Mercer Moss, F., Yeh, C-T., Zhang, F., Baddeley, R., & Bull, D. R. (2016). Support for reduced presentation durations in subjective video quality assessment. *Signal Processing: Image Communication*, 48, 38–49. <https://doi.org/10.1016/j.image.2016.08.005>

Peer reviewed version

License (if available):
Unspecified

Link to published version (if available):
[10.1016/j.image.2016.08.005](https://doi.org/10.1016/j.image.2016.08.005)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via Elsevier at <http://www.sciencedirect.com/science/article/pii/S0923596516301126> . Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

Support for Reduced Presentation Durations in Subjective Video Quality Assessment

Felix Mercer Moss^{a,*}, Chun-Ting Yeh^a, Fan Zhang^a, Roland Baddeley^b,
David R. Bull^a

^a*Department of Electrical and Electronic Engineering, University of Bristol, Bristol BS8 1UB, U.K.*

^b*Department of Experimental Psychology, University of Bristol, Bristol BS8 1UB, U.K.*

Abstract

Video content distributors, codec developers and researchers in related fields often rely on subjective assessments to ensure that their video processing procedures result in satisfactory quality. The current 10s recommendation for the length of test sequences in subjective video quality assessment, however, has recently been questioned. Not only do sequences of this length depart from modern cinematic shooting styles, the use of shorter sequences would also enable substantial efficiency improvements to the data collection process. Our previous work, using a double-stimulus methodology, indicated that shortening test sequences had a limited impact upon rating behaviour. Here, using a larger database and additional opinion score measures, we also explore the same effect within the popular single-stimulus approach. Two groups of viewers assessed reference and distorted videos ranging in length from 1.5s to 10s. Analyses confirmed our previous findings using the DSCQS paradigm, and were replicated when using a similar single-stimulus paradigm: while viewers' DMOS for 1.5s videos was significantly lower than for 10s, no significant variation was found between the groups of 10s, 7s and 5s videos. Together with our previous research, these data lead us to recommend the use of 5s, temporally-consistent video clips in quality assessment studies that employ either DSCQS or its single-stimulus

*Corresponding author

Email address: f.mercermoss@bristol.ac.uk (Felix Mercer Moss)

variant. The extension of our recommendation to further methodologies is also discussed.

Keywords: subjective testing, quality assessment, test conditions, reliability, methodology, video signal processing, video coding, video sequences, video databases, visual perception, statistical reliability, double stimulus continuous quality scale, single stimulus continuous quality scale, mean opinion scores, video coding, HEVC

1. Introduction

The improved availability of internet and visual displays, especially on mobile devices, has driven demand for digital video. In a Cisco Systems Inc. forecast, the proportion of all internet protocol traffic identified as video content is predicted to rise to 80% by 2019, up from 67% in 2014 [1]. This increased demand
5 for digital video has, in turn, stimulated research in related disciplines such as video quality assessment (VQA). This field aims to assess the perceptual quality of compressed video, with the aim of maintaining acceptable standards, while achieving highly significant bandwidth savings.

10 Video quality assessment falls into two principal classes: subjective and objective. The former involves human participants providing scores of perceived visual quality after watching test sequences in a controlled environment. Unfortunately, like all human experiments, subjective VQA studies are laborious, time-consuming and expensive. Alternatively, objective VQA attempts to solve
15 the same quality assessment problem by using computational performance metrics that model the human visual system, thereby reducing the need for human involvement.

Despite their clear benefits, the current level of sophistication in these computational models is not yet sufficiently advanced to be an acceptable substitute
20 for subjective assessment [2]. Furthermore, the development and benchmarking of objective VQA solutions will always depend upon human ground truth data obtained using subjective testing procedures. Consequently, subjective VQA

remains a vital tool to video researchers and the most reliable and accurate approach to assessing video quality [3] [4] [2] [5].

25 Subjective video quality assessment studies typically follow International Telecommunication Union (ITU) recommendations [6], resulting in a high level of standardisation across practices. However, we have previously proposed [5] how one particular guideline may be due for revision: the use of 10s test sequences. Aside from practical benefits, our motivations for exploring the effects
30 of reduced-length test sequences are based upon both theoretical and empirical reasons:

1. Limits to working memory mean observers tend to converge upon a quality assessment decision significantly before 10s of viewing [7].
2. Ten-second video sequences are not representative of current cinematic
35 shooting styles, where average shot lengths are in the region of 4.5s [8].
3. Observers tend to produce the most consistent viewing patterns in the first two to three seconds of scene-viewing [9], suggesting more consistent rating behaviour exists in shorter rather than longer clips.

Our previous results indicated that, specifically when using the double-
40 stimulus continuous quality scale (DSCQS) methodology, these motivations were well-founded. The accuracy of human observers was not significantly affected by reducing the length of temporally-consistent video sequences from 10s to as low as 3s [5]. While the DSCQS methodology is well-represented in the literature, researchers are increasingly turning to single-stimulus methodologies due to the
45 associated time-savings involved. Here we explore whether the practical benefits of reduced clip durations are as relevant within a single-stimulus (SS) methodology as we discovered them to be within a double-stimulus (DS) methodology. By doubling the size of our original dataset, we confirm that our original findings generalise to both new content and different testing paradigms. In doing
50 so, we ensure the efficiency benefits of our recommendations can be relevant to more researchers who use different methodologies and ratings statistics.

In this paper, Section 2 presents an overview of the previous research that

has motivated the current work, Section 3 provides a detailed account of the two main experiments and one pilot study, Section 4 presents the results and discussion alongside statistical analyses, while Section 5 discusses the implications of the current work with suggestions for future research directions.

2. Background

In the field of VQA, the question of how long an observer needs to make a reliable quality judgement on a video sequence is highly relevant. Two intuitive claims help inform the decision: (1) that longer clips provide more information and, therefore, higher accuracy; and (2) that shorter clips bring significant efficiency gains in the data collection process. Section 2.1 and Section 2.2 unpack and examine the validity of these two competing claims.

2.1. *Do longer clips produce more reliable assessments?*

Despite a relative scarcity of research on the topic, results from a handful of studies indicate that this is probably not the case. Motivated by the idea that typical video viewing periods are significantly longer than 10 seconds, Frölich et al. [10] explored the effects of extending the length of test sequences to the range of minutes, as opposed to seconds. However, contrary to their expectations, results suggested viewers were significantly more generous with their rating behaviour when viewing the group of longer clips (60s, 120s and 240s) than the group of shorter clips (10s, 15s and 30s), while no significant differences were found within each of the two groups. The results of this study directly challenge the claim cited earlier that longer test sequences lead to more accurate rating behaviour. On the contrary, the authors argued that longer sequences generated greater levels of immersion in viewers, which consequently reduced their sensitivity to distortions and artefacts. The result also raises an interesting point for researchers wanting to create natural viewing conditions in their subjective VQA experiments. While it is important for data to be collected in environments that do not substantially depart from those where

viewers typically consume video, care must be taken to ensure participants stay fully focussed upon the presentation quality of the video. Realism should be encouraged in VQA studies, but researchers should also be aware of the cost it may have upon observer criticality.

