CrossMark

RESEARCH ARTICLE

Symptoms analysis of 3D TV viewing based on Simulator Sickness Questionnaires

Kjell Brunnström[1,2] (ID) · Kun Wang[1] · Samira Tavakoli[3] · Börje Andrén[1]

Abstract Stereoscopic 3D TV viewing puts different visual demands on the viewer compared to 2D TV viewing. Previous research has reported on viewers' fatigue and discomfort and other negative effects. This study is to investigate further how severe and what symptoms may arise from somewhat longish 3D TV viewing. The MPEG 3DV project is working on the next-generation video encoding standard and in this process, MPEG issued a call for proposal of encoding algorithms. To evaluate these algorithms a large scale subjective test was performed involving Laboratories all over the world [(MPEG 2011; Baroncini 2012)]. For the participating Labs, it was optional to administer a slightly modified Simulator Sickness Questionnaire (SSQ) before and after the test. One of the SSQ data sets described in this article is coming from this study. The SSQ data from the MPEG test is the largest data set in this study and also contains the longest viewing times. Along with the SSQ data from the MPEG test, we have also collected questionnaire data in three other 3D TV studies. We did two on the same 3D TV (passive film pattern retarder) as in the MPEG test, and one was using a projector system. As comparison SSQ data from a 2D video quality experiment is also presented. This investigation shows a statistically significant increase in symptoms after viewing 3D TV primarily related to the visual or Oculomotor system. Surprisingly, 3D video viewing using projectors did not show this effect.

Introduction

It is quite clear now that the Hollywood strategy to re-introduce 3D movies has achieved a great success. The movie theaters have struggled a few years and losing spectators gradually to more and more potent home cinema systems. Now 3D film presentation has established itself as the most profitable movie category, where people are prepared to pay up to 50% more for the tickets. For 3D TV the situation is more complicated. At first, there was a big buzz from the TV-manufacturers hoping that consumers would immediately jump onto the new trend, but this was not the case. There are many factors involved which need to fall into place for 3D TV at home to have extensive usage. At the moment, the lack of 3D content to watch is a factor that makes it less attractive for consumers to invest in a new 3D TV. In the other end of the scale, the broadcasters have not yet launched so many 3D TV channels, although their numbers are also slowly increasing. The TV manufacturers have met this problem with bundling the 3D capability with the higher end TVs, so even if the targeted demand for 3D TVs is not that high, the number of 3D capable TV set are steadily increasing. Soon, it is, therefore, likely that the critical mass of the number of 3D capable TV set and the availability of content are high enough to make the market boost. Remember it has taken quite some time, 20–30 years, for HDTV to become a

✉ Kjell Brunnström
   kjell.brunnstrom@acreo.se

[1] Netlab: Visual Media Quality, Acreo Swedish ICT AB, Electrum 236, 164 40 Kista, Sweden

[2] Department of Information Technology and Media (ITM), Mid Sweden University, Sundsvall, Sweden

[3] Universidad Politécnica de Madrid, Madrid, Spain

commodity and the transition from standard definition TV is far from finished. The acceptance and final success of 3D TV are, among other things, depending on whether the viewing of 3D TV will induce any negative effects in the viewing experiences of the users or not.

Since the revival of the 3D movies, discussions and investigations about how to deliver and code 3D TV (e.g., Meesters et al. (2004), Wang et al. (2012)), as well as any potentially negative effects of viewing 3D video content (e.g., Lambooij et al. (2010) and Urvoy et al. (2013)), have been ongoing. In this context, we are only discussing stereoscopic 3D with eyeglasses. It may also apply to some autostereoscopic display systems.

Kennedy et al. (1993) developed a questionnaire for investigating the potentially negative effects of the usage of visual simulators (Kennedy et al. 1993), which was named Simulator Sickness Questionnaire (SSQ). They based it on the earlier developed Pensacola Motion Sickness Questionnaire (MSQ), where they recognized that some symptoms in MSQ were less relevant or could even be misleading, so Kennedy et al. (1993) deleted them in the SSQ. Furthermore, Kennedy et al. (1993) proposed how to group and analyze the SSQ based on a large number of data for simulators and factor analysis. 3D TV viewing has some similarities to visual simulators; we have, therefore, administered it as a part of some 3D TV subjective experiments performed at the research institute Acreo Swedish ICT in Sweden (Acreo Lab). We have also compared it to SSQ data collected at 2D TV subjective experiments.

The SSQ has been used in similar work previously. Takada and Matsuura (2013) used it in a comparison between viewing 3D movie on an LCD display, and a head mounted display. They did not find any significant differences based on SSQ among their different 3D movie stimuli. They found that sickness symptoms appeared more often after the test persons have been viewing the 3D movies, although there were substantial individual differences. Naqvi et al. (2013) compared 2D and 3D and found that there was a significant increase in the symptoms for 3D. The 3D viewing time was about 10 min in their study (Naqvi et al. 2013), which is shorter than in the current investigation (25 min). In Vlad et al. (2013) SSQ was used to compare 3D TV with immersive 3D glasses (a kind of head-mounted display) with a relatively large number of test subjects, which found a significant increase of the SSQ reported symptoms on the 3D viewing both for 3D TV and the immersive 3D glasses, although in a different way for the two 3D viewing technologies. In Jumisko-Pyykkö et al. (2010), SSQ was used for evaluating the visual discomfort in different dual-view autostereoscopic mobile screens with varying video quality, and under different viewing length. They

observed that in general short-term video viewing in these displays is not disturbing. In Wibirama and Hamamoto (2014), Visually Induced Motion Sickness (VIMS), an important safety issue in 3D technology, was investigated based on recording SSQ, heart rate variability, and depth gaze behavior. Their results indicated that nausea and disorientation symptoms increased as the dynamic motion increased in the presented video. Also, to reduce VIMS, the user should perform gaze fixation at one point when experiencing vertical and horizontal motion in 3D content. Using SSQ, Häkkinen et al. (2002) investigated the potential effects induced by watching the head-mounted display (HMD). The results showed that there was no general HMD symptomology, but the symptoms should always be related to specific tasks and technologies, e.g., in their study the stereoscopic game playing was relatively nauseogenic and induced postural sway, but the movie watching with the same technology was relaxing experience.

The terms fatigue and discomfort is often used to describe the negative effects induced by the 3D TV systems. These terms have been used quite differently by different authors, but we will use them following Urvoy et al. (2013).

