicitly can be an advantage. Interrupting test participants while experiencing multimedia content can disrupt their level of immersion and therefore can influence their quality of experience (QoE). Furthermore, converting a subjective opinion onto a scale might be a challenge. Using direct measures from the body potentially minimizes these issues.

In the past, electroencephalography (EEG) has been shown to be a valid measure of the level of QoE of participants in a variety of multimedia contents in laboratory environments. In 2D still images [38], video [44], audio [4], as well as audiovisual [7] presentations, the P300 wave has been shown to be a good indicator of the users' experience. In all three domains, a short stimulus with varying quality was presented to the subject. Based on the stimulus presentation, a so-called event-related potential (ERP) is elicited in the users' brain and can be detected using electroencephalography [40]. ERPs are direct stereotyped electrophysiological responses to a specfic sensory, cognitive or motor event [40]. Within the ERP, the P300 wave is considered to be representing a measurement of difference between a standard and a target stimulus [21]. During the QoE experiment, the standard representation was the undistorted stimulus and the deviant the distorted stimulus. The common result of these experiments, using different modalities, is that the stronger the degradation was, the larger and earlier the P300 amplitude rose to its maximum. The work presented in [48] investigated whether there is one specific dimension in audio distortions that contributes overproportional to the generation of the P300. Due to the stimuli selection, no such component could be identified in this study. However, EEG not only allows to discern perceived quality from perceived intensity level, but also in terms of distinctive quality dimensions, such as "discontinuity", "noisiness" and "coloration" [49]. Also other ERP components are candidates for neural markers of perceived quality, e.g., the P1 component for visual comfort due to vertical disparities in stereoscopic images [9]. Different to transient ERPs, steady-state visual evoked potentials (SSVEP) [41] are evoked by periodically changing visual stimulus. The feasibility of SSVEP for image quality assessment was shown conceptually in [12, 13]. In [1], SSVEP have been used for the neurally informed detection of perceived image distortions. In [14] it was shown that the prediction of the MOS from a single observer's SSVEP response is statistically indistinguishable the prediction from a single observer's overt rating.

In contrast to seeing how the brain is reacting towards an immediate change in quality, it is also possible to analyze how the brain state is changing when being exposed to longer sequences of low-quality multimedia content. Here, different sub-frequency bands of the EEG signal are analyzed. When recording and analyzing an EEG, the recorded data can be divided into different subbands. Each of the bands is associated with a different mental state. A variety of audio-only [3], and audiovisual experiments [6] was conducted in which the quality was varied within the presentation of the multimedia content. The general conclusion from these experiments was that a lower quality leads to a larger portion of alpha and delta activity compared to high quality sequences. An increase in these sub-bands is associated with subjects becoming mentally more fatigued, as a higher workload is required to follow the presented content [5]. Although EEG currently appears to be the psychophysiological method that is most widely used in quality assessment, also other measurement methods such as near-infrared spectroscopy are studied and show promising results [24]. For more thorough reviews on this topic we refer the reader to [8, 16, 22]. In [42] the potential on how to use different assessment methods in the context of immersiveness as a part of QoE are discussed.

What can be seen from this brief review is that different equipment, different paradigms and different analysis are used. The pure fact that these differ is not problematic per se. However, a standardized way of reporting is needed in order to be able to compare results from different studies. It is obvious that recordings from a consumer grade system will have a different data quality than those from clinical grade systems. Furthermore, although a variety of experiments have been conducted in different laboratories, no systematic cross-lab validation using physiological measures in the domain of QoE has been performed. In the case of inter-lab studies not only the different locations of the laboratories would be variable. This will give the opportunity to follow the exact same experimental protocol, and report systematically about differences in equipment used. Thus, investigating the potentially different outcomes would be of major interest, as these are not only affected by the general experimental design, but also by finer differences in the experimental setup, such as the interstimulus distance in ERP-based approaches [40], or the stimulation frequency in SSVEP-based approaches [15]. Furthermore and most crucially, in order to move towards practical applicability it is necessary to systematically evaluate to what extent and precision psychophysiological assessment methods actually work outside isolated and overly well-controlled experimental setups.