85 An alternative approach to increasing realism in VQA studies is to replicate, not the overall viewing time but instead, the cut-to-cut shot lengths observed in modern cinema. Shots are a fundamental unit in cinematic story-telling while they also represent important events to be considered during efficient video coding.¹ Analysis of their structure may therefore be useful when creating test
90 material. Average shot length (ASL) has dropped from almost 35 seconds in 1905 to between 4 and 5 seconds in the last decade [8]. The vast majority of modern movies (those released between 2011 and 2015) are dominated by shots that are around 5 seconds long [8] (for further analysis also see [11]). So to accurately reproduce the experience of audiences visiting modern cinemas,
95 shorter rather than longer sequences are to be encouraged.

Well-known limitations to human working memory [12] and attention [9, 13] represent further concerns with the idea that longer durations produce more accurate video quality assessments. Recency effects are those characterised by the tendency to recall items better, when placed at the end of a serial list.
100 This time-varying performance of working memory has been shown to have an impact in subjective VQA studies: viewers afford significantly more weight to a 10s sequence of impairments occurring at the end of a 30s sequence. Identical impairments shown at the beginning or middle of the 30s sequence typically produced scores 10% lower than when they were at the end [14].

105 The same effect has also been observed in shorter, 10s sequences. Research using a time-varying continuous rating scale has indicated that observers tend to form consistent quality judgements significantly before the end of a 10s sequence. By six seconds in, scores produced a 0.9 correlation with final scores recorded

¹The placement of I-frames at the start of shots is necessary to prevent significant visual artefacts.

at 10 seconds [15].

110 Evidence from eye movement studies indicate that attention between individuals diverges quickly after 2 to 3 seconds of static scene viewing [9, 13]. If consistent rating behaviour is dependent upon similar viewing behaviour, these findings suggest test sequences should not be significantly longer than five seconds.

115 Our previous work [5], exploring the impact of sub-10s sequences in the DSCQS paradigm, found that longer clips did tend to produce slightly higher levels of accuracy (higher differential mean opinion scores (DMOS)). However, compared to ratings from the 10s clips, it was only the most severe truncation (1.5s) that registered a significant negative impact upon scores.

120 2.2. Practical considerations of shorter clips

Data collection is expensive and the practical benefits to be gained from reducing the standard recommendation for test sequence duration are substantial. Theoretically, halving sequence durations would cut the entire experiment time in half or, equivalently, double the volume of collected data. However, 125 in practice, events such as video preparation screens, voting, and instructions are unaffected by sequence truncation. Therefore, the savings associated with halving test presentation times are closer to a third than a half.

However, such gains do not uniformly apply across all testing methodologies. Many different methodologies exist (for an almost exhaustive list, the reader is referred to ITU-R BT.500 [6], ITU-T P.910 [16], and ITU-R BT.1788 [17]) but 130 they can be principally divided into either single-stimulus or double-stimulus variants. In the SS variant, the experimenter asks participants to watch a single video sequence and then to make a response before continuing with the next trial. Alternatively, for the latter DS variant, the same video is displayed twice, at different levels of impairment before the observer makes a response. Due to 135 their presenting more content, DS methodologies benefit more from truncation than their SS counterparts. More specifically, replacing the 10s ITU standard with 5s would bring a 32% timesaving (or 48% volume increase) in DS studies

and a 28% timesaving (or 39% data volume increase) in SS studies².

140 Despite DS approaches being less sensitive to contextual effects (the unwanted influence of previously seen video upon ratings), SS approaches are often a more popular choice with researchers due to their increased time-efficiency [18]. While the theoretical efficiency gains of presentation reduction in SS methodologies are more modest than their DS counterparts, the savings that could
145 be made would still be very valuable. Given that previous research using the DSCQS methodology indicates rating behaviour is barely affected [5] by sequence truncation, what reasons are there to believe that this might not be the case with the SS variant? One reason is that, due to the lack of an immediate reference sequence, SS trials are intrinsically more difficult than DS trials. If, as
150 has been suggested [5], longer sequences improve observer accuracy only when the task is sufficiently difficult (or when information is sufficiently scarce), we may expect to see a stronger duration effect in SS rather than DS methodologies. Alternatively, we might expect SS methodologies to be relatively less impacted by a change in clip duration. Due to the reduced time between the start of
155 a trial and the start of voting, the concerns about the limitations of working memory capacity discussed in the previous section are less pertinent within a SS context.

In addition to different presentation styles, the type of rating scale also varies between different testing methodologies, and this too can affect the data
160 collected. Continuous quality scale designs (such as those in DSCQS) offer a finer resolution of opinion scores but, due to the influence of spatial biases, category rating methods (such as DSIS and ACR) have been found to produce more stable results [19]. Due to their sensitivity to small changes, marginal effects (such as those caused by duration [5]) may be more easy to detect using
165 continuous scales rather than more granular categorised scales.

²The assumptions used for these estimates are that the countdown time (CD) is 3s, and the voting time (VT) is 5s. The time-per-trial formulae for DS and SS (with presentation time denoted as PT) are $CD+PT+CD+PT+VT$ and $CD+PT+VT$, respectively.

3. Methodology

This section presents detailed specifications of the two main experiments reported in this paper. Both experiments were designed to gather data revealing observers’ rating behaviour when viewing different sub-10s sequence durations. For the double stimulus experiment, the DSCQS as described in [6] was used with a small number of minor alterations. Alterations from the documented DSCQS methodology include presenting each test sequences only once and asking an additional binary question with regard to which video is of a superior quality. The single-stimulus experiment used a SS variant of DSCQS with hidden reference removal. In addition to the two principal experiments, one shorter pilot study was conducted that used the same SS methodology described above.

3.1. Participants

A total of 48 participants were recruited for two experiments and were financially compensated for their time. Half of them participated in the SS experiment while the other half participated in the DS experiment. The male-to-female ratio for the DS study was 1.4:1 whereas this was 1:1.4 for the SS study.

Each participant took part in no more than one experiment. The average age was 23.6, with a minimum of 19 and maximum of 46. Snellen and Ishihara charts were used to confirm that all participants had normal visual acuity and colour vision, respectively.