The MPEG 3DV project was working on the next-generation video encoding standard, and in this process, MPEG issued a call for proposal (MPEG 2011) of encoding algorithms. To evaluate these algorithms a large scale subjective test was performed involving Laboratories all over the world. For the participating Labs, it was optional to administer a slightly modified Simulator Sickness Questionnaire (SSQ) before and after the test. One of the SSQ data sets described in this article is coming from this study, Brunnström et al. (2013). The SSQ data from the MPEG test is the largest data set in this study and also contains the longest viewing times.

Along with the SSQ data from the MPEG test, we have also collected questionnaire data in three other 3D TV studies. We did two on the same 3D TV (passive film pattern retarder) as in the MPEG test, and one was using a projector system. As comparison SSQ data from a 2D video quality experiment is also presented. Although for some of the experiments we have SSQ data collected in the break between the Sessions, we have here concentrated the analysis to the pre- and post-experiment SSQ data, since this data was available from all studies.

Method

For easier understanding and interpretation of the results, an overview of the test set-ups and methods for the different test will be given here and in Table 1.

Table 1 Overview of the test conditions of the different experiments

Experiment	1	2	3	4	5
Test method	Double stimulus impairment scale (DSIS)	Single stimulus—3 scales (3D Realism", "Depth Quantity" and "Video Quality)	Single stimulus—3 scales (Visual Quality, Visual Discomfort and Sense of Presence)	Double stimulus impairment scale (DSIS)	Single stimulus—2 scales (Quality + Impairment observation)
Screening	Visual acuity/Ishihara/Randot/dominant eye	Visual acuity/Ishihara/Randot	Visual acuity/Ishihara/Randot	Visual acuity/Ishihara/Randot	Visual acuity/Ishihara/Randot
Content	Poznan_Hall2;Poznan_Street;Undo_Dancer;GT_Fly; Kendo;Balloons;Lovebird1; Newspaper	NAMA3DS1—COSPAD1	Documentary and three movies	Movie	Movie, documentary, music, sports
Degradations	Coding and view synthesis; fixed bitrate	NAMA3DS1—COSPAD1	2D, compression, geometrical distortion, temporal mismatch	Crosstalk (0, 2, 7, 12, and 20%) + system crosstalk (passive and active)	Adaptive video streaming
SI	Min = 28, Max = 71, Mean = 49	Min = 36, Max = 101, Mean = 67	Min = 44, Max = 79, Mean = 62	Min = 38, Max = 115, Mean = 77	Min = 32, Max = 67, Mean = 48
TI	Min = 8, Max = 28, Mean = 18	Min = 4, Max = 56, Mean = 22	Min = 7, Max = 33, Mean = 18	Min = 11, Max = 84, Mean = 55	Min = 18, Max = 85, Mean = 52
DSI	Min = 0.8, Max = 18, Mean = 3.5	Min = 12, Max = 25, Mean = 20	Min = 0.6, Max = 6.2, Mean = 3.7	Min = 2.8, Max = 8.2, Mean = 5.0	N/A
DTI	Min = 0.5, Max = 38, Mean = 4.5	Min = 7, Max = 18, Mean = 13	Min = 0.6, Max = 5.7, Mean = 2.4	Min = 1.7, Max = 25, Mean = 12.7	N/A
Disparity uncrossed (D+)	Min = 20, Max = 0, Mean = −5.9, Median = −2.5	Min = −14, Max = 17, Mean = −6.2, Median = −6.5	Min = 12, Max = 31, Mean = 21.1, Median = 19.5	Min = −10, Max = 37, Mean = 24.6, Median = 30	N/A
Disparity crossed (D−)	Min = −49, Max = −8, Mean = −20.9, Median = −15	Min = −3, Max = 26, Mean = 11.4, Median = 9.5	Min = −24, Max = −5, Mean = −12.6, Median = −12	Min = −46, Max = 2, Mean = 23.7, Median = 25	N/A
Viewing distance	3.6 m (6H)	1.7 m (3H) and 2.8 m (5H)	2.3 m (4H)	3 m (3H)	2.3 m (4H)
Display device	Passive 3D TV (Hyundai S456D)	Passive 3D TV (Hyundai S456D)	Passive 3D TV (Hyundai S456D)	Passive + active 3D projector	2D HDTV (Hyundai S456D)
Ambient illumination	≈20 lx, 6500 K	≈20 lx, 6500 K	≈20 lx, 6500 K	≈20 lx, 6500 K	≈20 lx, 6500 K
Test duration	30–95 Min	38 Min	48 Min	50 Min	60 Min
Break time	5 Min	5 Min	10 Min	5 Min	5 Min
Number of sessions	2–8	2	2	2 (1 active and 1 passive)	2
Number of votes per session	28	55	63	35	66

Table 1 continued

Experiment	1	2	3	4	5
Max number of subjects per session	3	1	1	1	1
Number of subjects	70	28	24	26	23
Age range	16–72 (mean 34)	18–62 (mean 34)	16–61 (mean 29)	14–53 (mean 27)	18–68 (mean 30)
Gender ratio	20 (f)/48 (m)	9 (f)/19 (m)	7 (f)/17 (m)	12 (f)/14 (m)	7 (f)/16 (m)
Naive/expert	Naive	Naive	Naive	Naive	Naive
Excluded subjects	None screened	1 Pre-screened + 2.5 post-screened	1 Post-screened	None screened	None screened
References	MPEG (2011), Baroncini (2012), Brunnström et al. (2013a), Perkis et al. (2012)	Brunnström et al. (2013b), Urvoy et al. (2012)	Kulyk et al. (2013)	Wang et al. (2014)	Tavakoli et al. (2015), Tavakoli (2015)

Common for all the studies both 3D and 2D is that they are Laboratory studies of video quality based on standardized methods from the ITU, such as ITU-R Rec. BT.500-13 (2012), ITU-T Rec. P.910 (1999) and ITU-T Rec. P.913 (2014). The primary task for the test subjects has been to rate their experiences on rating scales based on viewing shorter video clips. Then in conjunction with these tests, the SSQ has been administered. The specific experiments have been all previously published and described, so we will therefore not go into detail on any of the results from these studies, apart from the SSQs. The different subject experiments were:

Subjective experiment 1 or Exp 1 The main target of the test was to collect subjective opinion scores for evaluating different 3D video coding algorithms for the MPEG 3DV project (Perkis et al. 2012).

Subjective experiment 2 or Exp 2 Test of different rating scales and viewing distance for 3D TV using an open 3D video database NAMA3DS1-COSPAD1 (Brunnström et al. 2013b).

Subjective experiment 3 or Exp 3 Test of different rating scales for 3D TV using video containing both coding impairments and geometrical distortions (Kulyk et al. 2013).