Data set of video sequences

A central aspect of the testplan is a data set of impaired videos that was specifically designed to be used for research on psychophysiological quality assessment and made publicly available. This section describes the selected source reference sequences (SRC), the hypothetical reference circuit (HRC) considered and the resulting processed video sequences (PVS). At the current state of research, the preparation of psychophysiological quality assessment studies is often very time consuming and many trials are needed. Therefore, the number of SRCs and HRCs, and thus the number of resulting PVS, is intentionally restricted. This prevents researchers from being forced to select (between different studies or laboratories potentially disjoint) subsets

(a) CityFly **(b)** CostumesRun **(c)** CostumesSearching

(d) ManInFountain **(e)** PeopleRun **(f)** RunInWoods

Fig. 1 First frames of the video sequences included in the testplan

from the dataset, but rather to enable them to study psychophysiological assessment method on the full dataset. The dataset can be downloaded as *PsyPhyQA Video Dataset* from https://www.cdvl.org/.

Source reference sequences

The target for the test plan is to get 10 s long PVSs, which allow for sufficient visual stimulation for analysis. Stimulus onset introduces transient neural responses as well as transient codec behavior and we need some time for the stimulus onset related transients to fade away, so the length of the SRC sequences has, therefore, been set to 12 s.

Six video sequences of Full-HD (1920 by 1080 pixels) resolution, a frame rate of 50 fps, and a duration of 12 s, i.e. 600 frames, were selected as SRC sequences. All of them are cut outs from a 6.5-min-long video produced by the Swedish Television (SVT) [25]. Fairytale, as the film is called, was professionally filmed and produced on 65 mm analogue film in 50 fps (slow motion up to 100 fps) and then scanned frame by frame while color correcting and applying film grain noise reduction, to produce the 4K (3840×2160 progressive, 16 bit per color) Master. The 1080p version was produced by downsampling the Master using a sinc filter. For more details on the production see [25], where image examples are showing that there is hardly any film grain noise left in the downsampled version. We have therefore judged that Fairytale, although a bit old, still is a very good source video material. There are 10 s cut outs suggested in their original distribution, but since in this test plan 12 s are targeted new cut outs with new names have been produced, for minimizing confusion with original cut outs.

Table 1 summarizes the selected SRC sequences. Three sequences have a large overlap with the original cut outs (10 s, 500 frames): PeopleRun overlaps with CrowdRun's frames (starting with frame 7111); CostumRun overlaps with PassingBy's frames (start frame 14131) and RunInWoods with PrincessRun (starting with frame 10429). Under the constraint of copyright considerations and with the goal of a practically sized testset, SRCs were selected

Table 1 Selected SRC sequences and respective values of coding difficulty, spatial (SI) and temporal (TI) information

Name	Start frame	Coding difficulty	SI [min, max]	TI [min, max]
CityFly	3001	Moderate	[46.8 54.5]	[9.6 15.7]
PeopleRun	7001	Difficult	[77.5 96.1]	[19.9 33.4]
CostumesSearching	11,161	Difficult	[49.0 79.9]	[19.0 33.]
CostumesRun	14,001	Easy	[29.9 50.5]	[11.4 20.2]
ManInFountain	9701	Easy	[17.3 49.9]	[7.7 15.2]
RunInWoods	10,431	Difficult	[87.1 115.1]	[16.3 38.1]

All are cut-outs from SVT Fairytale, with resolution 1920×1080 pixels, frame rate 50fps and duration of 12 s, i.e., 600 frames

to cover a sufficiently large perceptual space as possible. This is quantified in terms of temporal (TI) and spatial (SI) information [51] as shown in Table 1. Representative frames of the selected SRC sequences are shown in Fig. 1.

Hypothetical reference circuits

In practice, compression is one of the major sources of impairments in visual signals. Therefore, and to generate stimuli with properties that are of practical relevance, state-of-the-art video compression was chosen for introducing distortions and SRCs were encoded with High Efficiency Video Coding (HEVC) [46]. For compression the HEVC reference encoder HM-16.0 [35] was used with the random access main profile from the JCT-VC common test conditions [17] and an intraframe period of 48 frames.

Four target quality levels were defined as (a) perception threshold (MOS $\approx$ 4.5); (b) close below perception threshold (MOS $\approx$ 4); (c) bad broadcast quality (MOS $\approx$ 3); and (d) severe distortions (MOS $\approx$ 2).

To obtain these quality levels, the corresponding SRC-specific QP values were determined in a pre-study and are summarized in Table 2. The quality scores of the resulting processed video sequences (PVS) is presented in "Processed video sequences" section below.