3.2. Reference Sequences

The VQEG-HD [20] database and original videos from the Bristol Texture Database [21] were combined to form a pool of 113 high-definition (1920×1080), progressive scan reference sequence candidates. Each candidate had at least 10 seconds without a shot transition, contained only natural images (i.e. non-animated) and had no audio components. After transforming each candidate into the YUV 4:2:0 format and truncating to 10s, four low-level features were computed in order to quantitatively characterise each candidate.

Figure 1: Mean feature plots for all 10s video candidates with final selected references highlighted.

195 The mean spatial information (SI) estimates the frame edge density [18]. The mean temporal information (TI) is a commonly used measure of between-frame luminance change [22]. The texture parameter (TP) is a static texture properties descriptor and the dynamic texture parameter (DTP) describes dynamic texture properties [23]. These values for each reference candidate are plotted in Figure 200 1; the exact descriptions can be found in the Appendix of [5]. Due to the original DTP being highly concentrated at low values, a natural log scale was used.

Six videos were chosen from the larger set of candidates. The scene selection strategy proceeded in two steps. First, a short list of candidates was selected based upon maximising the total range along all four feature dimensions. This short list was then refined to six sequences by eliminating videos 205 featuring unusual camera dynamics and prioritising those that were the most temporally-consistent. The feature profiles of the final database of six videos are displayed in Figures 1 and 3. Example frames from each sequence can be seen in Figure 2, their dynamics are described in Table 1 and qualitative summaries 210 can be found below.

- **Divers** Two divers overturn an underwater motorbike.
- **Lobsters** Fishermen converse behind a table of live lobsters.
- **Toys** An Assortment of colourful toys and static feathers rotating on a

platform.

- 215 • **Tulips** A view of tulips and trees amongst street walkers and buskers.
- **Hamster** A hamster running in a wheel beside two toy blocks, surrounded by floating soap bubbles.
- **Splash** Flowing water splashed repeatedly.

Table 1: Description of types of movement featured in the four reference sequences.

Sequence	FPS	Camera	Structured Movement	Dynamic texture
Divers	30	Pan + tilt	Diver	Sea vegetation
Lobsters	30	Pan + tilt	People	NA
Toys	25	Static	Rotating toys	Feathers
Tulips	25	Static	People	Tree leaves
Hamster	60	Static	Bubbles	Spinning wheel
Splash	60	Static	NA	Rippling water

Of the six reference sequences, four came from the VQEG-HD database [20]
 220 (Divers, Lobsters, Toys, Tulips) and the other two from the Bristol Texture Database [21] (Hamster and Splash). All six references were used in the SS experiment but, due to time constraints, the Lobsters and Hamster videos were not used in the DS experiment.

3.3. Test Sequences

225 To produce the test sequences, each reference sequence was distorted at six different levels using High Efficiency Video Coding (HM 16.4) compression, with quantisation parameters (QP) of 22, 27, 32, 37, 42, and 47, and random access mode.

The total of 42 videos (including references) were then truncated to 1.5, 3, 5,
 230 and 7 seconds, with special care taken in the event that the video had any obvious temporal shift in content that might affect opinion scores (see Pilot reported in Section 3.4). Three such events were identified: one in Tulips; one in Toys;

Figure 2: A single frame from each of the six reference sequences.

and one in Divers. For these three cases, the truncations were made symmetrically around the anomalous event. Otherwise, videos were cut symmetrically
 235 around the 5s mark (see Figure 3). The final video database consisted of 30 reference sequences and 180 test sequences. The SS experiment employed the complete dataset whereas the DS experiment utilised just 20 reference sequences and 120 test sequences.

3.4. Pilot

240 A pilot study was conducted to check the temporal consistency of quality ratings across the six reference sequences. Each 10s reference was used to create eight smaller truncated versions: five 1.5s clips, the start of each offset by 2s; and three 3s clips, the start of each offset by 3s. Each of these clips were then encoded at QP27 and QP42. Ten participants viewed and rated each of these
 245 sequences within a SS methodology with hidden reference removal.

Figure 3: Time-varying feature values of the reference sequences. The grey horizontal bars depict which portions of the full 10s original video were truncated into shorter sequences.

Results from the 1.5s clips can be seen in Figure 4. One-way ANOVAs on the DMOS indicated no significant variance between the different offsets for the QP27 clips; however, for the QP42 clips significant effects were found for the Divers sequence, $F(4,36)=2.60$, $p=.05$, and the Tulips sequence, $F(4,36)=3.70$, $p=.01$. For the 3s clips, analysis of the QP27 ratings also yielded no significant effects while the QP42 ratings yielded one marginally significant effect associated with the Toys sequence, $F(2,18)=3.85$, $p=.04$.

These results present no evidence that user ratings are temporally inconsistent when these six sequences are encoded at a higher bitrate. When encoded at a lower bitrate, however, there is evidence that events in three of the sequences may affect user ratings. In the Tulips sequence, the 2s offset DMOS is significantly higher than the 4s offset DMOS. The reason for this is likely to be that the earlier clip contains a man walking towards and past the camera, while the later one is relatively still. Therefore, motion artefacts may be more visible at the 2s offset. In the Divers sequence, DMOS dips during the 4s offset clip. This clip ends with diver's flashlight facing the camera which may have distracted and reduced the criticality of observers. For the 'Toys' sequence, the 0s offset

clip produced significantly lower DMOS than the 3s offset. The prominence of
a high contrast structure in the earlier clip may have been attracting attention
265 and masking artefacts in the rest of the scene although this was not the case for
the 1.5s clips.

Figure 4: Results of pilot study to check the temporal consistency of the six original 10s reference sequences. Each sequence was split into five 1.5s clips, the beginning of each offset by 2s. Only the Divers and Tulips sequences showed significant differences in ratings across the different offsets. Error bars represent the standard error of the mean.

3.5. Environmental Setup

Both experiments were conducted in a darkened, living room-style environ-
ment [6]. Three observers participated in each session, which lasted no more
270 than one hour, including time for breaks. Participants sat in a position such
that the horizontal distance from their eyes to the display measured 1110mm
- three times the screen height (consistent with [5]). The display used was a
Panasonic BT-4LH310 LCD high definition monitor with 60Hz refresh rate and
1500:1 contrast ratio. It measures 700×370mm and supports viewing angles of
275 up to 178°. All videos were played at their native frame rate (see Table 1), con-

trolled by a connected PC running Windows Matlab R2012a with Psychtoolbox 3.0. A Tobii X300 remote eyetracker was located directly below the monitor and used to collect gaze data. The participant that sat centrally, placed their head on a chin rest directly in front of them. All participants provided their responses using a 9.7" iPad tablet computer running *SubTest*, an iOS app developed by the authors for the purposes of collecting subjective VQA data. The graphical user interface of *SubTest* contains either one (for SS) or two for (DS) vertical visual analogue scales, labelled at five equally-spaced intervals reading ‘Excellent’, ‘Good’, ‘Fair’, ‘Poor’, and ‘Bad’. Participants used an adjacent finger slider to register their response along a visual analogue scale. In the DS experiment, *SubTest* displayed an additional question with regard to which video was of a higher quality, to which participants answered using a binary tabbed button.