Subjective experiment 4 or Exp 4 Test of the impact of crosstalk on 3D video viewing (Wang et al. 2014).

Subjective experiment 5 or Exp 5 2D video quality experiment that was targeting HTTP adaptive video streaming (Tavakoli et al. 2015).

For all the experiments we had followed the common practice that before the actual test, each subject was given written instructions and also the opportunity to ask questions about the procedure if anything was unclear. A training session was performed to familiarize the subjects with the test method and give them a sense of the range of qualities that were involved in the test. Each test subject was greeted and guided to the pre-screening locations. If there were two or three test persons at the same time, they were kept separated during pre-screening, so that no-one could know the results of the others. Furthermore, the test subjects were asked not to discuss the test with other potential test subjects after they had performed the test. The name of test subject was also anonymous for the test leader. A separate person administrated the booking of the test persons. He/she attached a randomly generated identity code to the subject from a list, and also marked this code on all the papers, files or documents that belonged to that subject. We screened each test subject for visual acuity, color vision (Ishihara), and stereo acuity through a Randot test (not Exp 5). A test to find the dominating eye was also performed and recorded (not Exp 5). The SSQ was filled in before the test, and the instructions were given to the

subject to read. Sometimes, if there was a waiting time between the subjects the order in which they performed visual screening, reading the instructions and filling in the SSQ were different between them, to reduce the idle time before starting. Then all subjects in the test group were gathered in the lab room and asked if they had any questions about the instructions. Each viewer adjusted the height of their chair so that the position of his/her eyes was at about the same as the height of the center of the TV. We seated a maximum number of 3 viewers in front of the screen at the same time. (only Exp 1 had more than 1 test subject at the time). After answering any questions of the subjects, a training session, was performed. During the training session, the test leader was in the room, helping or answering questions if needed. Then the main viewing sessions took place (see further below about viewing and session durations as well as the number of sessions, etc.). After the test a new SSQ with the same questions as before was answered by the subjects. Afterward, the test subjects were rewarded with cinema tickets to a value corresponding to one or two visits to a 3D movie (different in different Experiments).

The tests were performed in the Acreo Lab, which conforms to ITU-R Rec. BT.500 (2012), using a Hyundai S46556D, a passive film pattern retarder stereoscopic 3D TV except for Exp 4 where a 3D projector was used (see more detail below). The peak white luminance of TV was 177 cd/m^2 (78 cd/m^2 through eye-glasses). The stereo views for the 3D TV were off-line vertically sub-sampled in half, spatially interlaced and added with a gray surround if needed to match the TV's native 2D resolution of 1920×1080. We did the spatial interlacing so that every second row corresponded to the correct left or right view and was playable as 2D videos. The ambient illuminance level in the room was about 20 lx using D_{65} high-frequency fluorescent tubes giving a color temperature of the light of 6500 K.

The viewers were of various social backgrounds, occupations and normally recruited through mail advertisement through a company contact register, personal contacts, advertisement on the web and the company's homepage. The age ranges were broad for all studies, and we tried to balance to gender ratio, but we was in most cases easier to recruit male test persons than females.

Subjective experiment 1

The area utilized for the Exp 1 was 5 m long and 3.6 m in width. The TV was placed 0.8 m from the back wall and the viewer 3.6 m (6H) from the front side of the TV.

In total 70 test subjects or viewers participated in the experiment.

Viewing time

A session took about 12–13 min to complete. The test persons typically completed two sessions continuously and then we enforced a break. No viewer was running more than two sessions without a break, which means that the maximum continuous viewing time was about 25 min. The participating viewers completed 2–8 sessions, ranging from a viewing time of 25 min up to 90 min and including the training session of about 5 min it was 30–95 min, see Table 2, for a more detailed distribution of the viewing times including the training session.

Subjective experiment 2

In Exp 2 we used the NAMA3DS1—COSPAD1 video dataset (Urvoy et al. 2012) and was designed for comparing three different rating scales and two viewing distances (Brunnström et al. 2013b). The three scales were: Visual Quality (VQ), Visual Discomfort (VD) and Sense of Presence (SP). We based our experimental design on the Absolute Category Rating (ACR) scale (ITU-T 1999) with five levels for the Visual quality scale and the Sense of Presence scale. We derived the Visual Discomfort scale on the Degradation Category Rating scale (ITU-T 1999). We divided the test into two sessions, and we then placed the test subjects on two different viewing distances, either 3H or 5H, in the two sessions (randomized order).

In an earlier analysis of the scaling data and the influence of viewing distance published in Brunnström et al. (2013), we did not find any statistically significant effect on the viewing distance. We have therefore chosen to analyze both viewing distances together in this study.

A modified version of a video player, AcrVQWin (Jonsson and Brunnström 2007), developed by the authors was used to present and retrieve the responses from the test subjects.

Viewers

The test subjects were of different background and age. There were 28 test subjects in total, and we post screened

Table 2 The number of sessions taken by how many subjects and the total viewing time including the training session

Number of sessions	Number of subjects	Viewing time (min)
2	1	30
4	10	55
6	3	80
7	53	92.5
8	1	95

2.5 test subject's data (1 test subject was post-screened in one session hence 0.5) based on the procedure used by VQEG in their HDTV test (VQEG 2010), and we discarded one test subject due to pre-screening of visual ability. There were 14 Swedish subjects and 14 international. The native Swedish speaking test subjects did the experiment in Swedish, and the international observers did it in English.

Viewing time

A total of 110 three-dimensional PVSs (10 SRCs × 11 HRCs), where the duration of each sequence was 16 s except for the eleven PVSs with SRC10 where they instead were 13 s long each. That gives a pure 3D video viewing time of 29 min and if we include the voting time as in Exp 1, which could be estimated here to about 5 s. then the total time was about 38 min.

Subjective experiment 3

Exp 3 (Kulyk et al. 2013) is to some extent similar to Exp 2, in that it uses three rating scales for voting, but there was a broader range of impairments and some that were more demanding to view than in Exp 2.

The voting scales used in the test were "3D Realism", "Depth Quantity" and "Video Quality," with discrete five level category scales. 13 source stereoscopic video sequences (SRC), chosen from one documentary and three movies. When we made the scene selection, we avoided scene changes. We divided them into three content types:
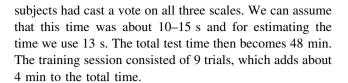
- Content 1—recorded with a still camera and containing small amount of motion (standing or sitting people)
- Content 2—recorded with a still camera and containing a moderate amount of motion.
- Content 3—recorded using a Zoom with or without a moving camera and containing a moderate/large amount of motion.