Table 2 SRC-specific set of quantization parameters (QPs) selected to obtain the desired quality levels

Source reference sequence	Selected QPs
CityFly	25, 34,40, 45
PeopleRun	27, 35, 41, 46
CostumesSearching	26, 36, 43, 49
CostumesRun	27, 35, 41, 46
ManInFountain	22, 30, 37, 43
RunInWoods	25, 34, 40, 46

Processed video sequences

The perceptual quality scores of the processed video sequences (PVS) resulting from the SRC sequences described in "Source reference sequences" section, affected by the HRCs described in "Hypothetical reference circuits" section, were validated in a psychophysical cross-lab study by Fraunhofer HHI, Kingston University London, and the University of West Scotland. Subjective assessment in all laboratories employed degradation category rating (DCR) on a 5-point degradation scale according to [32]. Mean opinion scores (MOS) were obtained by aggregating individual quality ratings of all observers collected in all three laboratories after screening according to [31]. The resulting perceptual qualities are plotted versus QPs in Fig. 2; the results for different SRC sequences are plotted in different colors, with vertical bars denoting the 95% confidence interval of the MOS. Perceptual qualities per quality level are t-tested to be statistically indistinguishable across different SRC sequences with $p < 0.05$. The quality scores of the obtained PVSs fulfill the desired requirements specified in "Hypothetical reference circuits" section.

Psychophysical tests

In order to allow for single subject analysis psychophysiological tests should be accompanied by conventional behavioral subjective assessment. The part of the behavioral quality assessment should precede the psychophysiological recording in order to avoid tiredness of the subject while giving overt responses. Every subject should be asked to give overt quality ratings in response to the presentation of the PVS. The psychophysical test setup should be based on ITU Recommendations BT.500 or P.910 [31, 32]. The presentation procedure (test method, rating scale) is to be documented.

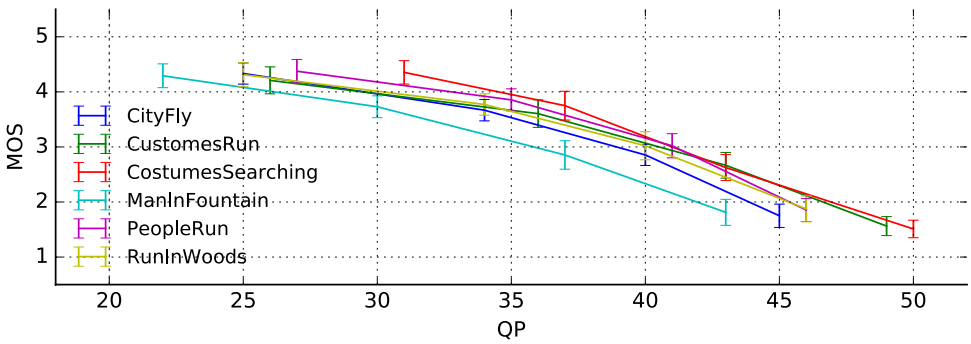

Fig. 2 Perceptual qualities of the PVS in dependence of the quantization parameter used for encoding. Vertical bars denote the 95%-confidence interval

Physiological measurement

A wide range of physiological measurement methods potentially feasible for the assessment of QoE exist [22]. As already indicated in "State-of-the-art of physiological measurements in quality assessment" section, the resulting experimental parameter space is tremendous and comprises the psychophysiological signal modality, the specific device used for signal recording, the stimulus presentation paradigm, and the specific experimental setup. Although the choice of a specific experimental designshould depend on the aim of a study, in practice laboratories face limitations with regard to the availability of recording devices and the experience of the experimenters. Therefore this test plan is not meant to prescribe any aspect of the experimental setup, but primarily intends to encourage laboratories to make use of the provided test material. However, PsyPhyQA decided to focus on electroencephalographic approaches in a first step and to evaluate ERP- and SSVEP-based assessment (cf. "State-of-the-art of physiological measurements in quality assessment" section) in cross-laboratory studies in order to gain insight into reproducibility and device dependency of these methods.

Documentation of test results

Given the variety of psychophysiological signals potentially carrying information about the perceived quality, this test plan leaves the choice and design of experimental setups to the laboratories. However, in order to obtain reproducibility and comparability of results, and to allow for systematic research, it is crucial that laboratories document the setup and data processing, and report results in a detailed manner.

Description of the experimental design

The insight that the experimental design has an impact on the test results in subjective quality assessment led to clear recommendations such as Recs. ITU-R BT.500 [31], ITU-T P.910 [32] and ITU-R BT.2022 [30]. These recommendations specify the experimental design for psychophysical quality assessment and are widely used by quality engineers and researchers. They comprise definitions of experimental parameters such as the ratio of inactive screen luminance to peak luminance (0.02), ratio of background luminance to picture's peak luminance (≤ 0.15), ratio of screen only black level luminance to peak white luminance (≈ 0.01), maximum observation angle relative to the normal (30°), background chromaticity of D65 and low other room illuminations. Besides these global aspects, different values of viewing distance are specified depending on the resolution of the test material.