3.6. Procedure

Each trial in the DS experiment started with a 3000ms grey screen count-down, showing a central crosshair that changed colour each second from red to yellow to green. Video A was then played, followed by another identical count-down screen before the commencement of Video B. When Video B was complete, a grey voting screen was shown to indicate that observers should record their responses to two requests. The first of these was phrased “Which video did you perceive as better quality?”. Participants registered their answers by tapping either ‘Video A’ or ‘Video B’. The second request was phrased “Please rate the perceived quality of the two videos.” Participants could use the sliders next to the visual analogue scales to record their response. The next trial did not begin until all present participants indicated they had finished voting. As a means of quality control, the *SubTest* app checks for inconsistencies between requests 1 and 2. For example, having selected Video A as superior to Video B in request (1), the app will only allow the user to progress to the next trial if slider A has been registered as higher than slider B.

Each trial in the SS experiment began with a countdown screen of 3000ms identical to the one described above and played a designated video. Once the

video was complete, viewers were asked: “Please rate the perceived quality of the video.” Participants then recorded their response using the touch-based visual analogue scale on the tablet computer.

The videos used for each experiment were grouped by duration to form five
310 blocks. The block order and the order within each block was randomised. The presentation order of the test and reference sequences in the DS experiment was also randomly assigned. To minimise ordering effects, care was taken to ensure that all such randomisations were counter-balanced and the frequency of a video being at a certain position approximated a uniform distribution.

315 For both experiments, a session began with participants being delivered clear instructions before commencing three practice trials in the respective formats. For each session, one participant had their eyes tracked. Before each block this participant engaged in an eye calibration process using Tobii software. At the end of each block, every participant was asked: “How confident do you feel about
320 your ratings for the videos you have just watched in the previous block? Draw on the scale from Very Unconfident to Very Confident”. The visual analogue scale participants used to respond was labeled at each end with “Very Unconfident” and “Very Confident”. Participants were given time to rest between blocks.

3.7. Analysis

325 All responses from visual analogue scales were recorded as points on a 0-100 scale. DMOS, raw mean opinion scores (MOS) and error percentages were used to measure rating accuracy. DMOS in the DS dataset was calculated as the absolute score difference between the reference and test video in each trial. DMOS in the SS was calculated as the absolute difference between each test sequence
330 score and that of the corresponding reference sequence (as in hidden reference removal). In each dataset, each observer produced as many DMOS values as there were test sequences. Raw MOS scores are reported for both SS and DS experiments but statistical analyses were limited to data from the SS paradigm as MOS is not suitable for DS methodologies. An additional performance metric
335 was calculated for each methodology by computing participant errors. Errors

were events where an observer had rated a distorted sequence higher than its reference counterpart.

Subjects were then screened for outliers according to ITU protocol [6]. A Kurtosis value was computed for DMOS in each possible test condition, across
340 participants. If a Kurtosis value was within the range of 2 and 4, then the score distribution in that particular test condition was considered normal. For normally-distributed conditions, outliers were defined as a score outside two standard deviations of the mean. For those not normally-distributed, scores were classified as outliers if they were outside $\sqrt{20}$ times the standard deviation
345 of the mean. Any participant whose data satisfied two rejection criteria in conjunction was excluded: first, more than 5% of DMOS were outliers; and second, the absolute difference in counts of their high and low outliers were below 30% of the sum of their outlier count.

In the DS dataset, one participant was rejected for 9.2% outliers and 9.1%
350 outlier difference ratio, while no participants were classified as outliers from the SS dataset.

The gaze data was analysed after dividing it into 500ms time periods. Two-dimensional probability distribution functions (PDFs) were then calculated for each time period. PDFs were calculated by counting the number of gaze loca-
355 tions landing upon each pixel on the screen before convolving with a Gaussian kernel (standard deviation of 100 pixels) and normalising so that it sums to one. The spread of each of these distributions was measured by calculating the entropy of each. Gaze data recorded up to first 500ms after stimulus onset was not used as inspection indicated many participants were not fixated centrally
360 on the cross.

4. Results and Discussion

The main statistical analyses used in this section were repeated-measures analysis of variance (ANOVA), pairwise comparisons adjusted for multiple comparisons using Tukey’s Least Significant Difference and Pearson’s correlation

coefficient [24]. The one-way ANOVA reveals whether duration significantly affects opinion scores in Section 4.1. Two-way ANOVAs were employed in Section 4.2 and 4.3 respectively to show whether variation in compression level (QP) or video content made significant contributions to the main duration effect. When degrees of freedom are reported as a decimal, they have been adjusted due to a violation of the sphericity assumption.

In the case of an ANOVA being significant, pairwise comparisons were examined to identify which duration groups were significantly different. Correlational analysis determined whether the opinion scores held a linear relationship with duration in any given test condition. For further details on the guidelines for statistical methods used in subjective VQA studies, the reader is referred to the ITU recommendation, BT.500 [6] and [24].

4.1. Global Analysis

By collapsing over all participants and test conditions, the global DMOS averages were 33.9 for the DS experiment and 23.9 for the SS experiment. A matched samples t-test indicated this difference was highly significant, $t(91)=10.54$, $p<.001$. This discrepancy demonstrates how the presence of a reference sequence provides an advantage when identifying compression artefacts. Global single-stimulus MOS (henceforth referred to as SS-MOS) was 55.3.

The six plots in Figure 5 present the global results for the two experiments. DMOS from the SS and DS (displayed in Figures 5a and 5b, and henceforth referred to as SS-DMOS and DS-DMOS) experiments yielded similar increasing trends from 1.5s to 7s, although the increase was steeper in the latter than the former. DMOS continued to rise in the 10s group for the SS data, yielding a significant linear correlation between duration and SS-DMOS ($r(22)=0.24$, $p=.009$) but declined in the DS experiment (no significant correlation, $r(21)=0.06$, $p=.51$). Furthermore, no significant variation was found between the groups of DS-DMOS, $F(2.39, 52.52)=.58$, $p=.60$ while significant variation was identified in the SS-DMOS, $F(4, 92)=3.91$, $p=.006$. Pairwise comparisons indicated that SS-DMOS was significantly lower than 10s only in the

Figure 5: Viewer ratings for each of the five duration groups, averaged over all content and all distortions. The plots in the first column (a,c,e) show data from the single stimulus experiment, while plots in the second column (b,d,f) show data from the double stimulus experiment. Plots in the top row (a,b) show DMOS data, plots in the middle row (c,d) show DMOS error data and plots in the bottom row show MOS data. Error bars represent the standard error of the mean.

