Signal Processing: Image Communication 89 (2020) 116005

Contents lists available at ScienceDirect

Signal Processing: Image Communication

journal homepage: www.elsevier.com/locate/image

Latency impact on Quality of Experience in a virtual reality simulator for remote control of machines

Kjell Brunnström [a,b,*], Elijs Dima [b], Tahir Qureshi [c], Mathias Johanson [d], Mattias Andersson [b], Mårten Sjöström [b]

[a] RISE Research Institutes of Sweden AB, Kista, Sweden
[b] Mid Sweden University, Sundsvall, Sweden
[c] HIAB AB, Hudiksvall, Sweden
[d] Alkit Communications AB, Mölndal, Sweden

ARTICLE INFO

Keywords:
Quality of Experience (QoE)
Virtual reality
Latency
Head-Mounted Displays (HMD)
Forestry crane

ABSTRACT

In this article, we have investigated a VR simulator of a forestry crane used for loading logs onto a truck. We have mainly studied the Quality of Experience (QoE) aspects that may be relevant for task completion, and whether there are any discomfort related symptoms experienced during the task execution. QoE experiments were designed to capture the general subjective experience of using the simulator, and to study task performance. The focus was to study the effects of latency on the subjective experience, with regards to delays in the crane control interface. Subjective studies were performed with controlled delays added to the display update and hand controller (joystick) signals. The added delays ranged from 0 to 30 ms for the display update, and from 0 to 800 ms for the hand controller. We found a strong effect on latency in the display update and a significant negative effect for 800 ms added delay on latency in the hand controller (in total approx. 880 ms latency including the system delay). The Simulator Sickness Questionnaire (SSQ) gave significantly higher scores after the experiment compared to before the experiment, but a majority of the participants reported experiencing only minor symptoms. Some test subjects ceased the test before finishing due to their symptoms, particularly due to the added latency in the display update.

1. Introduction

Virtual and Augmented reality (VR, AR) are emerging technologies for assisting or solving real world industrial problems. In this case we are considering immersive techniques, where the user is visually interacting with the physical environment using Head-Mounted Displays (HMD), also popularly denoted as “VR goggles”. Potentially, this will imply that workers will be using such goggles for extended periods of time; not only the same day, but most likely every working day for an extended period. Therefore, the quality related issues are crucial, not only because they are tied to performance and task completion, but also because they affect the well-being of the worker.

In this study, we investigate a VR simulator of a forestry crane used for loading logs onto a truck, mainly looking at Quality of Experience (QoE) [1,2] aspects that may be relevant for task completion, but also whether there are any discomfort related symptoms experienced during task execution. The target is an immersive video based system with the ambition to also become an AR system that lets the crane operator stay in the truck cabin or a remote location, while loading logs onto the truck, aided by a 270° HMD video view generated from four video cameras mounted on the crane (see Fig. 1). The benefits of this system are that the crane does not need to be equipped with an operator cabin, as well as improved safety and comfort for the operator. Connected to the development of the system, a desktop simulator has also been developed (see Fig. 2), which instead of displaying live video views, generates a virtual view using a 3D gaming engine. The VR simulator is used as an educational tool and should simulate the authentic crane system as closely as possible. The present QoE study has focused on the VR simulator, with the intention to also be a starting point for assessing the subjective experience of the AR system. Both the AR system and the VR simulator have the same crane control devices (joysticks) as the real ones used in the truck cabin, and an Oculus Rift HMD for the visual presentation.

QoE tests have been designed to capture the general subjective experience of using the simulator, and to study the task completion rate. Moreover, a specific focus has been to study the effects of latency on the subjective experience, with regards both to delays in the crane control interface as well as lag in the visual scene that is rendered in the HMD. Latency is of particular interest for two reasons: Firstly, it is a crucial design parameter for the AR system, since the processing of

[*] Corresponding author at: RISE Research Institutes of Sweden AB, Kista, Sweden.
E-mail address: kjell.brunnstrom@ri.se (K. Brunnström).

https://doi.org/10.1016/j.image.2020.116005
Received 11 December 2019; Received in revised form 21 August 2020; Accepted 9 September 2020
Available online 12 September 2020

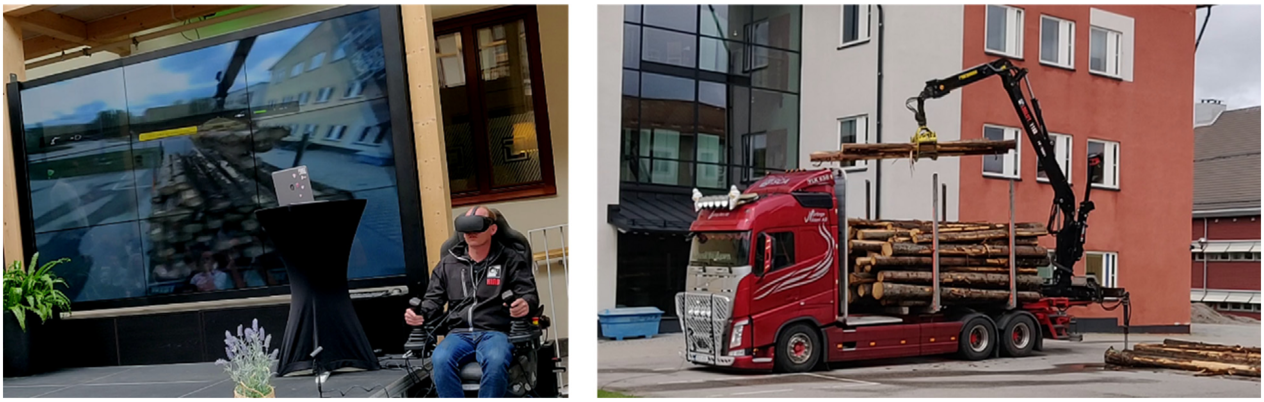


Fig. 1. Photo, provided by HIAB AB, of VR-goggle based crane operation from a remote location. Left: Operator with a VR headset and two joysticks. Operator's view is shown on the adjacent display. Right: The remotely operated crane.

video signals to generate the visual HMD scene is very CPU-consuming and the tolerable delay serves as a performance requirement for the processing hardware of the system. Secondly, we are interested in exploring the possibility of controlling a crane from a remote location, which requires the video signals, as well as the crane control signals, to be transmitted over a (typically wireless) network connection, which will introduce delays. Hence, the delay tolerance strongly influences the feasibility of such an approach. Subjective studies were performed where we have added controlled delays to the display update and hand controller (joystick) signals in the VR-simulator. The added delays ranged from 0 to 30 ms for the display update and from 0 to 800 ms for the hand controller. The selected range of delays were obtained from the literature review (see below), and is supported by pre-tests as well as the progression of the experiments where different ranges of delays have been incorporated. That is based on the experiences gained in previous studies [3,4] to cover a wider range of effective delays compared to related works.

This paper builds upon two previous papers published at the Human Vision and Electronic Imaging Conference [3,4]. It brings the results together, adds analysis not previously published, including a deeper comparison between inexperienced and experienced log lifters, and possible effects of delay inertia, learning effect and time-in-test. It also contains an extended introduction and a more comprehensive review of the background and related state of the art. The discussion and conclusions sections are also extended.

Our work is unique in the sense that the simulator provides the experience of the same real-world scenario, as the simulator is a digital clone of an actual product commercially available on the market. In addition to this, the study includes participants from both academia and industry.

2. Background

2.1. Augmented telepresence

To highlight the focus and the direction of our work we are using the term Augmented Telepresence (AT) to denote applications where high quality video-mediated communication is the enabling technology, but where additional data can be superimposed on or merged with the video as in Augmented Reality. It is not yet a commonly used term, but has been used by a few authors [5,6].

AT is similar to augmented reality in that it tries to present additional information on top of the image view as seen by the user. It primarily differs from augmented reality in that the user is present in a remote location and is observing the augmented view, but may also include the case where a two-way audio and/or audio-visual communication channel is being retained at the same time with the user seeing the augmented view.

2.2. Quality of experience

Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in light of the user's personality and current state, as defined by EU Cost Action 1003 Qualinet [2] and standardized by the International Telecommunication Union (ITU) [1]. A comprehensive overview of the field can be found in the recent QoE book by Möller and Raake [7].

The above definition of QoE, which is also pointed out by Möller and Raake [7], goes beyond the traditional QoE and Quality of Service (QoS) research and makes a clear overlap with the User Experience (UX) research tradition. These two fields originate from two different technoscientific communities, i.e. Telecommunications and Human Computer Interaction (interaction design) respectively. The QoE community is still in the process of embracing some of the more user-centric and UX-like methods.

Traditionally, in QoE research, the methods to gain insight into the delivered quality of a service and the users' experience of it have been conducted through controlled laboratory experiments, where the opinions of multiple panels of users have been collected. The results are typically reported as Mean Opinion Scores (MOS). These methods are usually referred to as subjective quality assessment methods and there are standardized ways of conducting them, e.g. for visual quality ITU-R Rec. BT.500-13 [8] or ITU-T Rec. P.910 [9]. They have been criticized for not providing enough ecological validity [10]. Improvements have been done for example in ITU-T Rec. P.913 [11]. Investigations into 3D video quality a few years ago, when the 3D TV hype was the most intense, resulted in new Recommendations from the ITU [12–14]. It was discovered that if care was not taken, several users experienced issues such as discomfort, and visual fatigue may occur. The Recommendations give some guidance on how to minimize these. An attempt to build an experimental framework for QoE of AR was made by Puig et al. [15] who advocate a combination of subjective assessment (e.g. questionnaires, subjective ratings) and objective measurements (e.g. task completion time, error rates). They only presented the results from a pilot study, so it still needs to be experimentally confirmed whether the framework gives scientifically reproducible results and if it can be extended to AT.