While these general recommendations also hold for psychophysiological assessment, to date it is not clear how the demographics of test participants (e.g. age, gender, experience with quality tests), the number of participants, the session length, the number of trials per condition or the length of the test material affect the test results. About 15–28 participants is a range of widely accepted values for the cohort size in psychophysical tests [18, 34, 52]. In most psychophysiological quality assessment studies experiments were conducted with around 10 subjects [4, 7, 44]. The determination of the optimal (or minimally necessary) number of subjects is an open research question and should therefore be explicitly reported. One approach is to plan the size using a priori statistical power analysis [18], but then good estimate of the variance to be expected in the data needs to be established, which could be an outcome of the experiments done based on this article. The number of trials (number of repetitions) per condition is also, similarly, an open research question. According to [33], with frequent breaks a maximum of 3 h can be spent on rating PVS. This duration is subject to research in psychophysiological quality assessment; frequent breaks, however, are strongly recommended. The experimental design choice with regard to the duration is especially sensitive in studies that are particularly exhaustive, e.g. if it can be expected that fatigue might occur from the stimulus material. Subjects should be checked on color vision, visual acuity, and language used. As the demography of the subjects may have effect on results, the personal information with respect to age, gender, education and occupation can be obtained from subjects after getting signed consent that follows the GDPR directives [47].

For psychophysical quality assessment, video sequence durations in the range of 5 to 20 seconds [33] have been recommended. While this range appears reasonable for studying visual impairments, it might not be feasible if long term effects such as fatigue are investigated. Currently it is not clear if a psychophysiological response might be influenced, e.g., biased by the source content. Source content should therefore be chosen spanning a wide range of temporal and spatial information and to have neutral impact on the subjects [23]. A precise description of the stimuli utilized is thus crucial.

Several test procedures such as absolute category rating (ACR) [32] or Degradation Category Rating(DCR) [32] (or Double Stimulus ImpairmentScale (DSIS) [31]) have been proposed and studied for traditional psychophysical quality assessment. For psychophysiological quality assessment the distinct properties of different stimulus presentation methods

are vastly unknown. While it is well understood that, e.g., SSVEP are elicited by a stimulus presentation that is distinctly different to the stimulus presentation used to elicit ERP, the influence of experimental parameters such as the interstimulus period is unknown. For a better understanding, test procedures should be rigorously controlled and described.

Psychophysiological quality assessment relies, in contrast to psychophysical quality assessment, on the use of a device for measuring and recording the relevant psychophysiological signals. These devices differ in the measured signal, e.g. EEG, ECG or EMG, and the quality of the specific device, e.g. clinical or consumer grade devices. Thus, for a better understanding of the impact of the device quality, the manufacturer and model of the used device should be reported. Further important signal acquisition-related aspects are the recording frame rate and, if applicable, the sensor positions, reference and ground electrode positions and possible re-referencing.

Description of processing of neurophysiological data

Although the analysis and processing of neurophysiological data such as EEG recordings has been an active research field for several years [39], it has not been studied conclusively for applications in quality assessment [16]. The signal processing of the recorded neural data has typically several aspects; for systematic research, a detailed description is crucial.

In a first step the signal is roughly cleaned from drift artifacts, line noise, and high frequency artifacts by band-pass filtering and potentially downsampled [14, 44]. Filtered data is then commonly epoched with regard to a given temporal trigger, i.e. the stimulus onset, and simple statistical properties of the epoched signal, i.e. maximal values [7] or ratio of samples exceeding a certain threshold [14] were used as simple rejection method for epochs affected by motion artifact. For reproducible research, details of these preprocessing steps should be reported.