Viewers

25 naïve test subjects participated; only one subject performed the test at a time. One subject was rejected and thus removed from the final analysis due to inadequate results in the stereo vision test. The total number of subjects after screening was 24.

Viewing time

The test consisted of at total of 126 PVS of 10 s each, plus voting time, which we divided into two sessions with a 10 min break in between. The voting time was flexible in that the test software did not play the next video until the

subjects had cast a vote on all three scales. We can assume that this time was about 10–15 s and for estimating the time we use 13 s. The total test time then becomes 48 min. The training session consisted of 9 trials, which adds about 4 min to the total time.

Subjective experiment 4

In Exp 4 we varied the crosstalk level in movie-like content. We used a 3D projection system which could be utilized both with active and passive eyeglasses. The purpose of the test was to evaluate passive 3D projector system, but also to get some insight into the relationship between crosstalk and how visible and annoying the ghosting distortions are.

We measured crosstalk objectively at the center of the screen. The measurement method adheres to ICDM standard (2012). The objective measured crosstalk from the projection system itself was about 0.3% for the system using active shutter eyeglasses and 2% for the system using passive polarized glasses (polarization modulator contributed less than 1%, the rest was due to other components in the system, e.g., silver screen).

We based the procedure used for adding the crosstalk on the measured system gamma function of the projector including the screen, which was found to be:

$$L = 31.53 \times \left(\frac{Y}{255}\right)^{2.15}$$

where L is the luminance that was measure and Y is the digital input Luma- or gray values (0 corresponds to black, and 255 to white). The crosstalk is light leakage between the views, so the video Luma-values were transformed into Luminance and the crosstalk were added in this domain using the following equations

$$L_{left}^{crosstalk} = L_{left}^{original} + C \cdot L_{right}^{original}$$
$$L_{right}^{crosstalk} = L_{right}^{original} + C \cdot L_{left}^{original}$$

where C is the added crosstalk. We applied the formulas per pixel and added an equal amount of crosstalk in both left and right views. Then the luminance values were transformed back using the inverse gamma function and stored in the images.

The experiment consisted of two main sessions: (a) passive projector system using passive polarized eyeglasses, and (b) active projector system using active shutter eyeglasses. The subjects saw the same test video set in both sessions.

The subjective experiment used Double Stimulus Impairment Scale (DSIS) as defined in ITU-R Rec. BT.500-13 (2012), using the five graded scale: imperceptible, perceptible but not annoying, slightly annoying,

annoying and very annoying. We selected seven stereo-scopic cinema contents and processed them in five simulated crosstalk levels (0, 2, 7, 12, and 20%) plus the 2% system crosstalk for the passive system and plus 0.3% system crosstalk for the active system for the subjective experiment.

The set-up consisted of a DepthQ® HD3D projector from LightSpeed with a polarizing modulator from LC-Tec in front of the projector lens and a silver screen to project the sequences on for the passive eyeglasses. For the active eyeglasses, we removed the polarization modulator. The active eyeglasses were NVIDIA Stereovision and were controlled by an NVIDIA graphics card.

Viewers

In this study, we recruited test persons from Stockholm University notice boards and different forums on Facebook, in addition to our normal way described above. The total number of test subjects that participated in the test was 26. Also in contrast to our normal age ranges used most participants were young students between 20 and 30 years old. Participants were non-expert or in fields not directly related to S3D video as part of their professional work.

Viewing time

We split the test into two sessions; each session was about 26 min and totally about 52 min. The sessions consisted of 35 trials. A trial there was initiated with a picture that showed the text "Reference Video" for 2 s followed by the actual reference video for about 15 s. Then a picture with text "Processed Video" appeared for 2 s, and the processed video sequence was presented. After which the voting interface was shown until the subject had given its rating. We observed that some people voted rather quickly while others took a longer time to vote. We are assuming a mean voting time of 5 s. The total time of a trial is then 39 s and with 35 PVS a total viewing time of 22.7 min per session and about a total of about 3 min voting time.

Subjective experiment 5

Exp 5 is a 2D video subjective experiment for assessing adaptive video streaming QoE and used as our 2D control experiment. For this experiment seven 6 min, 2D video contents in different types were chosen among commercial video contents. The characteristics of the contents were different containing from smooth to sudden motions, smooth scene change transitions to fast scene change, and recorded using a still, a zoom or a moving camera. On the other hand, the chosen sequences spanned a considerable portion of the spatial–temporal information plane.

We applied eight different HRCs simulating different adaptive streaming scenarios applied to the video content. The six minutes long videos were cut into smaller pieces with a length depending on the HRC type. A PVS with a gradual change with 10 s chunks was longer than a PVS with rapid change with 2 s chunks. Furthermore, we did apply all HRC to each of these smaller pieces. In total 132 PVSs were used in the experiment.

Following the ACR method specification, after presentation of each PVS, the subjects were asked to evaluate the sequence by voting for two different questions: the overall quality of the PVS ranging from Bad (1) to Excellent (5) and if they have perceived any change in the quality by stating the type of the change.

Viewers

The test subjects were of different ages and background. There were 7 female and 16 male, including 4 Swedish and 19 international. Four of them had subscriptions from the streaming media service providers (specifically Netflix).

Viewing time

Each PVS had a length ranging 14–45 s. The voting time in between was as long as the test subject wanted, but usually, they responded quite quickly. We assume an average of 5 s. There were in total 132 PVS. The total viewing time including voting was about 60 min.

Simulator sickness questionnaire

The simulator sickness questionnaire or SSQ we used in this study is shown in Table 3. This is a modified version as compared to the SSQ proposed by Kennedy et al. (1993), as it has one more level than the original. The participating Labs in MPEG 3DV used this modified version of the SSQ, and we have therefore continued to use it for being able to compare results.

Statistical analysis

The questionnaire answers were translated into a number in our case by None = 0, Slight = 1, Moderate = 2, Strong = 3, Severe = 4 for allowing parametric statistical analysis, but we performed a non-parametric analysis also on the voting of the individual symptoms. Pairwise T test, Kolmogorov-Smirnoff and Mann–Whitney tests were performed for the means of each symptom of the SSQ, testing for statistically significant difference for their values before

Table 3 Simulator Sickness Questionnaire (SSQ) used in the test

	1	2	3	4	5
General discomfort	None	Slight	Moderate	Strong	Severe
Fatigue	None	Slight	Moderate	Strong	Severe
Headache	None	Slight	Moderate	Strong	Severe
Eye strain	None	Slight	Moderate	Strong	Severe
Difficulty focusing	None	Slight	Moderate	Strong	Severe
Increased salivation	None	Slight	Moderate	Strong	Severe
Sweating	None	Slight	Moderate	Strong	Severe
Nausea	None	Slight	Moderate	Strong	Severe
Difficulty concentrating	None	Slight	Moderate	Strong	Severe
Fullness of head	None	Slight	Moderate	Strong	Severe
Blurred vision	None	Slight	Moderate	Strong	Severe
Dizzy (eyes open)	None	Slight	Moderate	Strong	Severe
Dizzy (eyes closed)	None	Slight	Moderate	Strong	Severe
Vertigo	None	Slight	Moderate	Strong	Severe
Stomach awareness	None	Slight	Moderate	Strong	Severe
Burping	None	Slight	Moderate	Strong	Severe

and after. We also calculated a repeated measure analysis of variance (ANOVA) followed by a Tukey HSD post hoc test, on whether there was a significant impact on time on the different questions.