Now we are in the early stages of large-scale deployment of fully immersive environments, supported by advances in HMD technology, e.g. Oculus Rift, PS4 VR, and HTC Vive. Furthermore, the deployment of 5G mobile telecommunications infrastructure will support higher bandwidth, and, perhaps even more importantly, lower latency communication. This means that we are now facing low latency and distributed immersive environments on a large scale, meaning that it is of utmost importance to understand the user experience issues connected to it. New types of interactions, especially those of a highly immersive nature, will put new demands on the correct way of designing the user environment. Therefore, increased efforts should be allocated to understanding the QoE, so that the new technology is not inducing negative perceived user experiences, discomfort or even simulator sickness. Furthermore, low latency can enable services and applications with an intensive interaction component, such as gaming or remote control of professional equipment [16], which will increase the cognitive load on the user. Although research in this field has been ongoing for some time, the rapid technological development and increasing availability of immersive low latency user environments make additional research efforts necessary and more relevant than ever before.

For the QoE of the remote control of machinery, there are certain aspects that set it apart from traditional multimedia QoE. Not only is it multimodal [17], interactive [18] and task based [19], but it is also, in contrast to gaming for example, mostly for professional use and not for recreation. The multimodal aspect most often involves visual, auditory and haptic experiences, but also other senses of the body could be important, e.g. feeling the movement of the machine when lifting something heavy. The interactivity aspect relates to the fact that the user needs to react and adapt based on what is happening in the remote location, and the task-based nature of the interaction reflects that the operation of the machine has a certain purpose, and hence the ease or difficulty to perform this task will have a great impact on the QoE. The professional aspect is also very important as it can be highly repetitive and sustained for many hours per day, and even every day for the worker, which means that even relatively minor negative effects (e.g. visual discomfort) the system has on the user, could have a severe impact over time. However, based on the review of temporal aspects of simulator sickness by Dużmańska et al. [20], there are studies indicating that users may adapt after long-term exposure to immersive environments, so that initial discomfort or immersive environment sickness (a.k.a. cyber sickness or simulator sickness) is reduced after some adaptation time . This would be similar in principle, to what happens to sailors working at sea.

2.3. QoE and delay aspects of VR simulators

This section presents some related work that deals with measuring quality of experience of VR-simulators in different perspectives and involving visual and/or haptic delay.

Debattista et al. [21] presents a subjective evaluation of high-fidelity virtual environments for driving simulations. The evaluation is based on providing 44 participants access to a purpose-built virtual environment with graphics quality settings of low, medium and high. The study concludes that graphics quality affects the perceived fidelity of visual and overall experience. However, the study was limited to only judging graphics quality in three fixed states and the authors acknowledge the complexity of the visual simulator.

Ni et al. [22] designed an excavator simulator of a virtual environment for training of human operators and evaluating control strategies. The paper mostly covers the algorithms for producing terrain deformation and predicting excavation forces, but results from a user satisfaction investigation is also presented.

Strazdins et al. [23] studied virtual reality in the context of gesture recognition for deck operation training. Since available simulators supported only keyboards and joysticks as input devices, the authors developed a prototype gesture recognition system and performed their study on 15 subjects. The study concluded that improving video quality affects the user experience positively; better quality improved scores from 3.63 to 4.83 on a scale from 1 to 5. However, participants' self-assessment scores, measuring how well they performed, were on average only 3.7 on a 5-point scale. It is worth mentioning that the study was performed on students with no actual crane operation experience.

Suznjevic et al. [24] compared the QoE of two different VR-goggle technologies, i.e. Oculus Rift and HTC Vive, in a pick-and-place task. They found a slight advantage for the HTC Vive.

Jay et al. [25] studied delays in haptic and visual feedback in collaborative virtual environments, in the range of 25 to 400 ms of added delay. They found that the latency in visual feedback had a strong influence on the haptic task performance. They studied the effect on a task requiring continuous haptic and visual exchange between participants to acquire a target.

Jay and Hubbold [26] investigated whether visual and/or haptic delay influenced task performance in reciprocal tapping tasks. They found that the haptic delay had low influence, but the visual delay and combined delay had a considerable impact. Here the range of studied added delay was up to 200 ms.

Knörlien et al. [27] studied the influence of visual and haptic delay on stiffness perception in AR. They found that haptic delay decreased stiffness perception whereas visual delay increased it.

Qian et al. [28] investigated the network latency impact on the haptic QoE for remote control of robot arms. A robot arm was remotely controlled using a haptic interface to push a rod onto a ball for sensing its softness. Latency in the range of 0 to 400 ms was added in the experiment. The purpose was to find methods to stabilize the control, and not to study the impact of latency as such.

Desai et al. [29] investigated the QoE of interactive 3D tele-immersion and found that other factors were equally or even more important for the QoE than the video quality, such as better immersion and realistic interactions.

Tatematsu et al. [30] studied the influences of network latency on the QoE for haptic media, sound and video transmission. They showed that not only is absolute latency important, but as the network jitter increases, the QoE will also decrease. They also demonstrated that QoE could be estimated based on the QoS parameters.

2.4. Conclusions on current knowledge gaps

The main and foremost lack in previous research, is closeness to the target use case i.e. crane operation from a remote location. There are studies with a similar range of delays [24,25,27] but as they are not applied to the target use case, it is not clear that the effect is similar. Furthermore, there are no formalized guidelines for these types of experiments and it is therefore necessary to investigate which experimental practices are useful or not.

3. Method

Three formal subjective studies have been performed: one with the VR-system as it is; one where we have added controlled delay to the screen update and to the joystick signals, and one with added controlled delay strictly to the joystick signals only. The first has been named the Baseline Experiment (BEXP), the second the Display and Joystick Delay Experiment (DJEXP) and the third the Joystick Delay Experiment (JEXP). In addition, a few complementary smaller experiments were performed, that will also be reported here.

3.1. Common procedures for the formal tests

Test subjects were invited to perform a log-loading task in the VR simulator. They were asked to read the instructions, which explained the required task and also gave a description on how to operate the crane in the simulator. As the test subjects were not required to have any previous experience in real truck crane operation, the instructions on how to move the crane with the two joysticks, see Fig. 2, were somewhat lengthy, but all participants did understand this quickly when trying in the training session.

In the instructions the following was pointed out:

"For some people, an immersive simulator may give some discomfort or nausea. If you want to stop and not finish the test you can do it at any time without giving a reason. All the data that are gathered during the test will be treated and analyzed strictly anonymously. We do not

Table 1
The symptoms in the Simulator Sickness Symptoms by Kennedy et al. [31].

Simulator Sickness Symptoms	
1.	General discomfort
2.	Fatigue
3.	Headache
4.	Eye strain
5.	Difficulty focusing
6.	Increased salivation
7.	Sweating
8.	Nausea
9.	Difficulty concentrating
10.	Fullness of head
11.	Blurred vision
12.	Dizziness (with eyes open)
13.	Dizziness (with eyes closed)
14.	Vertigo
15.	Stomach awareness
16.	Burping

keep record on who is participating in the test that can be connected to the data''.

Consent from the test subjects was obtained prior to conducting the experiments.

The test subjects were then asked to fill in a questionnaire with a few general questions about their experience in operating truck cranes and in using VR.

A Simulator Sickness Questionnaire (SSQ) [31,32] was administered. The questionnaire contained 16 symptoms, listed in Table 1, that were identified by Kennedy et al. [31] as relevant for indicating simulator sickness.

For each of the symptoms there are four possible levels of response: None, Slight, Moderate and Severe. At the beginning of the experiment, the test subjects were asked to put on the HMD and adjust the sharpness of the image as necessary. Then the training session started. The task for the training session was to load two logs onto the truck. If something was still unclear, the test subjects were allowed to ask, and the test leader attempted to answer to make sure that the task and operation of the crane were clear. After the training, the main test phases took place.

After the main tests, a post-questionnaire was filled in by the test subjects, covering questions on their impression of the system and then the SSQ was applied once more.

The aforementioned procedures were common for all the formal experiments included in this study. The unique specifics of each experiment (BEXP, DJEXP, JEXP) are further discussed in the respective following Sections 3.3–3.5.

3.2. Apparatus

The simulator is designed for training new customers and performing user experience studies related to the actual product. The simulator includes VR goggles (Oculus Rift) which provide stitched stereo camera views, joysticks for controlling the crane, and a simulation environment of lifting logs onto the truck. The computer used is a VR-ready ASUS ROG Strix GL702VM GTX 1060 Core i7 16 GB 256 GB SSD 17.3′′. The simulation software environment was built in Unity 2017.3. The input signals from the Joysticks are converted by a special purpose interface card to give gamepad signals over USB. It was estimated from the simulator software developer that the delays in the baseline system were about 25 ms in the screen update from the movement of the head to rendering and about 80 ms from movement of Joysticks to visual feedback on the screen.

3.3. Baseline experiment

3.3.1. Procedure

For the baseline experiment (BEXP), although no specific demands on any specific visual ability were specified before the study, the test subjects' vision was investigated and noted down, by performing a Snellen visual acuity test, a 14-chart Ishihara color blind test and a Randot stereo acuity test. The dominant eye was also investigated and noted down.

The main task consisted of loading two piles of logs with 16 logs each onto the truck. When one pile was completed, the test subjects had a short break, and the test leader noted down the task completion time and restarted the program. This task took about 15 min for one pile of logs.

After the main task was completed, the experience was investigated by letting the test subject indicate their responses on rating scales, shown in Fig. 3 as printed on the paper score sheet given to the participants. The scales have been constructed so that the apparent distance between the different levels is equal, for fulfilling interval scale properties and enabling quantitative parametric analysis of the scores.

3.3.2. Lab conditions and test subjects

The Baseline experiment was conducted at RISE Research Institutes of Sweden AB, Kista, Sweden, within a lab room specifically for subjective experiments. The room was kept quiet from disturbing noises and at a comfortable temperature. Note that lighting conditions were irrelevant, due to the use of HMD.

3.3.3. Test subjects

Eighteen test subjects internally recruited from RISE completed the test, 12 males and 6 females, with a mean age of 40. The youngest participants were 23 and the oldest 58. All but two test persons had normal or corrected to normal visual acuity. These two had slightly lower visual acuity but could still perform the task without any problems. Two other participants were either color-blind or had a reduced ability to perceive colors. The 3D acuity varied between 4 and 10, with a mean acuity of 8.

3.3.4. Analysis

The MOS was calculated as outlined in Section 3.6.1. There were no statistical tests performed between the MOS of the used scales, as each scale represents different questions and experience categories, which are not directly comparable to each other.

The SSQ was analyzed as described in Section 3.6.2.