395 1.5s block ($p=.009$) and 3s block ($p=.05$). No significant differences were found between the 10s block and the blocks of 7s ($p=.52$) or 5s ($p=.35$).

The corresponding error plots in Figure 5c (for SS) and Figure 5d (for DS), produced a similar, albeit more noisy, story. The pattern of errors in the DS
 400 errors occurred in the 10s group. The ANOVA confirmed there to be no significant differences between the groups $F(4, 88)=2.36, p=.60$. While the pattern of errors in the SS data appears to follow a more predictable trend with most

occurring in the 1.5s group and least in the 7s group, none of these differences were significant, $F(4, 92)=1.36$, $p=.25$.

405 The MOS data for the SS and DS experiments are displayed in Figure 5e and 5f, respectively. While analysing this data it is important to note that MOS is an absolute score and not a relative one. For this reason, and unlike DMOS, a higher or lower MOS has no relation to accuracy of compression detection. Nonetheless, the pattern emerging from the MOS data is consistent
410 with what is seen in the DMOS: that ratings are consistent between the groups of clips containing sequences over 1.5s. However, significant differences were found between the different duration groups, $F(4, 92)=4.15$, $p=.004$. Pairwise comparisons indicated that a large proportion of this variation may be explained by appealing to the fact that MOS values in the 1.5s group were significantly
415 lower than all other truncation groups (all $p<.0.02$ with the exception of 1.5s vs 10s, $p=.05$). The only other marginally significant difference was the increase in MOS from 10s to 7s groups ($p=.05$). Significantly lower MOS recorded for 1.5s clips compared to each of their longer duration counterparts suggests an interesting and counter-intuitive effect: when viewing very short duration clips,
420 observers become more critical of the quality. Such an insight may become increasingly relevant as the cinematic trend for average shot length is in sharp decline [11].

This set of results supports the notion that a reduction in the length of test sequences has a weak but significant effect upon the performance of observers
425 tasked with identifying compression artefacts. Despite the ANOVA yielding no significant effects, the striking resemblance of the pattern of DS-DMOS data to the results presented in our previous paper using the same methodology [5] provides strong support for our original conclusions, including the recommendation of 5s test sequences in DSCQS studies.

430 Despite the SS-DMOS data and DS-DMOS data being highly correlated, $r(98)=0.930$, $p=.02$, there were meaningful differences. Single-stimulus DMOS was significantly correlated with clip duration and the same metric was found to be significantly diminished when 10s clips were exchanged for 3s or 1.5s clips.

For the DS-DMOS dataset, however, these effects were much weaker and not
435 significant. While DS-DMOS peaked before 10s, SS-DMOS continued to rise up
to 10s. These data suggest the existence of a slightly stronger duration effect in
the SS paradigm than the DS methodology. Why might this be? It is sensible
to believe that the paradigm with the more difficult task would be the one that
benefits more from longer clips. The case can be made either way: that the
440 increased strain upon working memory load in DS methodologies or the lack of
reference in SS methodologies make the respective task more difficult. In this
case, the results suggest that the lack of reference in the SS methodology had a
greater impact upon difficulty than the additional working memory load in the
DS methodology, contributing to a stronger duration effect in the SS dataset.

445 It is, however, of note that duration did not significantly affect the number
of errors recorded by participants in either experiment. Duration therefore, ap-
pears to have more of an effect upon observers' ability to identify the magnitude
of a distortion than their ability to detect the existence of one.

4.2. *Impact of Compression Level*

450 Higher levels of compression produce videos with greater distortion. The
six plots in Figure 6 demonstrate how observers are sensitive to these varying
levels of distortion, but not uniformly so. In both experiments, the differences in
DMOS (Figure 6a and 6b) between the three highest levels of distortion (QP37,
QP42 and QP47) were more than those from the three lowest levels of distortion
455 (QP22, QP27 and QP32). Interestingly, this was the case to a lesser extent for
MOS (Figure 6e and 6f) and not the case for the error plots (Figure 6c and 6d):
an indication that at the lower end of the QP scale, perception of distortion
magnitude but not detection is compromised.

As expected, the two-way ANOVA (factor 1: duration, factor 2: QP) con-
460 firmed the significant separation of the QPs in the SS-DMOS, $F(1.23, 28.33)=256.66$,
 $p<.001$, the SS-MOS, $F(1.43, 32.93)=432.05$, $p<.001$, and the DS-DMOS, $F(1.51,$
 $33.10) = 442.10$, $p<.001$. Furthermore, the interaction effect for SS-DMOS,
 $F(6.88, 158.21)=2.64$, $p=.01$, and SS-MOS, $F(7.44, 171.02)=2.89$, $p=.006$ were

both significant: an indication that the duration effect identified in the previous
465 section varied significantly as a function of compression strength. Table 2 and
Table 3 unpack these findings by listing the details of ANOVA and correlation
statistics for individual QP values for SS-DMOS and SS-MOS, respectively. In-
triguingly, the QPs that produced the largest duration effects were not the same
in the DMOS and MOS data. Table 2 highlights how, for DMOS, the two high-
470 est levels of distortion (QP42 and QP47) produced significant duration effects
as well as significantly correlating with the sequence duration. The remaining
QPs produced no significant effects with the exception of a marginally signifi-
cant ANOVA for QP27. Conversely, Table 3 presents data indicating that for
the MOS data, 4 QPs (22, 27, 32 and 42) produced significant effects.

Table 2: SS-DMOS correlational statistics and ANOVA separated by QP. Bold with asterisk
denotes $p < .05$.

QP	r	p_r	$F\text{-test}$	p_f
22	.13	.15	$F(4, 92) = 0.14$	.35
27	.04	.67	$F(4, 92) = 2.56$	.04*
32	.03	.71	$F(3.71, 85.33) = .043$	.99
37	.18	.05	$F(2.87, 66.03) = 1.74$	.17
42	.26	.005*	$F(4, 92) = 3.31$	.01*
47	.22	.015*	$F(2.66, 61.14) = 6.62$	.001*

475 Our previous work [5] using the DSCQS methodology identified significant
duration effects only in midrange QP values (QP32 and QP37) leading us to
suggest this is where the greatest perceptual gains can be made by lengthening
test sequences. However, the current DS set of data is unable to support this
claim as no duration effects were found when using the DSCQS methodology
480 (including no significant interaction effect between duration and QP in a two-
way ANOVA, $F(6.85, 150.78)=1.00, p=.43$). One possible reason for the reduced
magnitude of the duration effects in the current set of DS-DMOS results is that
the current content produced artefacts that were easier to detect than those

Table 3: SS-MOS correlational statistics and ANOVA separated by QP. Bold with asterisk denotes $p < .05$.