Kennedy et al. (1993) suggested a statistical analysis for the SSQ by grouping the different symptoms into three groups: Nausea (N), Oculomotor (O) and Disorientation (D). They also calculated a total score (TS). The Nausea symptom group contained the symptoms nausea, stomach awareness, increased salivation and burping. The Oculomotor grouped eyestrain, difficulty focusing, blurred vision, and headache. The symptom group Disorientation included the symptoms dizziness and vertigo. They are not completely disjoint since a few of the variables are used when calculating the scores in more than one group, e.g., nausea and difficulty concentrating. In Table 4 it is indicated which of the symptoms that are grouped together. The calculation is done by summing together the values with a 1 in Table 4 and then multiply that sum with factors at the bottom of the table, using the conversion between severity and numbers described above.

Results

Subjective experiment 1

The results were analyzed as described in section "Statistical analysis". The mean scores for the individual symptoms before and after along with 95% confidence intervals are shown in Fig. 1. The symptoms Fatigue, Eye-strain, Difficulty Focusing and Difficulty Concentrating, were statistically significant considering both parametric test and

Table 4 SSQ score calculations as described in Kennedy et al. (1993)

SSQ symptoms		Weight		
		N	O	D
1	General discomfort	1	1	
2	Fatigue		1	
3	Headache		1	
4	Eye strain		1	
5	Difficulty focusing		1	1
6	Increased salivation	1		
7	Sweating	1		
8	Nausea	1		1
9	Difficulty concentrating	1	1	
10	Fullness of head			1
11	Blurred vision		1	1
12	Dizzy (eyes open)			1
13	Dizzy (eyes closed)			1
14	Vertigo			1
15	Stomach awareness	1		
16	Burping	1		
	Total	[1]	[2]	[3]

$$N = [1] \times 9.54$$
$$O = [2] \times 7.58$$
$$D = [3] \times 13.92$$
$$TS = ([1] + [2] + [3]) \times 3.74$$

non-parametric, see Table 5. As shown in Fig. 1, these also had the biggest increase in mean value. The symptom of General discomfort, Sweating, Fullness of head, Blurred
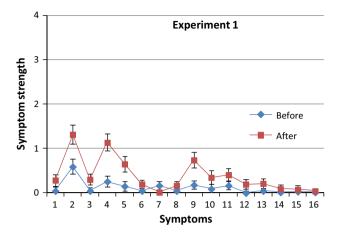
Fig. 1 The mean and 95% confidence interval for the different symptoms before and after. The numbers correspond to the order of the question in the questionnaire and are shown in Table 5

vision, Dizzy (eyes opened), Dizzy (eyes closed), were statistically significantly higher after than before in some tests. The symptom of Increased Salivation, Nausea, Vertigo, Stomach Awareness and Burping were not significant in any applied test. There was no-one that reported Severe symptoms (highest level), but several that indicated that they had strong symptoms (the second highest symptom strength). About 40% have not stated more than Slight symptom on any question.

The SSQ were also analyzed based on the procedure suggested by Kennedy et al. (1993). They suggest that the questionnaire could be analyzed in three groups: Nausea (N), Oculomotor (O) and Disorientation (D) as well as total score (TS).

The scores for the questionnaires before and after the sessions, including 95% confidence intervals, can be seen

in Fig. 2. A repeated measures ANOVA showed that the interaction effect between the grouping variable (N;O;D and TS) and time (before, after) was significant $F(3, 201) = 17{,}5$ p = 0.00, followed by the post hoc test Tukey HSD gave that the difference between before and after were significant ($p \ll 0.05$) for each of the grouping variables. The largest difference was in the Oculomotor dimension.

The effect of gender was also analyzed, but it was not found to be significant, as well as the main effect and the interaction effect. In fact, the means were very similar, so there was no tendency found.

Two and three age groups about equal size were defined to analyze if there were any difference due to age. The age boundaries for the division into two groups were: 16–30 and 31–72 years of age. There were 37 viewers in the younger group and 31 in the older group. For the division into three groups, the following age boundaries were used: 16–25, 26–40 and 40–72 years of age, resulting in 24 viewers in the youngest group, 25 in the mid-aged group and 19 in the older group. There was a tendency that the younger group in both age group divisions gave slightly higher scores both before and after the sessions. However, no effects were significant.

Subjective experiment 2

The mean scores for the individual symptoms before and after for Exp 2, along with the 95% confidence intervals are shown in Fig. 3. The results from a repeated measures ANOVA gave that the main effects of both the time, i.e., before compared to after and the symptoms were significant $F(1, 27) = 9.21$ $p = 0.005$ and $F(15, 405) = 8.06$

Table 5 Outcome of different statistical tests with 95% significance level

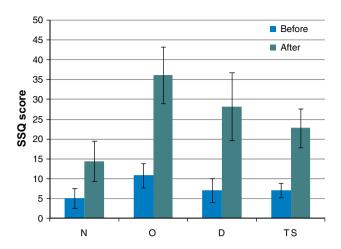|		T test	Kolmogorov–Smirnov	Mann–Whitney	Tukey HSD
1	General discomfort	0.25	$p > .10$	0.04	0.05
2	Fatigue	0.00	$p < .001$	0.00	0.00
3	Headache	0.00	$p > .10$	0.04	0.02
4	Eye Strain	0.00	$p < .001$	0.00	0.00
5	Difficulty focusing	0.00	$p < .025$	0.00	0.00
6	Increased salivation	0.05	$p > .10$	0.37	0.88
7	Sweating	0.01	$p > .10$	0.18	1.00
8	Nausea	0.09	$p > .10$	0.46	0.99
9	Difficulty concentrating	0.00	$p < .005$	0.00	0.00
10	Fullness of head	0.00	$p < .10$	0.02	0.00
11	Blurred vision	0.01	$p > .10$	0.05	0.00
12	Dizzy (eyes open)	0.00	$p > .10$	0.10	0.88
13	Dizzy (eyes closed)	0.02	$p > .10$	0.23	0.73
14	Vertigo	0.05	$p > .10$	0.46	1.00
15	Stomach awareness	0.30	$p > .10$	0.66	1.00
16	Burping	0.41	$p > .10$	0.77	1.00