3.4. Display and joystick delay experiment

3.4.1. Procedure

For the Display and Joystick delay experiment (DJEXP), we simplified the procedure for the visual abilities by letting the test subjects self-report their visual status.

The training session in DJEXP was conducted in the same way as in BEXP. It had the added purpose of giving the test subjects a sense of the baseline-delay case and this was pointed out in the instructions. The main task was to load logs onto the truck for about 20 min. The delay of the screen update and the Joysticks were adjusted every 2 min and the test subject was asked to give his or her quality ratings verbally after about 1.5 min (in practice it turned out that more time was needed to give the response, so almost the whole second minute was used for that). The simulator program was restarted after each completed 2 min trial. The scales used were the same as in BEXP, see Fig. 3, except that we added a scale about the experienced comfort. They were also shown as text sentences, as follows:

- How would you rate the picture quality?
- How would you rate the responsiveness of the system?
- How would you rate your ability to accomplish your task of loading the logs on the truck?
- How would you rate your comfort (as in opposite to discomfort)?
- How would you rate the immersion of the experience?
- How would you rate your overall experience?

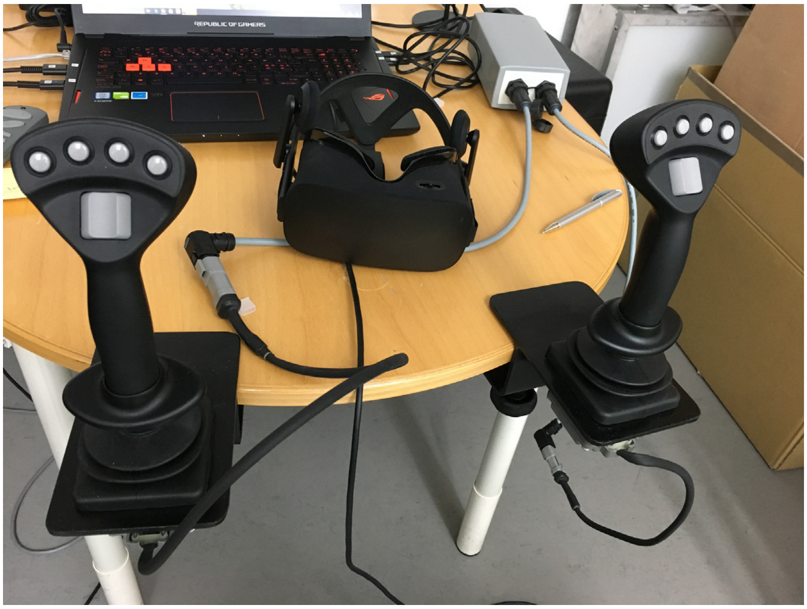


Fig. 2. The two joysticks for operating the crane in the VR simulator and the HMD (Oculus Rift).

A graphical representation of the scale was shown after these sentences, see Fig. 4, in the instructions, in order to give the test subjects a mental picture of the scale.

When the test subject was giving their ratings verbally, they gave the response vocally with the category labels: Bad, Poor, Fair, Good and Excellent. The questions above were repeated verbally by the test leader and the score noted down.

Ten delay conditions were used (nine with added delay and one baseline-delay). These were:

- Reference condition: baseline-delay (25 ms for Display and 80 ms for Joystick)
- Display delay (ms): 5, 10, 20 and 30
- Joystick delay (ms): 10, 20, 50, 100 and 200

The order was randomized per test subject. The choice of joystick and display delays was based on related works [24,25,27] and previous studies [3,4].

3.4.2. Lab conditions

The display and joystick delay experiment was conducted at Mid Sweden University, in a reserved office room. The room was kept quiet from disturbing noises and at a comfortable temperature. Note that lighting conditions were irrelevant, due to the use of HMD.

3.4.3. Test subjects

Thirty-five test subjects participated in the test, 26 males and 9 females, with a mean age of 39, the youngest participant being 23 and the oldest 61. They were recruited from the Mid Sweden University and were a mixture of students and staff. The visual status of the test subjects was self-reported. The experiment had a mixture of participants with corrected-to-normal and uncorrected vision. None of the participants reported problems with performing the experiment task due to vision issues. Only five subjects had prior experience with crane operation prior to the test. One of those subjects had 23 years of experience and was operating such a crane every day at work. Most test participants (28) had little to no prior experience with VR or HMDs.

3.4.4. Analysis

Scale analysis and SSQ analysis was performed as described in Section 3.6. The comparisons and statistical tests between all involved conditions were performed for each scale and delay type separately. For the Display delay we have $5 \times 4/2 = 10$ comparisons and for the Joystick delay $6 \times 5/2 = 15$ comparisons.

3.5. Joystick delay experiment

3.5.1. Procedure

As in DJEXP, in the Joystick delay experiment (JEXP) the training session had the additional purpose of giving the test subjects a sense of the no-delay case and this was made clear in the instructions. The main task was to load logs onto the truck 6 times in 2 min periods. The joystick delay was adjusted every 2 min, and the test subject was asked to give his or her quality ratings verbally, as was done also in DJEXP. In contrast to DJEXP, the answers given in JEXP were after each 2 min session when the test subjects had been asked to stop. The simulator program was restarted after that. The questions asked and rating scale used were as shown below:

- How many logs did you load these two minutes?
- How would you rate the responsiveness of the system?
- How would you rate your ability to accomplish your task of loading the logs on the truck?
- How would you rate your comfort (as in opposite to discomfort)?
- How would you rate the immersion of the experience?
- How would you rate your overall experience?

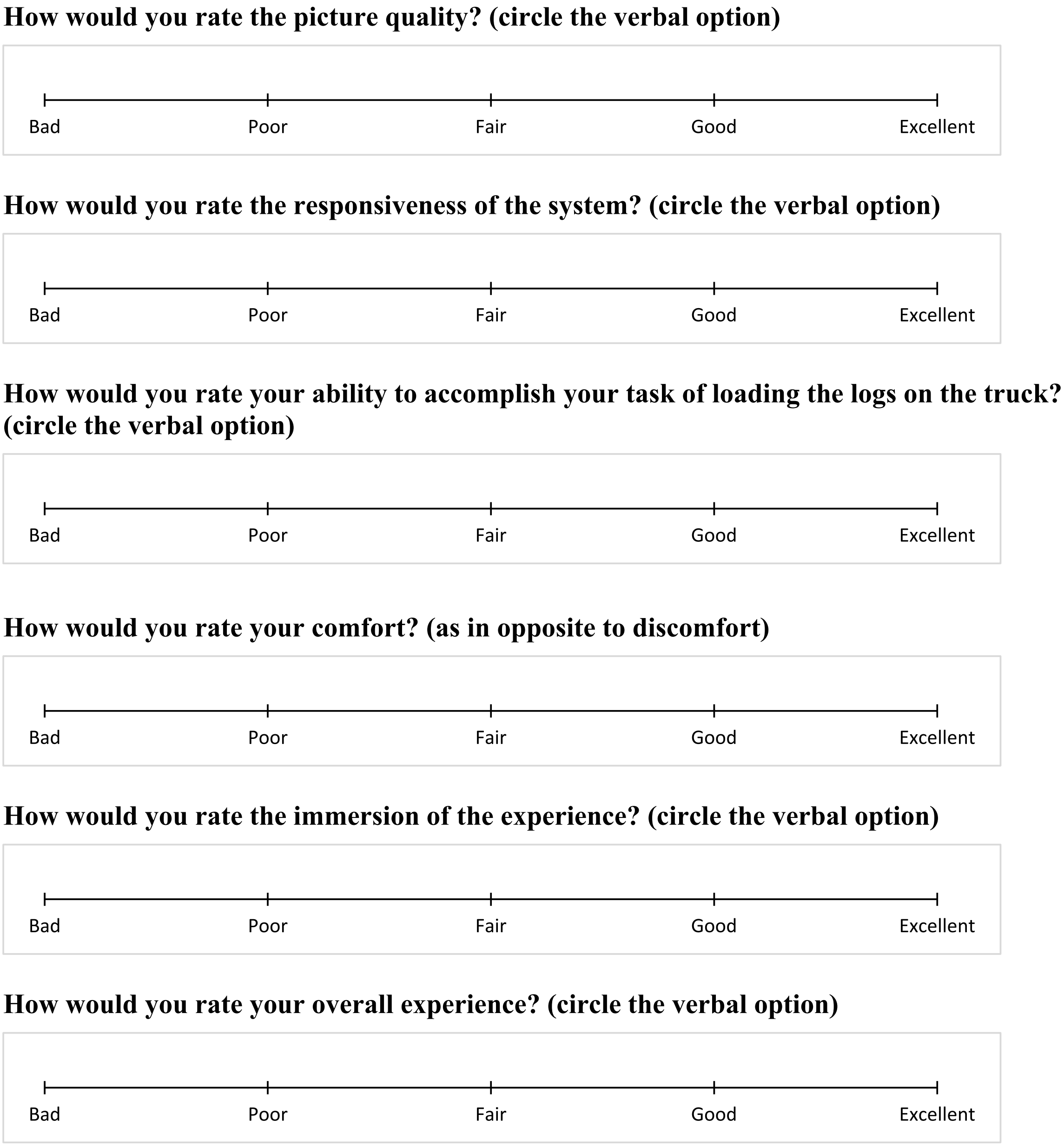
Like in DJEXP, a graphical representation of the scale was shown after these sentences, see Fig. 4, in the instructions, to give the test subjects a mental picture of the scale.

When the test subjects were giving their ratings verbally, they gave the response with the category labels: Bad, Poor, Fair, Good and Excellent. The questions above were repeated verbally by the test leader and the score noted down.

Six delay conditions were used (five with added delay and one no-delay, applied on top of the inherent operational delay within the system). These were:

- Reference condition: no delay
- Joystick delay (ms): 50, 100, 200, 400 and 800

The order was randomized per test subject. The choice of the delay used was based on other studies [24,25,27] as well as own pre-studies [3, 4] and the preliminary findings from DJEXP, which motivated the addition of higher delay levels.