QP	r	p_r	$F\text{-test}$	p_f
22	.20	.033*	F(4, 92) = 7.00	.001*
27	.13	.16	F(2.92, 67.12) = 4.15	.01*
32	.19	.04*	F(2.92, 67.12) = 5.40	.002*
37	.08	.41	F(4, 92) = 0.58	.68
42	-.09	.32	F(4, 92) = 2.99	.02*
47	-.13	.15	F(2.81, 64.73) = 2.19	.10

produced by the previous encoded content. Supporting this theory, average DS-
485 DMOS in the current study is almost 10 points more than the same statistic
in our previous study, despite the original experiment using sequences with, on
average, a higher level of distortion.

Furthermore, the current SS-DMOS data produced the strongest duration
effects not in the midrange, but in the most highly compressed sequences (or
490 easiest trials). It may be the case that the benefit of longer durations, currently
seen in highly distorted SS sequences, is replaced by the benefit of having a
direct reference sequence in the DS methodology. In this situation, it would be
expected that the duration effect would be evident at higher QP values in the
SS paradigm and lower QP values in the DS paradigm.

495 As MOS is an absolute rather than relative measure, the QP separated MOS
plots communicate different information, and hence, a slightly different story.
While only four levels of distortion produced significant test statistics (22, 27, 32
and 42), closer inspection of Figure 6e reveals that the pattern of MOS scores
reverses as the level of distortion increases, an observation supported by the
500 correlation statistics in Table 3. This interaction hints at a simple effect: that
for highly distorted content, observers are more critical of longer sequences,
while for high quality content, observers are more critical of shorter sequences.

The results of this section indicate that QP has a small but significant influ-

ence upon the duration effect in the SS paradigm, but that effect varies dependent upon whether DMOS or MOS is analysed. The lack of a significant effect in the DS dataset does not invalidate our previous claims about the impact of the duration effect being dependent upon the difficulty of the task but it provides no added support for the hypothesis, while reinforcing the overall claim that duration has a very weak effect upon rating behaviour.

4.3. Impact of Video Content

It is clear from Figure 7 and previous studies [5] that different reference sequences represent more or less difficult challenges for critical observers in a VQA context. Two-way ANOVAs (factor 1: duration, factor 2: reference sequence) confirmed that DS-DMOS, $F(3, 66)=87.61, p<.001$, SS-DMOS, $F(2.95, 67.80)=29.69, p<.001$ and SS-MOS $F(5,115)=37.18, p<.001$. varied significantly between reference sequences.

Interestingly, the interaction effects emerged significant for the DS-DMOS, $F(5.95, 130.90)=3.13, p = .007$ and the SS-MOS, $F(8.10, 186.30)=2.79, p=.006$, but not the SS-DMOS $F(9.22, 211.98)=1.51, p=.15$. This indicates that the pattern of DMOS was relatively consistent between sequences in the SS experiment but inconsistent in the DS experiment. Figures 7b and 7d clearly illustrate how the Divers clip in the DS video database distinguishes itself in two ways: first, it produced higher DMOS and fewer errors than the other sequences: an indication that it contained more visible artefacts; and second, the 5s, 7s and 10s groups have higher DS-DMOS and fewer errors than the 1.5s and 3s groups. Inspection of the pairwise comparisons confirmed that the inconsistency in the DS dataset was largely due to a distinct and significant pattern of variation in the Divers sequence. More specifically, the pairwise comparisons displayed in Table 4 show how a decrease in DS-DMOS from the 10s group was significant only in the 3s and 1.5s groups.

While no significant interaction effects were found in SS-DMOS, nevertheless, two sequences stood out as having stronger duration effects than the rest: Divers and Hamster. Independent analysis of the Divers SS-DMOS produced

both a significant ANOVA, $F(4, 92)=3.65$, $p=.008$ and a significant correlation
 535 with duration $r(22) = .22$, $p=.016$. Similarly, independent analysis of the Ham-
 ster SS-DMOS also produced a significant ANOVA, $F(4, 92)=4.13$, $p=.004$, and
 a significant correlation with duration, $r(24)=.23$, $p=.012$. The significant pair-
 wise comparisons of these data can be seen in Table 4. Further independent
 analysis of the SS-DMOS sequences yielded no more significant effects.

540 The significant interaction of duration and sequence in the SS-MOS is neatly
 illustrated in Figure 4e. There is general agreement between all six sequences
 with regard to SS-MOS increasing from 1.5s to 3s, with this also reflected in
 Table 4. However, the two most critically-rated sequences (Divers and Splash)
 increased from 3s to 10s while the remaining sequences displayed a decrease
 545 in SS-MOS during the same period. The greater number of SS-MOS entries in
 Table 4 may be an indication that duration has a greater impact upon MOS than
 DMOS. However, the strong interaction effect suggests it is difficult to predict
 whether the truncation of a sequence will produce an increase or decrease in an
 observers evaluation.

Table 4: Pairs of clip durations with significant differences in opinion scores.

Dataset	Sequence	Significant Durations Pairs
DS-DMOS	Divers	1.5s vs. 7s, 1.5s vs. 10s, 3s vs. 7s, 3s vs. 10s
SS-DMOS	Divers	1.5s vs. 7s, 1.5s vs. 10s, 3s vs. 7s
SS-DMOS	Hamster	1.5s vs. 3s, 1.5s vs. 5s, 1.5s vs. 7s, 1.5s vs. 10s
SS-MOS	Tulips	1.5s vs 7s, 7s vs 10s,
SS-MOS	Divers	1.5s vs 3s, 1.5 vs 5s, 1.5s vs 7s, 1.5s vs 10s
SS-MOS	Toys	3s vs 10s
SS-MOS	Splash	1.5s vs 10s, 1.5s vs 7s
SS-MOS	Lobsters	1.5s vs 3s, 1.5s vs 5s, 1.5s vs 7s 10s vs 3s, 10s vs 5s, 10s vs 7s
SS-MOS	Hamsters	1.5s vs 3s, 1.5s vs 5s, 1.5s vs 7s, 1.5s vs 10s