Fig. 2 SSQ scores calculated according to Kennedy et al. (Kennedy et al. 1993). *N* Nausea, *O* Oculomotor, *D* Disorientation, *TS* Total Score
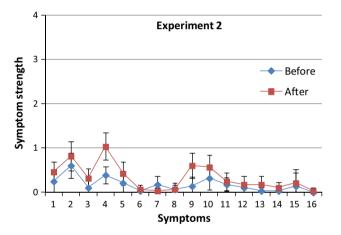

Fig. 3 The mean and 95% confidence interval for the different symptoms before and after for Exp 2

$p = 0.000$, as well as the interaction $F(15, 405) = 3.16$ $p = 0.000$. The post hoc shows this comes from that the symptoms Eye-strain ($p = 0.000$) and Difficulty Concentrating ($p = 0.004$) were significant.

Subjective experiment 3

The mean scores for the individual symptoms before and after for Experiment 3, along with the 95% confidence intervals are shown in Fig. 4. The results from a repeated measures ANOVA gave that the main effects of both the time, i.e., before compared to after and the symptoms were significant F(1, 27) = 21.3 $p = 0.000$ and F(15, 405) = 4.83 $p = 0.000$, as well as the interaction F(15, 405) = 2.36 $p = 0.003$. The post hoc shows this comes from that the symptoms Eye-strain ($p = 0.0003$), Difficulty Concentrating ($p = 0.032$) and Fullness of Head ($p = 0.008$) were significant.

Fig. 4 The mean and 95% confidence interval for the different symptoms before and after for Exp 3

Subjective experiment 4

The mean scores for the individual symptoms before and after for Exp 4, along with the 95% confidence intervals are shown in Fig. 5. The results from a repeated measures ANOVA gave that the main effects of both the time, i.e., before compared to after and the symptoms were significant F(1, 23) = 11.53 $p = 0.02$ and F(15, 345) = 6.13 $p = 0.000$, but not the interaction. No symptom was even close to being significant in the post hoc test.

Subjective experiment 5

The average scores for the individual symptoms before and after for Exp 5, along with the 95% confidence intervals are shown in Fig. 6. The results from a repeated measures ANOVA gave that the main effect of time, i.e., before compared to after was not significant, but the main effect
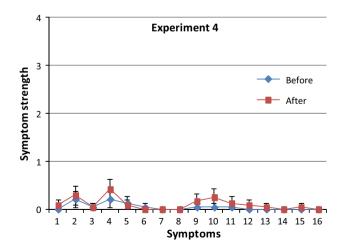

Fig. 5 The mean and 95% confidence interval for the different symptoms before and after for Exp 4

Fig. 6 The mean and 95% confidence interval for the different symptoms before and after for Experiment 5
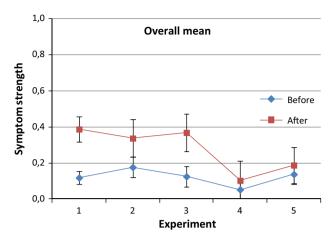

Fig. 7 Overall mean taken over all symptoms for the different experiments before and after

Fig. 8 The overall mean of the difference between the symptoms for each experiment

for the symptoms still were significant $F(15, 450) = 6.67$ $p = 0.000$. The interaction was not significant either. As in Exp 4, no symptom was even close to being significant in the post hoc test.

Cross-experiment

A repeated measures ANOVA was performed with the different experiments as between-group factor and the symptoms and time as within factor, which showed that the main effect of experiments was significantly different $F(4, 173) = 5,25$, p = 0.0005, as well as the interaction between before and after, and the different experiments $F(4, 173) = 6,06$, $p = 0.0001$. The means, and their 95% confidence intervals are shown in Fig. 7. By analysis the post hoc test (Tukey HSD), it was shown that the overall means before the experiments were not significantly different. For the overall mean after the experiments, Exp 1 was significantly different from both Exp 4 ($p = 0.0000$) and Exp 5 ($p = 0.002$). Exp 2 was only significantly different from Exp 4 ($p = 0.0062$). Exp 3 was also only significantly different from Exp 4 ($p = 0.0008$).

If we consider the difference between the symptom strength reported before and after then the overall mean of Exp 1 and 3 are significantly different from Exp 4 ($p = 0.0025$ and $p = 0.0008$) and Exp 5 ($p = 0.047$ and $p = 0.031$). The overall means are shown in Fig. 8. The symptoms giving rise to these significant effects are for Exp 1 compared to Exp 4: Fatigue ($p = 0.0029$), Eye strain ($p = 0.000$) and Difficulty focusing ($p = 0.008$). For Exp 1 compared to Exp 5 it were just the symptoms Fatigue ($p = 0.0001$) and Eye strain ($p = 0.000$) that were significantly different. The Fatigue in Exp 1 was also significantly different from the Fatigue in Exp 2 ($p = 0.037$). However, for Exp 3 no individual symptom was significantly different from the corresponding symptom in the other tests, but the overall significance was borderline.

We can also analyze the strength of symptoms based on the analysis suggested by Kennedy et al. (1993). The results are shown in Fig. 9. Tukey HSD post hoc tests indicate that the symptom group of Nausea, Oculomotor, Disorientation and Total Scores were significant on an at least a 95% confidence level after compared to the same symptom group in the same experiment before, in Experiment 1–3, but not for Exp 4 and 5. However, disorientation for Exp 5 has a significant difference after compared to before.

If we compare the difference between the experiments and symptom groups, that Exp 4 stands out as lower than the other. We found a significant difference based on Tukey HSD between Exp 1 and Exp 4 ($p = 0.00011$) and Exp 5 ($p = 0.026$) for the Oculomotor symptom. For Disorientation there were significant differences between Exp 3 and Exp 4 ($p = 0.00011$) and Exp 5 ($p = 0.010$). Here we also found a significant difference between Exp 1
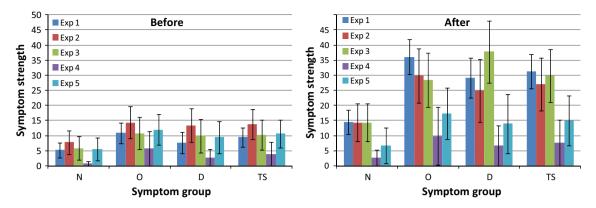
Fig. 9 The mean of each Kennedy symptom group before and after the experiments

Table 6 Viewing time of subjects having the test on the passive TV, i.e., Exp 1–Exp 3

Number of session	Number of subjects	Viewing time (min)	Group
2	1	25	Short
4	66	50	Short
6	3	75	Long
7	53	87.5	Long
8	1	90	Long

and Exp 4 ($p = 0.010$). For the Total Score, the only significant difference we found was between Exp 1 and Exp 4 ($p = 0.0013$). For Nausea no significant differences were found based on Tukey HSD.