Neurophysiological signals are multidimensional time series with features that consist of spatial and temporal (and/or spectral) components [11, 27]. The properties of these features depend on the stimulus material (i.e. auditory and visual neural processing units have different locations and thus different spatial signatures), and the stimulus presentation (i.e. ERPs, evoked by an isolated, discrete stimulation, are commonly described by its temporal properties, whereas SSVEPs, evoked by a periodically repeated stimulation, are described by its spectral properties). Spatial feature extraction commonly leads to dimensionality reduction [26]. Research in brain–computer interfacing (BCI) shows that the selected method for spatial filter extraction has significant impact on system performance [11, 26, 27]. In the context of neurophysiological quality assessment different spatial decomposition techniques such as spatio-spectral decomposition [14], common spatial filters [1], linear discriminant analysis, or pre-selected single channels [4, 7] have been used. Temporal features, such as the delay and the magnitude of the ERP waveform [4, 7, 44], and spectral features such as the amplitudes of elicited SSVEP responses [14], have been found to be related to the perceived quality. However, for neurophysiological quality assessment, it is not clear yet what the advantages and disadvantages of different combinations of spatial decomposition/channel selection methods and temporal/spectral features are. A common set of stimulus material helps to answer these questions and researchers are strongly encouraged to study different feature extraction methods. For a better understanding of psychophysiological quality assessment a thorough evaluation of these methods is essential and adaptations and enhancements potentially required. Thus, detailed descriptions should be reported in order to allow for systematic research.

In quality assessment outliers can occur on trial-level and on subject-level. For conventional quality assessment a heuristic outlier removal method has been prescribed in [31]; recently, more sophisticated statistically-motivated approaches to outlier detection have been proposed [37]. As mentioned earlier in this subsection simple trial-wise outlier rejection methods are widely used in neurophysiological quality assessment. However, these methods rely on rather simple heuristic assumptions with regard to the recorded signal and thus lead to sub-optimal results. From BCI it is known that the user's profile has an impact on system performance [2] and the identification of the relation between users' EEG features and performance is ongoing research [2, 10, 36, 43]. In the context of quality assessment a subject-wise outlier rejection method was proposed based on subject-wise estimated spatial activation patterns [14]. While for practical psychophysiological quality assessment the rejection of outlier subjects is in general already beneficial, a predictor of subjects' performance a priori to an assessment session could greatly reduce costs and time. A common framework will allow for a comparative study of outlier rejection methods; for this, methods used should be described in detail. Feature extraction or regression schemes used in neurophysiological quality assessment may rely on supervised learning methods that make a cross-validated evaluation necessary. For comparative research, the details of cross-validation should be reported.

Summary of test results

It is expected that the results are reported in terms of correlation of the physiological measurement with MOS (Pearson correlation and Spearman rank) and standard error of

the processed psychophysiological signals, used as quality predictor.

Laboratories are further encouraged to report correlations between physiological and behavioral responses on a subject-basis. This would enable a more detailed analysis and comparison of the data. Based on the study of the relationship between physiological measurements and subjects' opinion scores, prediction models can be derived for mapping physiological responses to behavioral responses. Laboratories are encouraged to report such models derived from the results.

Challenges and limitations

As mentioned earlier, psychophysiological assessment has potentially several substantial advantages over traditional psychophysical approaches, including overcoming subjective rating scales and the interpretation thereof, instantaneous and implicit responses that do not require an explicit rating task, and more direct insights into internal processing that might allow for the assessment of pre-concious near-threshold artifacts in a signal, and a generally reduced influence of psychological biases.

However, before arriving at reliable practical application scenarios of psychophysiological quality assessment, a variety of challenges have to be addressed. Many of these challenges manifest themselves as variations of challenges known from traditional psychophysical quality assessment. Although we argue that, in contrast to psychophysical methods, psychophysiological assessment may provide insight into the perceptual and cognitive processes of quality formation, it is important to note that a neurophysiological response does not necessarily constitute a precise measurement of a subjective experience. The difficulty to probe this hides in the fact that the *true* experience itself is not available and any form of ground truth suffers from label noise.

Although providing an objective measurement, psychophysiological assessment is not free from systematic inter-subject variances. Screening methods analogous to psychophysical tests can help to identify subjects for which psychophysiological assessment is not feasible. Signal processing methods such as spatial filtering show promising results to understand and overcome inter-subject differences. On the other hand, this difference might also be understood as a marker of individual experience.

As discussed earlier, different types of stimulus presentation can elicit different neural responses. For any practical application it is important to understandwhich neural response, and thus, which stimulus presentation is the most appropriate one.

The choice of a device is typically based on a quality-cost trade-off under the constraint of limited financial resources. Recording a neurophysiological signal is inherently a noisy process. To avoid the derivation of erroneous conclusions based on the noise rather than on the stimulus related signal it is important to understand the quality of the response provided by the given measurement device and the impact of the noise characteristics on the assessment task.