550 So why did we see the strongest duration effect in the Divers sequence, both

in the SS and DS experiments? The 10s Divers sequence is a dark underwater scene, punctuated by a single event around five seconds in, where a torch is shined towards the camera. This event is quantitatively characterised in the time-varying feature plots as a peak in dynamic texture (the DTP feature) while
555 it was also shown to negatively affect DMOS in the pilot study (see Figure 4). The 3s and especially the 1.5s clips are truncated tightly around this event meaning these sequences are likely to suffer a content-based decrease in DMOS, independent of the impact of duration. The three longer clips all contain more than a second of footage after the event in question, therefore are less likely to
560 be affected. As previously stated, the event may have distracted users from the task leading them to be less critical, but the corresponding spike in dynamic texture may also provide a clue. Theoretically, an increase in dynamic texture makes artefact detection more difficult as the chaotic nature of the original content should make distortions harder to identify. It is likely that the relative
565 increase in dynamic texture in the shorter Diver clips produced a masking effect leading to significantly lower DMOS in both experiments.

Despite these speculations, the majority of reference sequences for the DS experiment produced no duration effect while the one exception is not strikingly dissimilar. Furthermore, SS-DMOS did not significantly interact with duration
570 while the effects for SS-MOS between 3s and 10s were minimal. Overall, while it appears that singular, time-specific events may affect rating behaviour, there appears to be very little evidence that a duration effect is greatly affected by different sequence content in either the SS or DS procedures.

4.4. Participant Assessment Confidence

575 The two plots in Figure 8 reveal how confident observers felt after rating videos in each of the five different duration blocks. The first point of note is that both plots for SS and DS experiments bear a striking resemblance to their DMOS counterparts in Figure of 5a and Figure 5b. Confidence rose from the 1.5s group to the 5s group in the DS experiment while confidence peaked at 7s
580 for the SS experiment before levelling off. The differences between the groups

were significant both in the DS, $F(1.86, 40.84)=13.17$, $p<.001$, and SS, $F(2.49, 57.23)=10.33$, $p<.001$ procedures. Pairwise comparisons indicated that assessment confidence dropped significantly from the 10s level, only while viewing 1.5s videos (DS, $p=.001$; SS, $p=.006$).

585 These data indicate that shorter sequences do reduce the assessment confidence of observers; however, this only has a significant impact when clips are below 3s.

4.5. Fixation Distribution Analysis

Figure 9a illustrates how entropy (the spread of gaze data) increased as a function of viewing time. The increase was not a linear one, with the steepest
590 incline appearing from 1-3s of viewing, after which, values began to level off. Interestingly, Figure 9b shows how this pattern of entropy increase over time, was not the same when viewing different length clips. This indicates the fixation strategies, employed by observers in each of the five duration blocks, were not the same. For example, when comparing data from the 5s and 10s group, observers' gaze appears more consistent (lower entropy) while watching the shorter rather than the longer clips. One interpretation that is consistent with this pattern of data is that when observers are aware they have less time to assess a video, they make more fixations to a smaller number of predictable, information-rich
600 areas. Conversely, observers viewing longer clips know they have more time to make more exploratory fixations to less predictable locations. However, since the content of different clip lengths is not identical, it is difficult to surmise this effect is indeed due to presentation time and not variation in the visual stimuli.

One of the motivations for this paper was that, during static scene-viewing, consistency in observer viewing behaviour reduces as viewing time increases.
605 We suggested this was also likely to be the case while watching (temporally-consistent) video sequences. The data presented in this section represents the first evidence that this is indeed the case.

5. Conclusions

610 Our previous research indicated that reducing presentation time in DSCQS experiments from the standard 10s to 5s does not significantly affect observer rating behaviour. Here, using a new and larger video database, we have replicated and extended these findings, using not only the DS paradigm, but also the equivalent SS paradigm.

615 The results of our current DS experiment present an even stronger case than our previous work that, for DSCQS studies, there is little to no significant benefit in keeping sequence lengths as long as ten seconds. Truncating temporally-consistent video sequences produced significant effects neither globally nor at any specific compression level. Independent analyses indicated that observer 620 criticality was not significantly affected by truncation in three of four sequences, while the fourth produced no evidence of an effect on truncations as low as five seconds. These results reinforce our previous claims that significant efficiency gains could and should be made in VQA studies using the DSCQS methodology by using shorter video sequences.

625 We found stronger duration effects in our SS experiment but, for DMOS at least, these were only significant for the two most aggressive truncations of 1.5 seconds and 3 seconds. Critically, when compared to the 10s group, the 7s and 5s truncations produced no significant change in the accuracy of observers. Both the level of distortion and the content of the sequence were found to have 630 a small but significant influence upon the duration effect, but the weakness of these effects indicate that they should be of little practical concern to VQA researchers. The MOS data produced more variation than the DMOS, but not enough to weaken our principle recommendation. When participants were asked about assessment confidence, their responses mirrored the trial data: observers 635 only felt significantly less confident in their assessments when presentation time was reduced to 1.5s. We also present eye-tracking data that help contribute to an explanation for some of these results: gaze patterns between different observers were more consistent while viewing shorter rather than longer sequences.

While we present an increasingly strong case for the use of shorter clips in subjective VQA, it is important to note that we do so under specific conditions. Firstly, all of the clips we have used have been temporally-consistent and feature no significant shift in content throughout their duration (and the clip that deviated the most from this constraint - Divers - also produced the most anomalous results). Our findings, therefore, only apply to studies that contain clips that maintain a high level of consistency. Our recommendations apply directly to DSCQS and its SS variant, that is, double- and single-stimulus designs that use the continuous quality scale. Our research has not used categorical scales such as those used in the absolute category rating (ACR) and the double stimulus impairment scale (DSIS). There are theoretical reasons why the translation of results gained using continuous scales to contexts that use discrete scales can be problematic [25, 26]; however, recent work has demonstrated very strong correlations between the continuous and discrete variants of multiple different rating scales [3, 27]. For example, the average correlation between scores obtained in DSCQS and eight other methodologies (including DSIS, ACR and SAMVIQ) was 0.98 [27]. Therefore, while we reserve our strongest recommendations for studies that use the continuous quality scale, we also believe our results will be useful to those that use double and single presentations with a discretised rating scale.

Future research in this area may be directed towards extending these findings to other quality assessment paradigms such as those that use simultaneous or multiple presentation schemes. Additionally, it is also not clear from the current set of results what the optimal presentation time is for subjective image quality assessment studies. A similar set of experiments exploring the impact of presentation time on static image quality assessment may also lead to valuable findings leading to efficiency gains in the associated field.