Viewing length

In Exp 1 there was a mixture of viewing durations, but most test subjects had quite a long viewing duration. When session length was analyzed in this experiment alone no significant difference was found for longer and shorter viewing time (Brunnström et al. 2013a). The most likely explanation for that was that the group having shorter viewing duration was small (11 subjects) compared to the group with longer viewing duration (57 subjects). If we analyze Exp 1 to Exp 3 together, where we used the same 3D TV, the number of subjects having a shorter viewing time increased to 67, is shown in Table 6, where we labeled viewing durations longer than 50 min as Long and viewing durations shorter than that as Short. The overall mean (see Fig. 10) score of the group with fewer sessions was higher than before, but not as high as for the group with longer viewing time. However, also with a more even number of the two groups, it was not found that the overall means of symptoms after was significantly different from each other, based on a repeated measures ANOVA followed by Tukey HSD post hoc test ($p = 0.24$). The post hoc test revealed that the fatigue symptom was significantly higher ($p = 0.000$) for the longer sessions than for
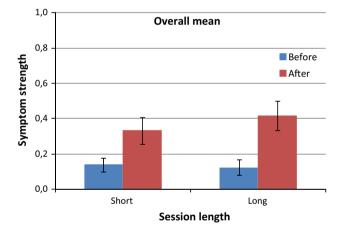

Fig. 10 The overall mean of session length was not found to be significant

the shorter, but no other individual symptom was significant.

Discussion

One aspect that is important to consider when interpreting the result in this study is that the situation for the test person is different when coming to a lab concentrated to provide scores for the main purposes of the experiments that those studies were based upon. Usually, video or movie viewing is done in a more relaxed atmosphere which

may make the symptoms less severe. However, the effect of some symptoms is clearly higher, so it is very likely that they will be similar even in a lean back situation.

Exp 1 was the largest experiment which also contained the longest viewing times. The total viewing time ranging between 30 min to about one and half hour, which is comparable to a feature length movie. From this experiment, we also see the largest effect on the symptoms, which is not surprising since it had the longest viewing time. However, we did not show in this study that overall mean of the symptoms for the longer viewing time was statistically different from the overall mean of shorter viewing time. It may be because the time difference in viewing time between the two cases was not big enough. The fatigue was significantly higher for the longer viewing time, which means that there is an effect partly but not large enough on all symptoms.

Looking at the cross-lab comparison, we can see those symptoms for 3D TV viewing were statistically significantly higher than for 2D viewing. An interesting result was received from Exp 4, where the effect of symptoms was even lower than 2D viewing (although not statistically significant) and significantly lower than the other 3D viewing experiment. This experiment was different in the sense that it was 3D using a projector system as compared to a 3D TV. The viewing distance cannot explain the difference as it was shorter than Exp 1 and almost the same as one of viewing distances of Exp 2. At this point, we cannot provide a proper explanation for the difference, however, suggesting that 3D projection system may be less demanding. Although, we could not establish an age-related effect, but the test persons in this study were dominated by younger persons, which may have affected the result.

The SSQ consists of 16 different symptoms that have been identified as important for indicating simulator sickness. When analyzing the individual symptoms it was found, mainly based on Exp 1 that Fatigue, Eye-strain, Difficulty Focusing and Difficulty Concentrating were significantly worse after the viewing compared to before, regardless whether the test used a parametric or nonparametric model. However, increased Salivation, Nausea, Vertigo, Stomach Awareness and Burping were not significant in any of the applied tests. There was no-one that reported any symptoms as Severe, but several that said that they had Strong symptoms. However, about 40% have not indicated more than Slight symptom on any question, which would suggest that a large population is largely unaffected by viewing 3D TV.

The SSQ analysis was done according to the model proposed by Kennedy et al. (1993), which classifies the symptoms into groups relating to Nausea, Oculomotor, and Disorientation. We found that the scores were significantly higher after the sessions compared to before the test, with the biggest impact on the Oculomotor system.

There was no significant effect of the gender or age found on the scores. Both of these cases would most likely need a much larger test population for showing any effect since the differences are small.

We measured the stereo acuity for all participating subjects with a Randot test. Although significant effects were found on the Oculomotor system for mid-range of stereo acuity, i.e., 20 ($p = 0.00006$), 30 ($p = 0.00006$), 40 ($p = 0.002$) and 50 ($p = 0.00006$), with a Tukey HSD post hoc test. Although, we cannot draw any strong conclusions from this since there were too few test subjects having very good stereo acuity and very poor.

The task itself may have induced the fatigue, and this was also pointed out by Kennedy et al. (1993) and from this analysis we cannot deduce exactly the cause of it.

Screening has been performed based on the scaling data according to standardized procedures of pre- and post-screening. We did not screen based on the SSQ-data. It is very hard to judge, whether someone claims they have a symptom and in fact do not. Several people have reported no symptoms before and after, but it is again very hard to judge if this is because they did not care so much about the questionnaire or just did not feel any symptoms. We have taken the position that if the test subjects have performed their tasks seriously enough otherwise, we do not have any reason to believe that the test subjects did not fill in their SSQ in a serious way.

Conclusion

In this article, we have presented that we administered the Simulator Sickness Questionnaires during a series of 3D subjective video quality tests. The purpose was to get an indication of the overall effects of symptoms that 3D TV viewing can induce. We collected the SSQ data in five different subjective experiments, from the test subjects, before and after the experiment. We performed three of the experiments on the same 3D TV, one on a 3D projector and one 2D experiment for comparison. We observed that 3D TV has a negative effect on some symptoms in the questionnaire; however, the results also indicate that the 3D video presented through a projection system does not have the same effect.

We did not find a significant overall effect by splitting the data in longer vs. shorter viewing time, although there was an individual symptom, Fatigue, which was significant. A larger difference between the longer and shorter viewing time may give a different result.