The quality of the recording device and the strategy of how to deal with inter-subject variability are closely connected to the question of how many subjects and how many trials are necessary in order to arrive at statistically significant results. A recommendation regarding these values that holds empirically is currently not available. In order to avoid fallacies, the current lack of knowledge demands statistical and scientific rigor, as, in contrast to a priori power analysis, post hoc power analysis is fundamentally flawed [28]. Thus, power analysis to estimate the sufficient number of trials has to done before the actual data is recorded [20].

Addressing, studying and eventually overcoming these challenges is a major motivation for proposing the presented evaluation framework. However, it is important to acknowledge several important fundamental limitations of psychophysiological quality assessment that go beyond the presented framework. An obvious limitation and probably one of the biggest obstacles towards the usage of physiological assessment is the intrusiveness of physiological measurements and the duration of the preparation of the subjects, i.e. by attaching sensors. This renders viewing conditions, as it is also the case for traditional assessment, rather unnatural. The resulting burden, potentially experienced by subjects might reduce the availability of subjects. A thoughtful experimental design, e.g. incorporating little detection tasks, and the creation of a comfortable and pleasant atmosphere in the lab can mitigate but not eliminate these concerns. However, less intrusive future technology may bring a solution.

Compared to psychophysical assessment, the experimental setup and the data analysis required for psychophysiological assessment is very complex. Additional experimental factors and aspects render experiments much more error-prone, i.e. by improperly attached electrodes or by a lack of accuracy in the synchronization between stimulus presentation and data recording. Dealing with this increased complexity requires highly skilled experimenters.

An inherent problem of psychophysiological signal acquisition is the interference with noise and artifacts arising from subjects' muscle activity and body movement. This is a fundamental challenge when interactive multimedia services such as (cloud) gaming and virtual or augmented reality are combined with psychophysiological methods.

Conclusion

This paper presented and outlined the VQEG test plan on psychophysiological quality assessment. While the main purpose of the test plan is to structure and coordinate the work within the PsyPhyQA project of VQEG, other researchers in the field of psychophysiologicalvideo quality assessment are invited to use the test plan for general guidance of their studies for systematic, reproducible and comparable research. The test plan especially contributes to the quality research community by providing a video dataset that is specifically designed to study psychophysiological quality assessment methods. This dataset, comprising SRC sequences, PVS, and associated quality ratings in terms of MOS, is made publicly available for research. This will greatly improve the quantitative comparability of psychophysiological approaches to visual quality assessment and help to identify specific strengths and weaknesses thereof. However, the presented dataset can only be a starting point. The considered selection of SRCs is limited with regard to its content as it does not capture many categories, such as cartoon, screen-recording, handheld capturing, artistic, CGI or mixed reality. The same holds for higher resolutions, i.e. 4K.

In addition, multimedia signals are not limited to videos. In parallel to move along the proposed test plan, it is important to evaluate physiological quality assessment for emerging modalities such as virtual and/or augmented reality. These are clearly aspects to be considered for the design of future datasets and evaluation frameworks.

Acknowledgements The authors would like to thank Pierre Lebreton, Werner Robitza and Steve Göring for providing their software for calculating spatial and temporal information at https://github.com/Telecommunication-Telemedia-Assessment/SITI, Mikołaj Leszczuk for running the pre-tests with his students that allowed us to estimate the relavant rate points of our PVS, and Margaret Pinson for uploading our dataset. Furthermore, the authors would like to thank the anonymous reviewers who helped to significantly improve the quality of the manuscript.

Compliance with ethical standards

Conflict of interest Authors declare that they have no conflict of interest.