How would you rate the picture quality? (circle the verbal option)

Bad Poor Fair Good Excellent

How would you rate the responsiveness of the system? (circle the verbal option)

Bad Poor Fair Good Excellent

How would you rate your ability to accomplish your task of loading the logs on the truck? (circle the verbal option)

Bad Poor Fair Good Excellent

How would you rate your comfort? (as in opposite to discomfort)

Bad Poor Fair Good Excellent

How would you rate the immersion of the experience? (circle the verbal option)

Bad Poor Fair Good Excellent

How would you rate your overall experience? (circle the verbal option)

Bad Poor Fair Good Excellent

Fig. 3. The rating scales used to investigate the Quality of Experience (QoE), as printed on the paper score sheet.

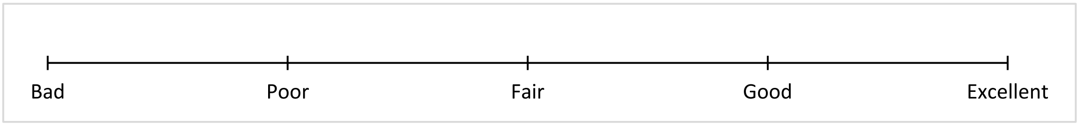


Fig. 4. Scale used in the Display and Joystick Delay experiment and in the Joystick Delay experiment, as visualized in the written instructions.

3.5.2. Lab conditions

The Joystick delay experiment was conducted at RISE's lab in Kista, Sweden. The room was kept quiet from disturbing noises, and at a comfortable temperature. Note that lighting conditions were irrelevant, due to the use of HMD.

3.5.3. Test subjects

Thirty-one test subjects participated in the test, 22 males and 9 females, with a mean age of 39 where the youngest participant was 22 and the oldest 64. They were recruited from RISE in Kista, Sweden, and made up of a variety of staff, visiting researchers and students. None of the test subjects were experienced in operating a real truck crane. Most of the subjects had little or no experience in using VR-systems (26) and some had gained their experience by participating in one of our previous studies. The visual status of the test subjects was self-reported. None of the participants reported any problems with performing the task based on vision issues. Some of the test subjects wore their glasses while using the HMD.

3.5.4. Analysis

Scale analysis and SSQ analysis performed as described in Section 3.6. The comparisons and statistical test between all involved conditions were performed for each scale separately, which gives 6 × 5/2 = 15 comparisons. In addition, for the JEXP experiment we conducted additional analysis on:

- Comparison with experienced log lifters (see Section 4.3.1)
- Effect of delay inertia (see Section 4.3.2)
- Learning effect (see Section 4.3.3)
- Time-In-Test (see Section 4.3.4)

3.6. Analysis

3.6.1. Scale analysis

The scale responses were given numerical values when analyzed using the following: Bad = 1, Poor = 2, Fair = 3, Good = 4 and Excellent = 5. The instructions displayed a graphical representation of the scales with equal distances between the categories. It was also explained in the written text within the instructions. We have therefore assumed that we can analyze the scales as interval scales. The mean opinion scores (MOS) were calculated from the scale responses of the test subjects.

First a Normality test was performed based on the method of Shapiro and Wilks [33].

We adopted the Bonferroni method [34] for compensating for multiple comparisons, as the planned number of comparisons were rather few. In this method, the considered significance level (α) is divided by the number of comparisons (n) so that the significance level for each comparison will be α/n. For 95% confidence $\alpha = 0.05$. For the Display delays, there were 10 comparisons in DJEXP and then we used $p \leq 0.05/10 = 0.005$ as the per comparison significance level. For the Joystick delays in both DJEXP and JEXP there were 15 comparisons and the significance level per comparison then became $p \leq 0.05/15 = 0.0033$.

The statistical test performed was the dependent T-test for paired samples [35]. As the result will demonstrate, the Normality will not hold in most cases (see Section 4), but the T-test has been shown to be very robust under deviations from Normality [36]. Nevertheless, to prevent reaching the wrong conclusion, the Wilcoxon signed-rank test was also performed [35], which is the non-parametric counterpart of the dependent T-test for paired samples and does not rely on the Normality assumptions.

3.6.2. SSQ analysis

The questionnaire answers were interpreted into a number, in our case by None = 0, Slight = 1, Moderate = 2, Severe = 3 for allowing parametric statistical analysis.

Kennedy et al. [7] 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. The groups are not completely separated since a few of the variables are used when calculating the scores in more than one group, e.g. nausea and difficulty concentrating. Table 2 indicates which of the symptoms are grouped together within Nausea, Oculomotor or Disorientation groups. The group scores are calculated by summing up the values with a '1' coefficient in Table 2 and multiplying that sum by the factors at the bottom of the table. Each symptom value is obtained using the conversion between severity and numerical value as described above.

After the scores for N, O, D & TS were calculated, the participant simulator sickness was calculated, taking the mean of all the participants once before, and once after the experiment.

Table 2
SSQ score calculations as described in Kennedy et al. [31].

	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] × 9.54
O = [2] × 7.58
D = [3] × 13.92
TS = ([1] + [2] + [3]) × 3.74.

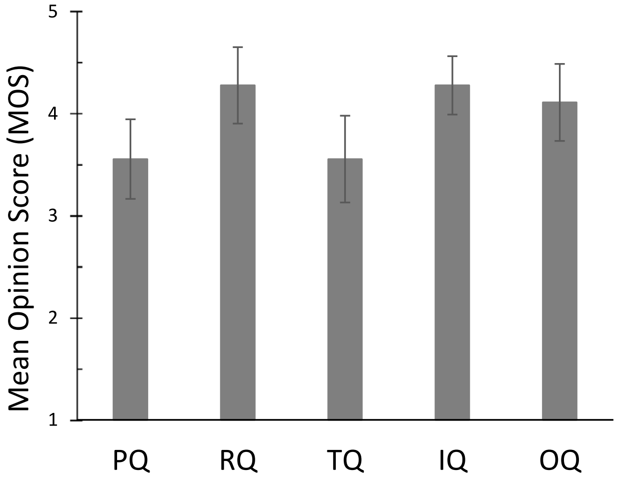


Fig. 5. The Mean Opinion Scores (MOS) for the baseline experiment. From the left along the x-axis the Picture Quality (PQ), Responsiveness Quality (RQ), Task Accomplishment Quality (TQ), Immersive Quality (IQ) and Overall Quality (OQ) are shown. The error bars indicate 95% confidence intervals.

The number of interesting comparisons performed were between each symptom group before and after, four comparisons in total. This gives $\alpha = 0.05$ $p \leq 0.0125$ as the significance level. Here, the statistical test was performed with a one-tailed dependent T-test for paired samples.

4. Results

The results of the three primary experiments (BEXP, DJEXP, JEXP) are presented in this section, and are based on the analysis method mainly outlined in Section 3.6.

4.1. Baseline experiment

The mean and the 95% confidence intervals of the ratings of the different scales used can be seen in Fig. 5. In Fig. 3 the full rating questions as well as the rating scales are shown, as presented to the test subjects.

The Picture Quality (PQ) was experienced as being between both Fair and Good. For the Responsiveness Quality (RQ) the scores were higher, and the mean resulted in being just above Good. The Task Accomplishment Quality (TQ), much like PQ, was also rated between Fair and Good. The Immersive Quality (IQ) and Overall Quality (OQ) were experienced as higher than Good.

The mean task completion time was 26.5 min, with a standard deviation of 8.7 min.

The SSQ showed only a minor increase in the symptom strength, see Fig. 6. However, the statistical test shows significant increase for Disorientation, as p = 0.004 < 0.01. The other symptom groups were not as significant, with Nausea having p = 0.03, Oculomotor having p = 0.17 and the Total score p = 0.02. Most test subjects reported only slight symptoms if any, and only one participant reported experiencing a moderate symptom. One interesting case featuring an especially sensitive person was encountered. The person in question did, just after 2 min, report Severe discomfort, including Nausea, Vertigo and Stomach awareness, as well as Moderate Sweating and Dizziness with his or her eyes open. This person was not participating in the actual test, but tested the simulator in a demo session. It seems that there is a small fraction of very sensitive people, but the majority had no major problems with this simulator.

4.2. Display and joystick delay experiment

The analysis was performed as described in Section 3.6.1. The Normality tests rejected the hypothesis of a Normal distribution at a 95% confidence for all tested cases i.e. the performance metric as well as for all the scales and all delays.

Ten test subjects ceased the test and therefore, did not complete all of the test conditions. The reason to stop was because of encountering discomfort and nausea. In most cases, this was related to the experience of higher added Display delay conditions just before stopping i.e. added Display delay ≥ 20 ms with baseline delay ≥ 45 ms. The test leader was present during the whole test and could monitor and give feedback to the test subject to continue or not if they felt discomfort or nausea. The recommendation in most cases was to discontinue the test. The ratings detailed up to the point of stopping have been included in the analysis, and the ratings not given have been treated in the analysis as missing data. In all cases, the SSQ were filled in for these test subjects, so these scores have been included in the analysis.

The results from the rating scales are shown in Figs. 7 to 12. To the left, the MOS for different Display delays (DD) are drawn and to the right, the MOS for different Joystick delays (JD). The total delays are specified in the graphs, which is baseline delay (discussed and confirmed by the manufacturer of the simulator) plus added delay. For DD it is 25 ms + (5 ms, 10 ms, 20 ms, 30 ms) = 30 ms, 35 ms, 45 ms and 55 ms. For JD it is 80 ms + (10 ms, 20 ms, 50 ms, 100 ms, 200 ms) = 90 ms, 100 ms, 130 ms, 180 ms and 280 ms. The error bars indicate 95% confidence intervals.

In Fig. 7, the MOS of the Picture Quality is shown. There is a trend for lower PQ at higher DDs, there is no clear trend for the JDs. Unexpectedly, 20 ms added Display delay (45 ms) was rated worse than 30 ms (55 ms) added delay, but the difference could not be determined as statistically significant.

In Fig. 8, the MOS of the Responsiveness Quality is shown. There is a trend for lower RQ at higher delays, but the differences are not significant.

In Fig. 9, the MOS of the Task Accomplishment Quality is shown. No clear trend can be noticed.

In Fig. 10, the MOS of the Comfort Quality is shown. The comfort is reduced by longer delay and this trend is clearer for the Display delay. The 30 ms added Display delay (in total 55 ms) is significantly lower (T-test: p = 0.0019 < 0.005, Wilcoxon: p = 0.0033 < 0.005), than the comfort quality for baseline-delay.