The individual symptoms Fatigue, Eye-strain, Difficulty Focusing and Difficulty Concentrating, had significantly

higher severity after than before. However, increased Salivation, Nausea, Vertigo, Stomach Awareness and Burping were not significant. The test subjects did not indicate any severe symptoms although some reported strong symptom. Many were also totally unaffected.

Based on the analysis suggested by Kennedy et al. (1993), it was shown that the biggest impact is on the Oculomotor system.

All in all this investigation shows a statistically significant increase in symptoms after viewing 3D video especially related to visual or Oculomotor system. However, we find that for most people stereoscopic 3D TV, especially when projected, has a very low impact on the experienced symptoms.

This work gives just one piece in our overall understanding of Quality of Experience in general and stereoscopic 3D TV QoE in particular. We are happy to share our data and collaborate with any researcher getting in contact with us, since we know that collecting data is both time consuming and expensive.

References

Baroncini V (2012) Design and logistics in formal subjective test the MPEG Case. In: Proc of Radioelektronika (RADIOELEKTRONIKA), 2012 22nd international conference, IEEE Explore

Brunnström K, Wang K, Andrén B (2013a) Simulator sickness analysis of 3D video viewing on passive 3DTV. SPIE, Bellingham

Brunnström K, Ananth IV, Hedberg C, Wang K, Andrén B, Barkowsky M (2013) Comparison between different rating scales for 3D TV. In: Proc of SID display week 2013, May 21–24, 2013, paper 36.4. Society of Information Displays, Vanvouver, Canada

Häkkinen JP, Vuori T, Paakka M (2002) Postural stability and sickness symptoms after HMD use. In: Proc of IEEE international conference on systems, man and cybernetics, pp 147–152

ICDM (2012) Information Display Measurements Standard (IDMS) (1 (Version 1.03c)). International Committee for Display Metrology (ICDM), Society for Information Display (SID). www.icdm-sid.org/. Accessed 2 Dec 2016

ITU-R (2012) Methodology for the subjective assessment of the quality of television pictures (ITU-R Rec. BT. 500-13). International Telecommunication Union, Radiocommunication Sector

ITU-T (1999) Subjective video quality assessment methods for multimedia applications (ITU-T Rec. P. 910). International Telecommunication Union, Telecommunication standardization sector

ITU-T (2014) Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment (ITU-T Rec. P.913). International Telecommunication Union, Telecommunication standardization sector

Jonsson J, Brunnström K (2007) Getting started with ArcVQWin (acr022250). Acreo AB, Kista

Jumisko-Pyykkö S, Utriainen T, Strohmeier D, Boev A, Kunze K (2010) Simulator sickness—five experiments using autostereoscopic mid-sized or small mobile screens

Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG (1993) Simulator sickness questionnaire: an enhanced method of quantifying simulator sickness. Int J Aviat Psychol 3(3):203–220

Kulyk V, Tavakoli S, Folkesson M, Brunnström K, Wang K, Garcia N (2013) 3D video quality assassment with multi-scale subjective method. In: Proc of fifth international workshop on quality of multimedia experience, QoMEX 2013, paper 60, IEEE Xplore Klagenfurt am Wörthersee, Austria

Lambooij M, Fortuin M, IJsselsteijn WA, Evans B, Heynderickx I (2010) Measuring visual fatigue and visual discomfort associated with 3-D displays. J SID 18(11)931–943

Meesters LMJ, IJsselsteijn WA, Seuntiëns PJH (2004) A survey of perceptual evaluations and requirements of three-dimensional TV. IEEE Trans Circuits Syst Video Technol 14(3):381–390

MPEG (2011) Call for Proposals on 3D Video Coding Technology (N12036). Moving Pictures Experts Group (MPEG), International Organisation for Standardisation, ISO/IEC JTC1/SC29/WG11, Coding of Moving Pictures and Audio

Naqvi SAA, Badruddin N, Malik AS, Hazabbah W, Abdullah B (2013) Does 3D produce more symptoms of visually induced motion sickness? In:Proc of 35th annual international conference of the IEEE EMBS. Osaka, Japan, pp 6405–6408

Perkis A, You J, Xing L, Ebrahimi T, de Simone F, Rerabek M, Nasipoulos P, Mai Z, Pourazad MT, Brunnström K, Wang K, Andrén B (2012) Towards certification of 3D video quality assessment. 2012. Scottsdale, AZ, USA

Takada H, Matsuura Y (2013) Comparison of form in potential functions while maintaining upright postures during exposure to stereoscopic video clips. In: Proc of 2013 IEEE international conference on systems, man, and cybernetics (SMC 2013). Manchester, UK, pp 2140–2145

Tavakoli S (2015) Subjective QoE analysis of HTTP adaptive streaming applications. Universidad Politecnica de Madrid, Madrid

Tavakoli S, Brunnström K, Gutiérrez J, Garcia N (2015) Quality of experience of adaptive video streaming: investigation in service parameters and subjective quality assessment methodology. Sig Process Image Commun. doi:10.1016/j.image.2015.05.001

Urvoy M, Gutiérrez J, Barkowsky M, Cousseau R, Koudota Y, Ricordel V, Le Callet P (2012) Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. In: Proc fourth international workshop on quality of multimedia experience. Yarra Valley

Urvoy M, Barkowsky M, Le Callet P (2013) How visual fatigue and discomfort impact 3D-TV quality of experience: a comprehensive review of technological, psychophysical, and psychological factors. Ann Telecommun 68(11–12):641–655

Vlad R, Nahorna O, Ladret P, Guerin A (2013) The influence of the visualization task on the simulator sickness symptoms: a comparative SSQ study on 3DTV and 3D immersive glasses. In: Proc of 3DTV vision beyond depth (3DTV-CON). Aberdeen, UK

VQEG (2010) Report on the validation of video quality models for high definition video content. Video Quality Experts Group (VQEG), www.vqeg.org. Accessed 2 Dec 2016

Wang K, Barkowsky M, Brunnström K, Sjöström M, Cousseau R, Le Callet P (2012) Perceived 3D TV transmission quality assessment: multi-laboratory results using Absolute Category Rating on Quality of Experience scale. IEEE Trans Broadcast 58(4):544–557. doi:10.1109/TBC.2012.2191031

Wang K, Hussain M, Andrén B, Brunnström K, Osterman J (2014) Perception and annoyance of crosstalk in stereoscopic 3D projector systems. IS&T/SPIE, Bellingham

Wibirama S, Hamamoto K (2014) Investigation of visually induced motion sickness in dynamic 3D contents based on subjective judgment, heart rate variability, and depth gaze behavior. In: Proc of 36th IEEE annual international conference of Engineering in Medicine and Biology Society (EMBC), pp 4803–4806