References

1. Acqualagna L, Bosse S, Porbadnigk AK, Curio G, Müller KR, Wiegand T, Blankertz B (2015) EEG-based classification of video quality perception using steady state visual evoked potentials (SSVEPs). J Neural Eng 12(2):026012
2. Ahn M, Jun SC (2015) Performance variation in motor imagery brain–computer interface: a brief review. J Neurosci Methods 243:103–110
3. Antons JN, Schleicher R, Arndt S, Möller S, Curio G (2012) Too tired for calling? A physiological measure of fatigue caused by bandwidth limitations. In: Proceedings of the quality of multimedia experience (QoMEX)
4. Antons JN, Schleicher R, Arndt S, Möller S, Porbadnigk A, Curio G (2012) Analyzing speech quality perception using electroencephalography. J Select Topics Signal Proc 6:721–731
5. Arndt S (2016) Neural correlates of quality during perception of audiovisual stimuli. Springer, Berlin
6. Arndt S, Antons JN, Schleicher R, Möller S (2016) Using electroencephalography to analyze sleepiness due to low-quality audiovisual stimuli. Signal Process Image Commun 42:120–129
7. Arndt S, Antons JN, Schleicher R, Möller S, Curio G (2014) Using electroencephalography to measure perceived video quality. IEEE J Sel Top Signal Process 8:366–376
8. Arndt S, Brunnström K, Cheng E, Engelke U, Möller S, Antons JN (2016) Review on using physiology in quality of experience. Electron Imaging 2016(16):1–9
9. Avarvand FS, Bosse S, Müller KR, Schäfer R, Nolte G, Wiegand T, Curio G, Samek W (2017) Objective quality assessment of stereoscopic images with vertical disparity using EEG. J Neural Eng 14(4):046009
10. Blankertz B, Sannelli C, Halder S, Hammer EM, Sanelli C, Halder S, Hammer EM, Kübler A, Müller KR, Curio G, Dickhaus T (2009) Predicting BCI performance to study BCI illiteracy. BMC Neurosci 10(1):84
11. Blankertz B, Tomioka R, Lemm S, Kawanabe M, Müller KR (2008) Optimizing spatial filters for robust EEG single-trial analysis. IEEE Signal Process Mag 25(1):41–56. https://doi.org/10.1109/MSP.2008.4408441
12. Bosse S, Acqualagna L, Porbadnigk A, Blankertz B, Curio G, Müller KR, Wiegand T (2014) Neurally informed assessment of perceived natural texture image quality. In: Proceedings of IEEE international conference on image processing (ICIP), pp 1987–1991
13. Bosse S, Acqualagna L, Porbadnigk AK, Curio G, Müller KR, Blankertz B, Wiegand T (2015) Neurophysiological assessment of perceived image quality using steady-state visual evoked potentials. In: Applications of digital image processing XXXVIII vol 9599, pp 959914–959914
14. Bosse S, Acqualagna L, Samek W, Porbadnigk AK, Curio G, Blankertz B, Müller K, Wiegand T (2018) Assessing perceived image quality using steady-state visual evoked potentials and spatio-spectral decomposition. IEEE Trans Circuits Syst Video Technol 28(8):1694–1706. https://doi.org/10.1109/TCSVT.2017.2694807
15. Bosse S, Bagdasarian MT, Samek W, Curio G, Wiegand T (2018) On the stimulation frequency in ssvep-based image quality assessment. In: 2018 Tenth international conference on quality of multimedia experience (QoMEX), pp. 1–6. https://doi.org/10.1109/QoMEX.2018.8463381
16. Bosse S, Müller KR, Wiegand T, Samek W (2016) Brain–computer interfacing for multimedia quality assessment. In: Proceedings of the IEEE international conference on systems, man, and cybernetics (SMC), pp 2834–2839