In Fig. 11 the MOS of the Immersion Quality is shown. There is a trend for lower Immersion Quality at higher delays. 30 ms added delay IQ (55 ms) is very close to being significant (T-test: p = 0.0056 > 0.005 Wilcoxon: p = 0.009 > 0.005) compared to the baseline-delay case.

In Fig. 12, the MOS of the Overall Quality is shown. The OQ has a similar trend to the IQ but is not as clear. No significance was discovered.

The SSQ analysis for the delay revealed a large increase in the symptom levels (Fig. 13), all of which were statistically significant i.e. <0.0125; where Nausea had p = 0.00005, Oculomotor p = 0.007, Disorientation (p = 0.00008) and the Total Score (p = 0.0002). However, only 2 test subjects reported symptoms on a Severe level. In this analysis, all test subjects were included, even those not completing the primary session.

4.3. Joystick delay experiment

The analysis was performed as described in Section 3.6.1. The Normality tests rejected the hypothesis of a Normal distribution at a 95% confidence for all tested cases i.e. the performance metric as well as for all the scales and all delays.

Two test subjects aborted the test and did not complete all test conditions. The reason to stop was discomfort and nausea. The test leader was present during the entire test and could monitor and also give feedback to test subjects whether to continue or not if they felt discomfort or nausea. In two cases, the recommendation to participants was to stop. The ratings given up to the point of stopping have been included in the analysis. In all cases the SSQ were filled in for these test subjects, so these scores have been included in the analysis.

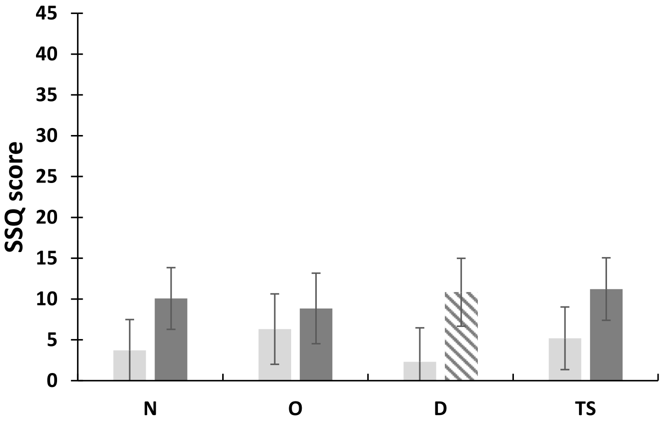


Fig. 6. Simulator Sickness Questionnaire (SSQ) scores for the baseline experiment, where the left (light gray) bars represent the symptom levels before the experiment and the right (dark gray and striped indicating statistically significant difference) bars the symptom levels after the experiment. The different symptom groups along the x-axis are: Nausea (N), Oculomotor (O), Disorientation (D) and the Total Score (TS). The error bars indicate 99% confidence intervals.

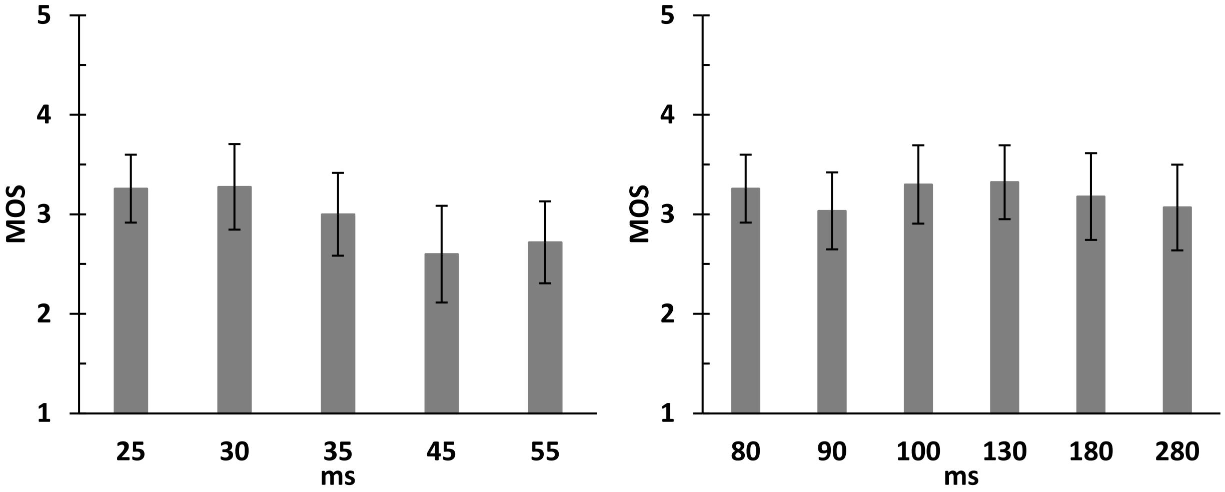


Fig. 7. The Mean Opinion Scores (MOS) for Picture Quality for different Display delays (left) and for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals.

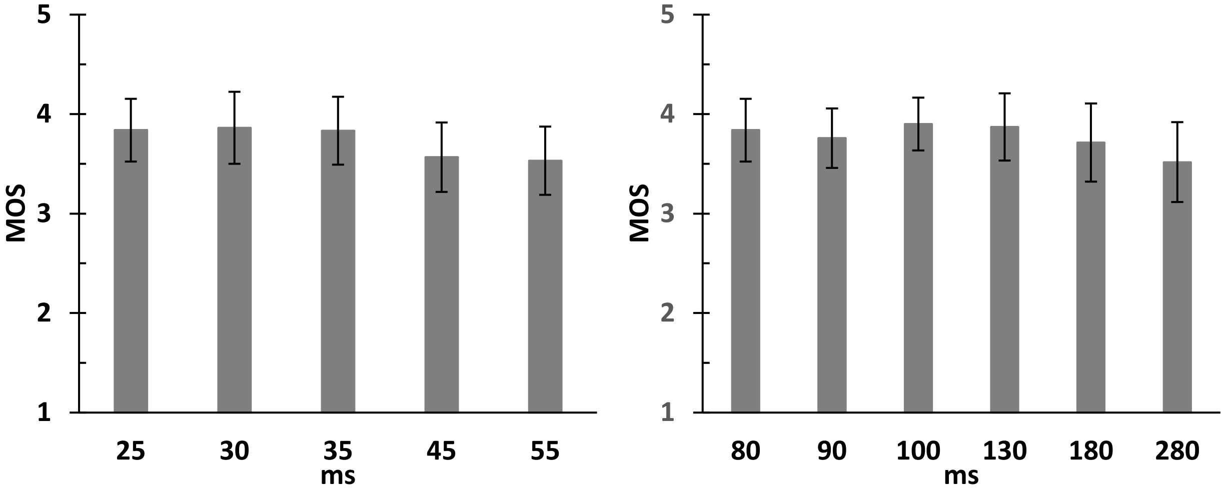


Fig. 8. The Mean Opinion Scores (MOS) for Responsiveness Quality for different Display delays (left) and for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals.

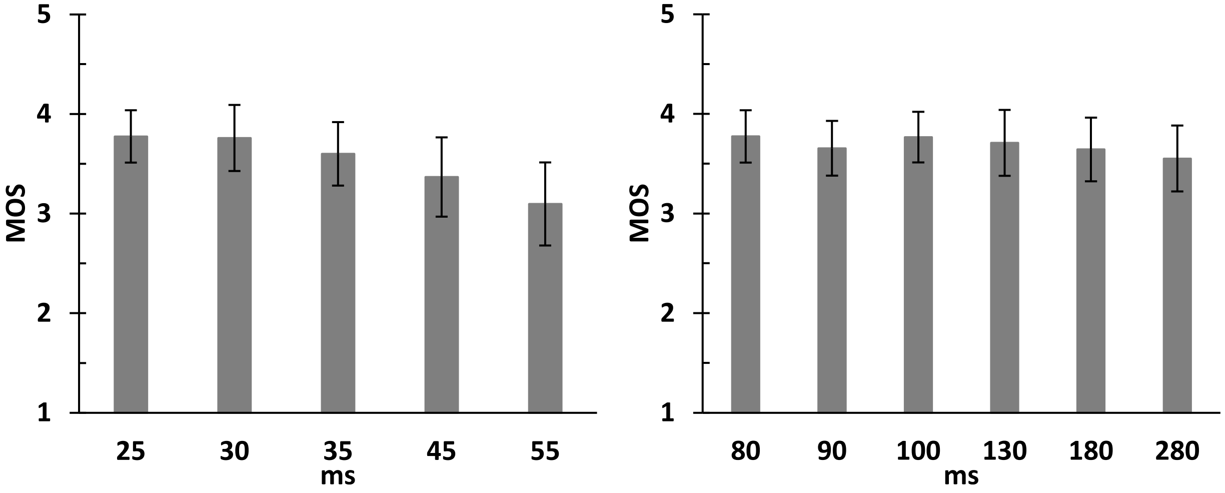


Fig. 9. The Mean Opinion Scores (MOS) for Task Accomplishment Quality for different Display delays (left) and for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals.

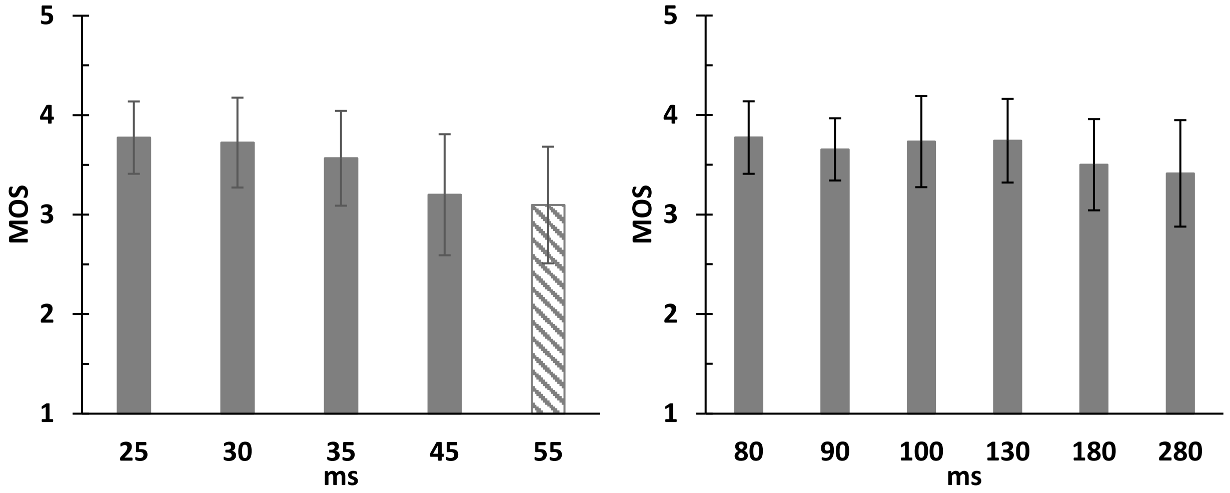


Fig. 10. The Mean Opinion Scores (MOS) for Comfort Quality for different Display delays (left) and for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals.