By providing mounting empirical justification for the use of shorter sequences in subjective video quality assessment studies, we hope researchers will benefit from the significant associated efficiency gains in time, labour and money.

References

- [1] C. V. N. Index, The zettabyte era—trends and analysis, Cisco white paper.
- [2] D. R. Bull, Communicating pictures: A Course in Image and Video Coding, 1st Edition, Academic Press, Oxford, UK, 2014.
- [3] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, A. Raake, Study of Rating Scales for Subjective Quality Assessment of High-Definition Video, *IEEE Transactions on Broadcasting* 57 (1) (2011) 1–14.
- [4] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, L. K. Cormack, Study of Subjective and Objective Quality Assessment of Video, *IEEE Transactions on Image Processing* 19 (6) (2010) 1427–1441.
- [5] F. J. Mercer Moss, K. Wang, F. Zhang, R. Baddeley, D. Bull, On the optimal presentation duration for subjective video quality assessment, *Circuits and Systems for Video Technology*, *IEEE Transactions on PP* (99) (2015) 1–1. doi:10.1109/TCSVT.2015.2461971.
- [6] Methodology for the subjective assessment of the quality of television pictures, Tech. Rep. ITU-R BT.500-13, International Telecommunication Union, Geneva, Switzerland (Jan. 2012).
- [7] M. H. Pinson, S. Wolf, Comparing Subjective Video Quality Testing Methodologies, in: T. Ebrahimi, T. Sikora (Eds.), *Visual Communications and Image Processing 2003*, SPIE, 2003, pp. 573–582.
- [8] Y. Tsivian, G. Civjans, *Cinematics: Movie measurement and study tool database* (2011).
- [9] B. W. Tatler, R. J. Baddeley, I. D. Gilchrist, Visual correlates of fixation selection: effects of scale and time, *Vision Research* 45 (5) (2005) 643–659.
- [10] P. Fröhlich, S. Egger, R. Schatz, M. Muhlegger, K. Masuch, B. Gardlo, QoE in 10 seconds: Are short video clip lengths sufficient for Quality of Experience assessment?, in: *2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012)*, IEEE, 2012, pp. 242–247.
- [11] J. E. Cutting, K. L. Brunick, J. E. DeLong, C. Iricinschi, A. Candan, Quicker, faster, darker: Changes in Hollywood film over 75 years, *i-Perception* 2 (6) (2011) 569–576.
- [12] B. B. Murdock Jr, The serial position effect of free recall., *Journal of Experimental Psychology* 64 (5) (1962) 482–488.
- [13] F. J. Mercer Moss, R. J. Baddeley, N. Canagarajah, Eye Movements to Natural Images as a Function of Sex and Personality, *PLoS ONE* 7 (11) (2012) e47870.

- [14] R. P. Aldridge, D. Hands, D. Pearson, N. Lodge, Continuous quality assessment of digitally-coded television pictures, in: Vision, Image and Signal Processing, IEE Proceedings-, Vol. 145, IET, 1998, pp. 116–123.
- [15] M. H. Pinson, M. Barkowsky, P. Le Callet, Selecting scenes for 2D and 3D subjective video quality tests, EURASIP Journal on Image and Video Processing 2013 (1) (2013) 50.
- [16] Subjective video quality assessment methods for multimedia applications, Tech. Rep. ITU-T P.910, International Telecommunication Union, Geneva, Switzerland (Sep. 1999).
- [17] Methodology for the subjective assessment of video quality in multimedia applications , Tech. Rep. ITU-R BT.1788, International Telecommunication Union, Geneva, Switzerland (Feb. 2007).
- [18] S. Winkler, Analysis of Public Image and Video Databases for Quality Assessment, Selected Topics in Signal Processing, IEEE Journal of 6 (6) (2012) 616–625.
- [19] E. Svensson, Comparison of the quality of assessments using continuous and discrete ordinal rating scales, Biometrical Journal 42 (4) (2000) 417–434.
- [20] VQEG, Report on the Validation of Video Quality Models for High Definition Video Content, Tech. rep. (Jun. 2010).
- [21] M. A. Papadopoulos, F. Zhang, D. Agraftotis, D. R. Bull, A video texture database for perceptual compression and quality assessment, in: Proc. IEEE Int Conf. on Image Processing, 2015.
- [22] American National Standard for Telecommunications - Digital Transport of One-Way Video Signals - Parameters for Objective Performance Assessment., Tech. Rep. ANSI T1.801.03-1996, Institute for Telecommunication Sciences, Boulder, Colorado, USA (2003).
- [23] F. Zhang, D. R. Bull, A Parametric Framework for Video Compression Using Region-Based Texture Models, IEEE Journal of Selected Topics in Signal Processing 5 (7) (2011) 1378–1392.
- [24] D. C. Howell, Statistical Methods for Psychology, 8th Edition, Cengage Learning, 2012.
- [25] B. L. Jones, P. R. McManus, Graphic Scaling of Qualitative Terms, Smppte Journal 95 (11) (1986) 1166–1171.
- [26] M. T. Virtanen, N. Gleiss, M. Goldstein, On the use of evaluative category scales in telecommunications, in: Proc. Human Factors Telecommun, Melbourne, Australia, 1995, pp. 253–260.

- [27] T. Tominaga, T. Hayashi, J. Okamoto, A. Takahashi, Performance comparisons of subjective quality assessment methods for mobile video, in: Quality of multimedia experience (QoMEX), 2010 second international workshop on, IEEE, 2010, pp. 82–87.

(a)

(b)

(c)

(d)

(e)

(f)

Figure 6: Viewer ratings for each of the five duration groups, averaged over all content but separated by distortion level. The plots in the first column (a,c,e) show data from the single stimulus experiment, while plots in the second column (b,d,f) show data from the double stimulus experiment. Plots in the top row (a,b) show DMOS data, plots in the middle row (c,d) show DMOS error data and plots in the bottom row show MOS data. Error bars represent the standard error of the mean.

Figure 7: Viewer ratings for each of the five duration groups, averaged over all distortions but separated by sequence. The plots in the first column (a,c,e) show data from the single stimulus experiment, while plots in the second column (b,d,f) show data from the double stimulus experiment. Plots in the top row (a,b) show DMOS data, plots in the middle row (c,d) show DMOS error data and plots in the bottom row show MOS data. Error bars represent the standard error of the mean.

Figure 8: Confidence scores for SS (a) and DS (b) experiments, plotted against duration. Error bars represent the standard error of the mean.

(a)

(b)

Figure 9: Entropy of fixation distributions plotted as a function of time. (a) includes all data in the distributions while (b) plots the data from each duration block separately. Error bars represent the standard error of the mean.