17. Bossen F (2013) Common test conditions and software reference configurations Output. In: Joint collaborative team on video coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG1, JCTVC-L110, San José CA, USA
18. Brunnström K, Barkowsky M (2018) Statistical quality of experience analysis: on planning the sample size and statistical significance testing. J Electron Imaging 27(5):11. https://doi.org/10.1117/1.JEI.27.5.053013
19. Brunnström K, Hands D, Speranza F, Webster A (2009) VQEG validation and itu standardisation of objective perceptual video quality metrics. IEEE Signal Process Mag 26(3):96–101
20. Cohen J (1988) Statistical power analysis for the behavioral sciences. Lawrence Erlbaum Associates, New Jersey
21. Coles MS, Rugg M (1995) Event-related brain potentials: an introduction. In: Coles MS, Rugg M (eds) Electrophysiology of mind: event-related brain potentials and cognition. Oxford University Press, Oxford
22. Engelke U, Darcy DP, Mulliken GH, Bosse S, Martini MG, Arndt S, Antons JN, Chan KY, Ramzan N, Brunnström K (2017) Psychophysiology-based qoe assessment: a survey. IEEE J Sel Top Signal Process 11(1):6–21
23. Gonzalez P, Althobaiti A, Katsigiannis A, Ramzan N (2017) Perceptual video quality evaluation by means of physiological signals. In: Proceedings of 9th international conference on quality of multimedia experience (QoMEX 2017), pp 1–6
24. Gupta R, Laghari K, Arndt S, Schleicher R, Moller S, O'Shaughnessy D, Falk T (2013) Using fnirs to characterize human perception of tts system quality, comprehension, and fluency: preliminary findings. In: International workshop on perceptual quality of systems (PQS)
25. Haglund L (2006) The SVT high definition multi format test set. Report, Sveriges Television AB (SVT)
26. Haufe S, Dähne S, Nikulin VV (2014) Dimensionality reduction for the analysis of brain oscillations. Neuroimage 101:583–597. https://doi.org/10.1016/j.neuroimage.2014.06.073
27. Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, Bießmann F (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87:96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067
28. Hoenig JM, Heisey DM (2001) The abuse of power: the pervasive fallacy of power calculations for data analysis. Am Stat 55(1):19–24
29. Huynh-Thu Q, Webster A, Brunnström K, Pinson M (2015) VQEG: shaping standards on video quality. In: Proceedings of 1st international conference on advanced imaging, Tokyo, Japan
30. ITU-R Rec. BT.2022: general viewing conditions for subjective assessment of quality of SDTV and HDTV television pictures on flat panel displays (2012)
31. ITU-R Rec. BT.500-13: methodology for the subjective assessment of the quality of television pictures (2012)
32. ITU-T Rec. P.910: subjective video quality assessment methods for multimedia applications (2008)
33. ITU-T Rec. P.913: methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment (2016)
34. Janowski L, Pinson M (2015) The accuracy of subjects in a quality experiment: a theoretical subject model. IEEE Trans Multimed 17(12):2210–2224
35. JCT-VC: Subversion repository for the HEVC test model reference software (2014). https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware
36. Jeunet C, Jahanpour E, Lotte F (2016) Why standard brain–computer interface (BCI) training protocols should be changed: an experimental study. J Neural Eng 13(3):36024
37. Li Z, Bampis CG (2017) Recover subjective quality scores from noisy measurements, pp 52–61. https://doi.org/10.1109/DCC.2017.26
38. Lindemann L, Magnor M (2011) Assessing the quality of compressed images using EEG. In: 2011 18th IEEE international conference on image processing (ICIP), pp 3109–3112. IEEE
39. Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A, Yger F (2018) A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. J Neural Eng 15(3):31005. https://doi.org/10.1088/1741-2552/aab2f2
40. Luck SJ (2005) Ten simple rules for designing ERP experiments. In: Handy TC (ed) Event-related potentials: a methods handbook. The MIT press, US, pp 17–32
41. Norcia AM, Appelbaum LG, Ales JM, Cottereau BR, Rossion B (2015) The steady-state visual evoked potential in vision research: a review. J Vis 15(6):1–46
42. Perkis A, Arndt S (2018) Trends in qoe for immersive experiences. IEEE MMTC Front 2018(13):27–33
43. Sannelli C, Vidaurre C, Müller KR, Blankertz B (2019) A large scale screening study with a SMR-based BCI: categorization of BCI users and differences in their SMR activity. PLoS ONE 14(1):e0207351
44. Scholler S, Bosse S, Treder MS, Blankertz B, Curio G, Müller KR, Wiegand T (2012) Toward a direct measure of video quality perception using EEG. IEEE Trans Image Process 21(5):2619–29
45. Stevens S (1946) On the theory of scales of measurement, vol 103. Bobbs-Merrill, Indianapolis
46. Sullivan GJ, Ohm JR, Han WJJ, Wiegand T (2012) Overview of the high efficiency video coding (HEVC) standard. IEEE Trans Circuits Syst Video Technol 22(12):1649–1668
47. The EU General Data Protection Regulation. www.eugdpr.org. Accessed 10 March 2019
48. Uhrig S, Arndt S, Möller S, Voigt-Antons JN (2017) Perceptual references for independent dimensions of speech quality as measured by electroencephalography. Qual User Exp 2(1):10
49. Uhrig S, Mittag G, Möller S, Voigt-Antons JN (2018) Neural correlates of speech quality dimensions analyzed using electroencephalography (eeg). J Neural Eng. http://iopscience.iop.org/10.1088/1741-2552/aaf122
50. Video Quality Experts Group (VQEG). www.vqeg.org. Accessed 25 April 2018
51. Webster AA, Jones CT, Pinson MH, Voran SD, Wolf S (1993) Objective video quality assessment system based on human perception. Human vision, visual processing, and digital display IV, vol 1913. International Society for Optics and Photonics, Bellingham, pp 15–27
52. Winkler S (2009) On the properties of subjective ratings in video quality experiments. In: 2009 international workshop on quality of multimedia experience. IEEE, pp 139–144