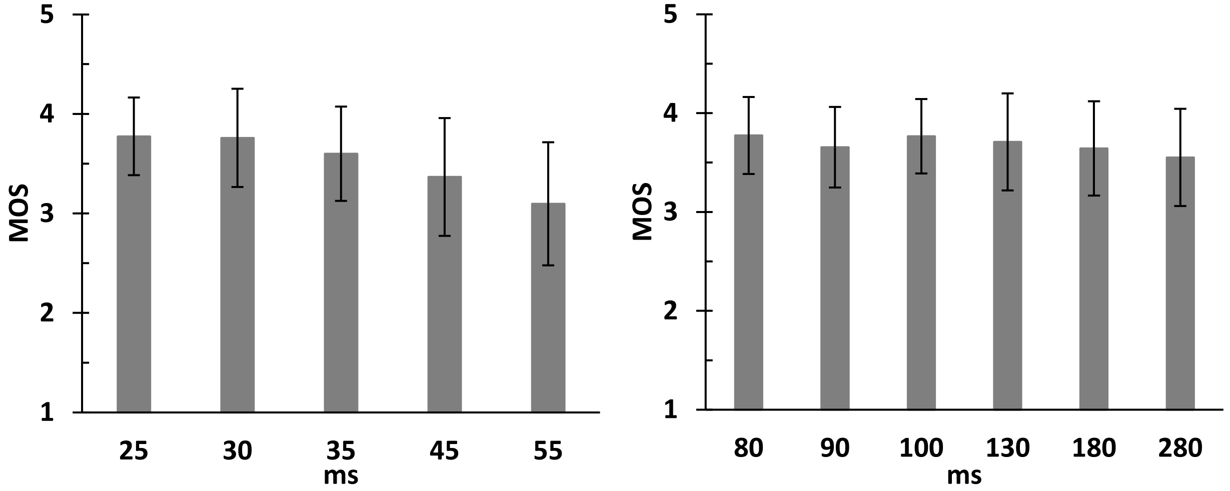


Fig. 11. The Mean Opinion Scores (MOS) for Immersion Quality for different Display delays (left) and for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals.

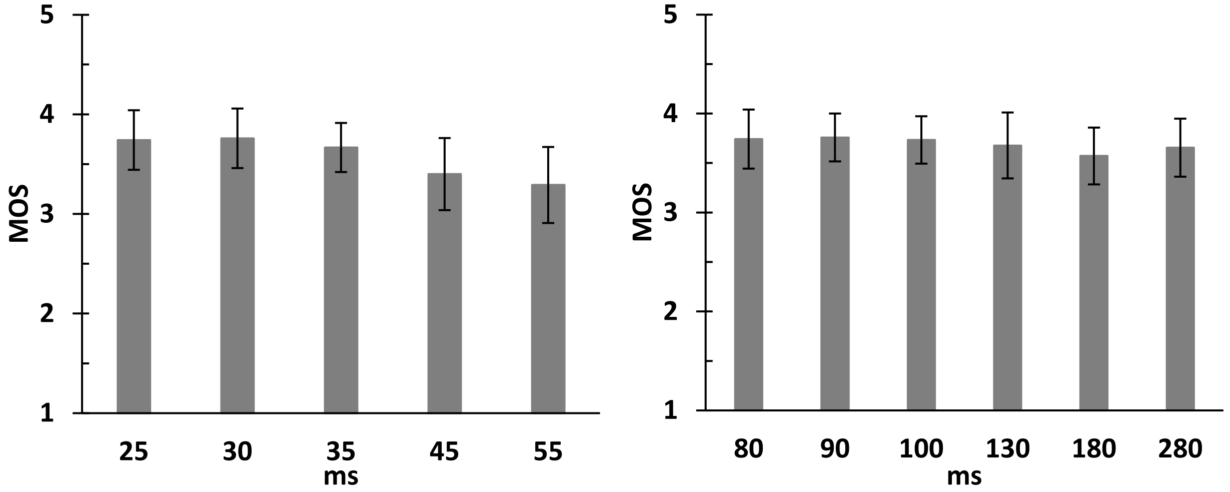


Fig. 12. The Mean Opinion Scores (MOS) for Overall Quality for different Display delays (left) and for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals.

The task performance results, in terms of the number of logs successfully loaded by test subjects, are shown in Fig. 14. The height of the bars shows the mean number and the error bars indicates 95% confidence intervals. Striped bars indicate significantly different mean amount of logs. The total delays are given in the graphs, that is baseline delay plus added delay. For Joystick delay it is 80 ms + (50 ms, 100 ms, 200 ms, 400 ms, 800 ms) = 130 ms, 180 ms, 280 ms, 480 ms and 880 ms.

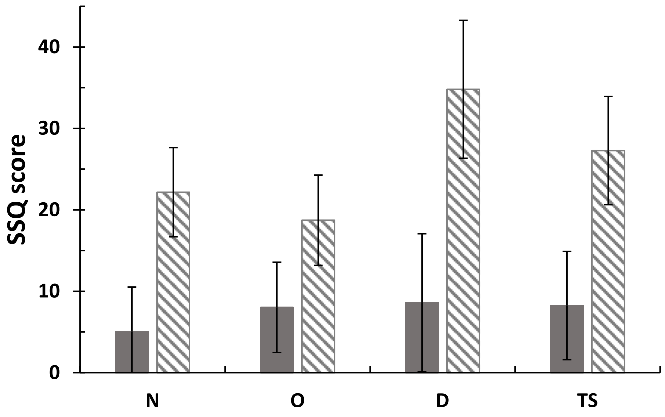


Fig. 13. Simulator Sickness Questionnaire (SSQ) scores for the delay experiment, where the left (light gray) bars represent the symptom levels before the experiment and the right (dark gray and striped indicating statistically significant difference) bars the symptom levels after the experiment. The different symptom groups along the x-axis are: Nausea (N), Oculomotor (O), Disorientation (D) and the Total Score (TS). The error bars indicate 95% confidence intervals.

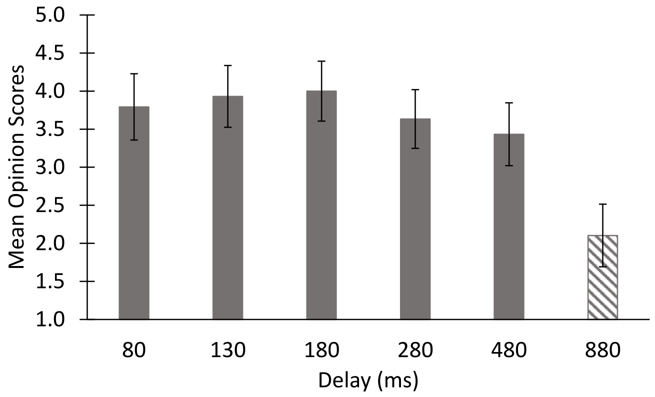


Fig. 14. The mean number of logs loaded per 2 min session as a function of the added Joystick delay in milliseconds (ms). The error bars indicate 95% confidence intervals. Striped bars marks statistically significantly different mean values.

The statistical tests showed that the 880 ms delay case was significantly different from all the other cases, with $\alpha < 0.0033$, see Table 3.

In Fig. 15, we can see the MOS of the Responsiveness Quality in the height of the bars as ordered along the x-axis with increasing added Joystick delay. Striped bars indicate significantly different MOS. The statistical tests detailed that 880 ms was statistically significantly different from the other cases, with $\alpha < 0.0033$, see Table 3. In addition, the 480 ms delay case was also significant towards 80 ms in the T-test, but not in the Wilcoxon test (T-test: p = 0.0023, Wilcoxon: p = 0.0067).

Fig. 16 illustrates the effect on the Task Accomplishment Quality. The statistical tests performed here also specified that 880 ms was statistically significantly different from the other cases with $\alpha < 0.0033$, see Table 3.

The impact of the delay on the Comfort Quality is shown in Fig. 17. In contrast to the previous cases, 880 ms was not statistically different from all other cases with (alpha) > 0.0033, but was statistically significantly different from some, as shown in Table 3. 80 ms was only significant in the T-test and 480 ms was not significant in either test.

In Fig. 18, the MOS of the Immersion Quality is shown. The Immersion has no significant cases. The only exception is that the T-test for 880 ms towards 130 ms was significant, but this was not conclusively confirmed with the Wilcoxon test, see Table 3.

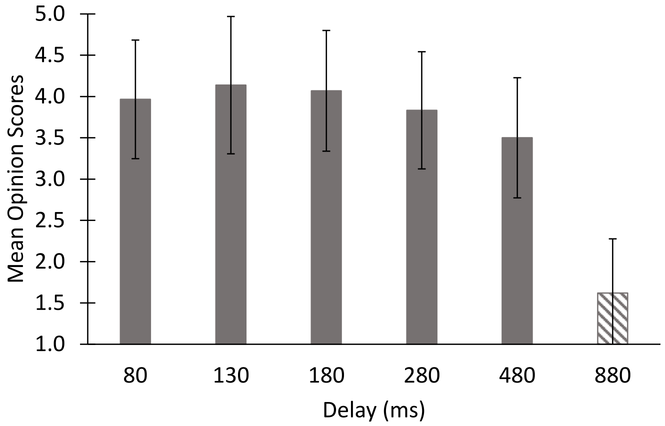


Fig. 15. The MOS for Responsiveness Quality for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals. Striped bars mark statistically significant different mean values.

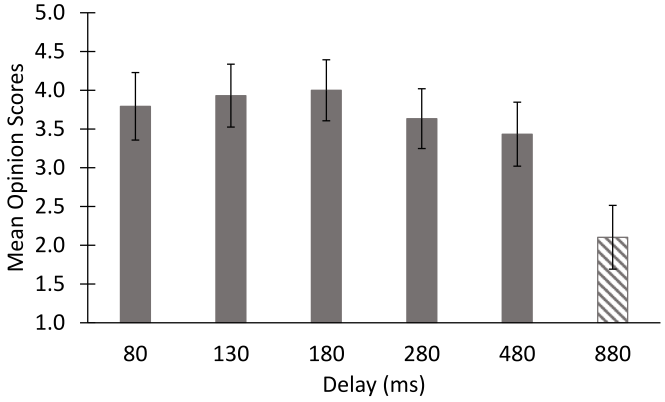


Fig. 16. The MOS for Task Accomplishment Quality for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals. Striped bars mark statistically significant different mean values.

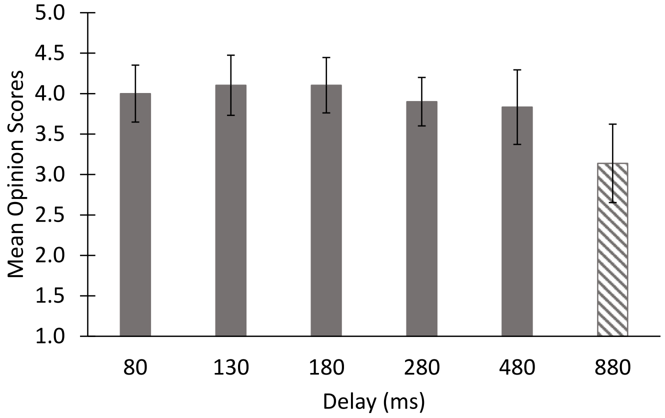


Fig. 17. The MOS for Comfort Quality for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals. Striped bars mark statistically significant different mean values.

In Fig. 19, the MOS of the Overall Quality is shown. The statistical tests gave that 880 ms was statistically significantly different from the other cases with $\alpha < 0.0033$, see Table 3.

The SSQ analysis for the delay revealed large increase in the symptom levels (Fig. 20), all of which were statistically significant i.e. p < 0.0125; where Nausea had p = 0.00024, Oculomotor p = 0.000014, Disorientation (p = 0.000022) and the Total Score (p = 0.0000063).

Table 3
P-values of statistical tests between 880 ms and the other delay cases.

Delay	80 ms		130 ms		180 ms		280 ms		480 ms	
P-value	T	W	T	W	T	W	T	W	T	W
NLOGS	0.00026*	0.00036*	0.00042*	0.00075*	$1 \cdot 10^{-5}$*	0.00016*	0.00098*	0.0024*	$5 \cdot 10^{-5}$*	0.00033*
RQ	$5 \cdot 10^{-12}$*	$6 \cdot 10^{-6}$*	$4 \cdot 10^{-11}$*	$5 \cdot 10^{-6}$*	$1 \cdot 10^{-12}$*	$4 \cdot 10^{-6}$*	$1 \cdot 10^{-12}$*	$3 \cdot 10^{-6}$*	$2 \cdot 10^{-9}$*	$1 \cdot 10^{-5}$*
TQ	$1 \cdot 10^{-6}$*	$7 \cdot 10^{-5}$*	$1 \cdot 10^{-6}$*	$6 \cdot 10^{-5}$*	$6 \cdot 10^{-10}$*	$1 \cdot 10^{-5}$*	$1 \cdot 10^{-5}$*	0.00016*	$2 \cdot 10^{-5}$*	0.00040*
CQ	0.0013*	0.0037	0.0013*	0.0022*	0.00025*	0.00095*	0.0018*	0.0030*	0.0098	0.018
IQ	0.017	0.021	0.0029*	0.0077	0.0054	0.0093	0.24	0.26	0.088	0.012
OQ	$1 \cdot 10^{-6}$*	$6 \cdot 10^{-5}$*	$3 \cdot 10^{-7}$*	$4 \cdot 10^{-5}$*	$3 \cdot 10^{-9}$*	$2 \cdot 10^{-5}$*	$5 \cdot 10^{-5}$*	$3 \cdot 10^{-5}$*	$1 \cdot 10^{-6}$*	0.00012*

*Significant cases $\alpha < 0.0033$.

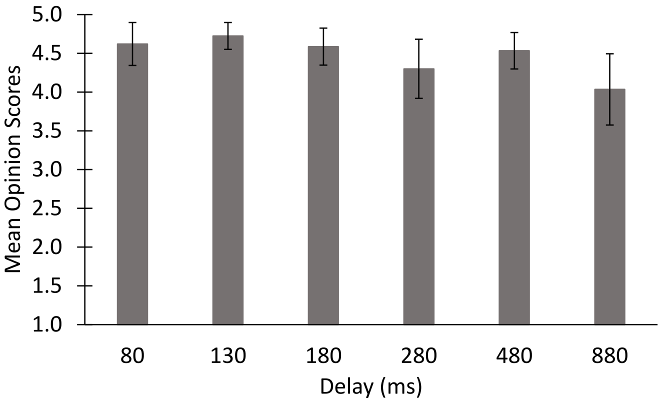


Fig. 18. The MOS for Immersion Quality for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals. Striped bars mark statistically significant different mean values.

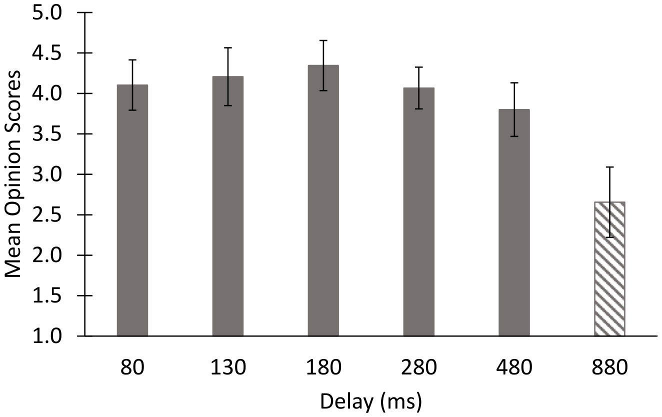


Fig. 19. The MOS for the Overall Quality for different Joystick delays (right) in milliseconds (ms). The error bars indicate 95% confidence intervals. Striped bars mark statistically significant different mean values.

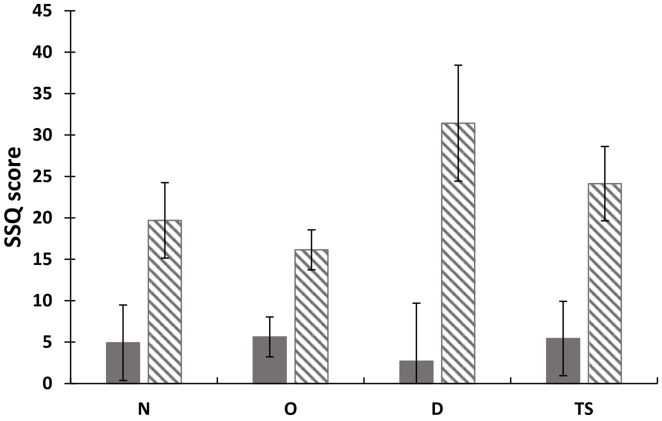


Fig. 20. Simulator Sickness Questionnaire (SSQ) scores for the delay experiment, where the left (medium gray) bars represent the symptom levels before the experiment and the right (striped bars indicating statistically significant difference) bars the symptom levels after the experiment. The different symptom groups along the x-axis are: Nausea (N), Oculomotor (O), Disorientation (D) and the Total Score (TS). The error bars indicate 95% confidence intervals.

However, one test subject reported symptoms on a Severe level (highest in the SSQ) and he/she also stopped the test. In this analysis, all of the test subjects were included, even those not finishing the main session.

4.3.1. *Comparison with experienced log lifters*

To investigate whether the results obtained are different if the test subjects have extensive experience in truck crane operation, we let seven experienced truck crane operators perform the experiment as well. We reduced the number of scales to rate after each 2 min period, in order to make the test quicker for this group, otherwise the experiment was performed as previously. The questions asked after each 2 min period were:

- How many logs did you load these two minutes?
- How would you rate the responsiveness of the system?
- How would you rate your ability to accomplish your task of loading the logs on the truck?

In Fig. 21 (left), the mean number of logs are shown, and it can be observed that the mean is distinctly higher, which can be expected. It can also be observed that the overall impact is the same, in that the results are very similar up to and including 480 ms, but then drops drastically for 880 ms. Fig. 21 (right) shows a scatterplot between the mean number of logs reached by the naïve users (x-axis) compared to the mean number of logs reached by the experts (y-axis). The Pearson linear correlation is 0.96 (i.e. $R^2 = 0.93$).

The rated experiences for the Responsiveness Quality (Fig. 22, left) and the Task Accomplishment Quality (Fig. 23, left) shows similar trends in regard to the inexperienced test subjects, with clearly lower ratings for the 880 ms case. It may be noted that for the experienced test persons, the drop for 480 ms is slightly deeper, although this is very uncertain due to the small number of experienced test subjects. Figs. 22 and 23 (right) shows scatterplots between the MOS of Responsiveness Quality and Task Accomplishment Quality of the naïve users (x-axis), compared to the MOS of the experts (y-axis). The Pearson linear correlation is 0.93 (i.e. $R^2 = 0.88$) for the Responsiveness Quality and 0.95 (i.e. $R^2 = 0.90$) for the Task Accomplishment Quality.

The SSQ results are also in line with what was obtained for the inexperienced test subjects, see Fig. 24.

4.3.2. *Effect of delay inertia*

During the subjective tests, a potential trend was noticed by the test leader, wherein participants seemed to give unexpectedly low scores to 80 ms, 130 ms delay scenarios, if such scenarios had been preceded by a high-delay scenario (880 ms). This effect may have been manifested as a kind of inertia in participant accommodation to the delays — when participants accommodated to a high delay scenario, then a

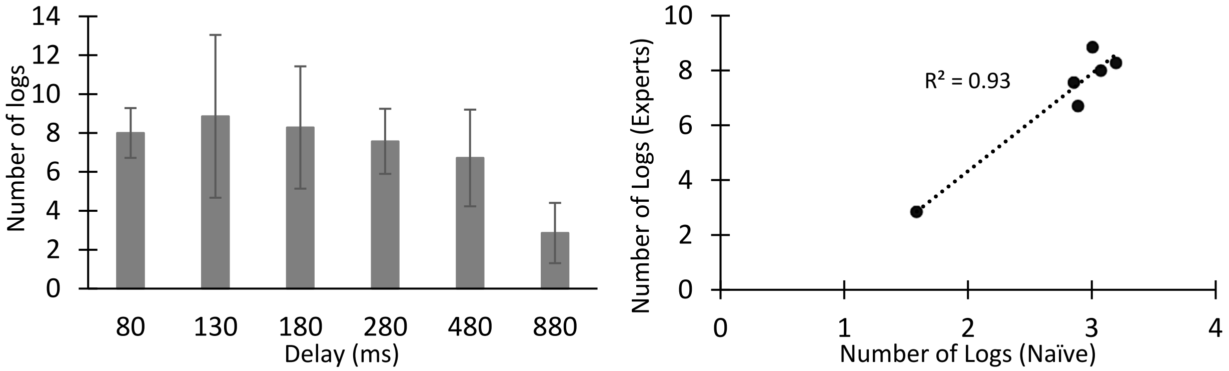


Fig. 21. (left) The mean number of logs loaded by the experienced log lifting test subjects per two min session as a function of the added Joystick delay in milliseconds (ms). The error bars indicate 95% confidence intervals. (right) Scatterplot between the mean number of logs reached by the naïve users (x-axis) compared to the mean number of logs reached by the experts (y-axis).

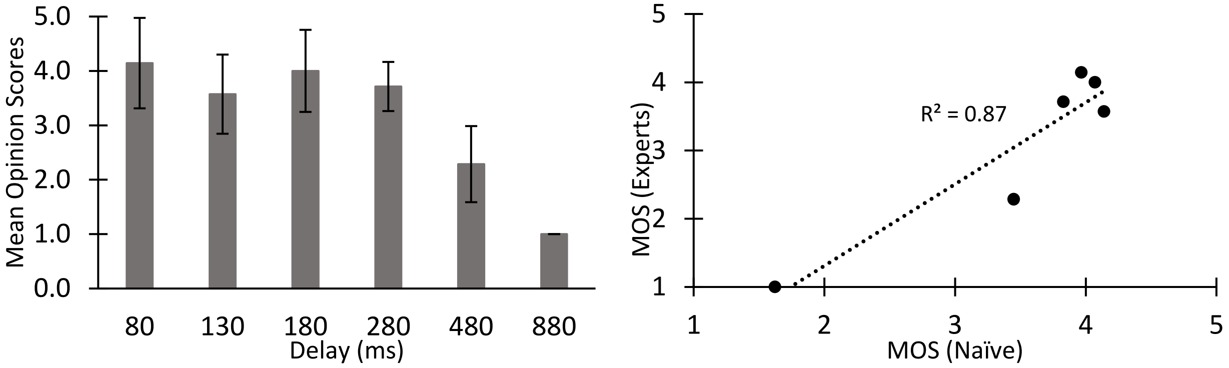


Fig. 22. (left) The MOS for Responsiveness Quality for different Joystick delays (right) in milliseconds (ms) rated by the experienced log lifting test subjects. The error bars indicate 95% confidence intervals. (right) Scatterplot between the MOS for Responsiveness Quality of the naïve users (x-axis) compared to the MOS for Responsiveness Quality of the experts (y-axis).

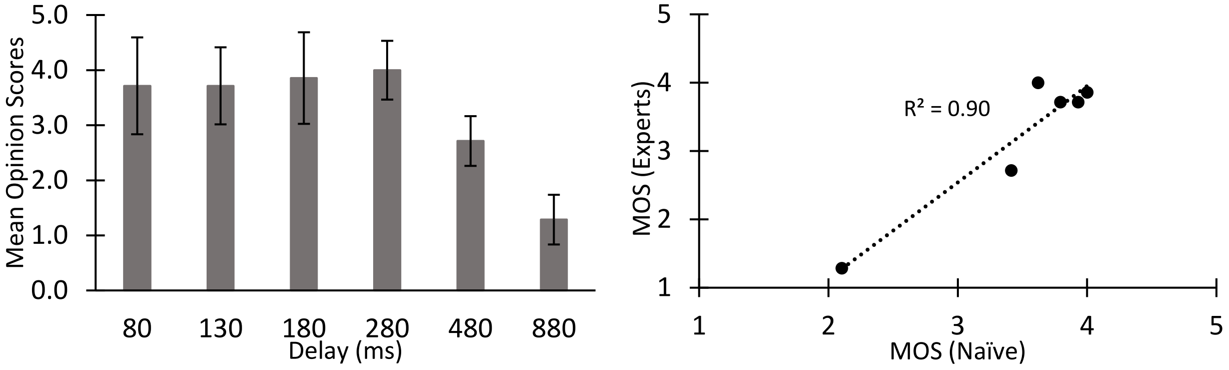


Fig. 23. (left) The MOS for Task Accomplishment Quality for different Joystick delays (right) in milliseconds (ms) rated by the experienced log lifting test subjects. The error bars indicate 95% confidence intervals. (right) Scatterplot between the MOS for Task Accomplishment Quality of the naïve users (x-axis) compared to the MOS for Task Accomplishment Quality of the experts (y-axis).

sudden transition into a low delay scenario might have been jarring and therefore reduced the participant QoE within the low delay scenario.

To check whether this suspected trend was actually manifested in the collected data, we performed two repeated-measures ANOVA tests. Test 1 considered the 80 ms and 130 ms scenario ratings as the intercept (main factor), and a binary predictor "preceded by 880 ms delay", which categorized the main factor scenarios as either having been, or not been, preceded by an 880 ms delay scenario. The predictor's value was set according to the experiment sequencing logs. The RM-ANOVA was performed for each of the five response scales (Overall, Comfort, Immersion, Responsiveness, Task Accomplishment). Test 2 considered only the 80 ms scenario, using the five response scales as the intercept (main factor), and the binary predictor "preceded by an 880 ms delay".

In test 1 (80 ms and 130 ms responses), the "preceded by an 880 ms delay" factor did not have a statistically significant ($p < 0.05$) effect on any of the response scales (Overall: F1,27 = 0.607, p = 0.442. Comfort: F1,27 = 0.058, p = 0.811. Immersion: F1,27 = 0.306, p = 0.584. Responsiveness: F1,27 = 1.518, p = 0.228. Task Accomplishment: F1,27 = 0.572, p = 0.455.)

There were also no statistically significant ($p < 0.05$) joint interactions between the "preceded by 880 ms delay" factor and the main delay factor. The interaction results for Overall scale were F1,27 = 1.390, p = 0.248, for Comfort scale F1,27 = 3.389, p = 0.076, for Immersion scale F1,27 = 0.352, p = 0.557, for Responsiveness scale F1,27 = 1.298, p = 0.264, and for Task Accomplishment scale F1,27 = 0.715, p = 0.404.

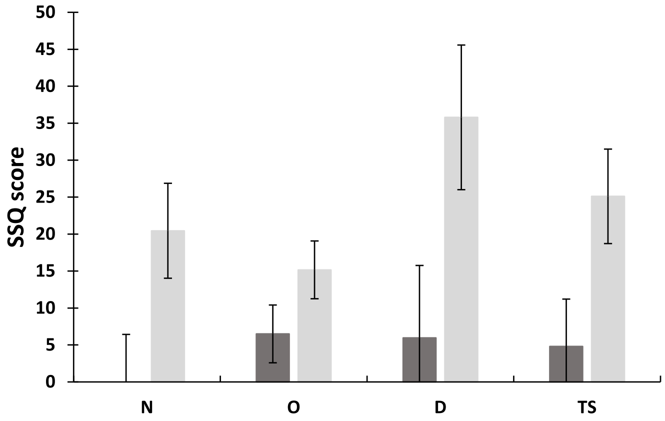


Fig. 24. Simulator Sickness Questionnaire (SSQ) scores for the delay experiment, where the left (medium gray) bars represent the symptom levels before the experiment and the right (light gray) bars the symptom levels after the experiment for the experienced log lifting test subjects. The different symptom groups along the x-axis are: Nausea (N), Oculomotor (O), Disorientation (D) and the Total Score (TS). The error bars indicate 95% confidence intervals.

In test 2 (only the 80 ms scenario responses, preceded or not preceded by the 880 ms delay scenario), the scales themselves had a statistically significant effect (F1,27 = 584.085, p = 7.9 $*$ 10^{-20}) on the results, suggesting that the scales were not treated as interchangeable by our participants. However, the "preceded by 880 ms delay" factor had no significant effect on its own (F1,27 = 0.004, p = 0.945), nor did it have any significant interaction with the main, scale factor (F4,108 = 0.502, p = 0.734).

Due to the comparatively high p-values, we can conclude that within the collected data there is no evidence of any delay-handling inertia. The results of participant scores on the low-delay scenarios were not significantly affected by whether the participants had just completed a high-delay (880 ms) scenario or not.

4.3.3. Learning effect

The test participants had not encountered the delay simulator prior to the testing process. Because each participant spent time using the simulator during the test, we wanted to check whether there was any task-learning effect exhibited by the participants. We categorized the responses in each scale according to the time already spent in the test, for all scenarios with delay < 280 ms. Due to results from preceding tests, showing a negative effect of high delay on participant QoE and task performance, we omitted the large delay scenarios (>280 ms) from this attempt to find a learning effect over time, and only considered delays from 80 ms to 180 ms, which – according to previous results – should not have a significant effect on participant responses. The cumulative results for the participant responses based on time spent are shown in Fig. 25a–e, and their task performance is shown in Fig. 25f. As can be seen in Fig. 25, in several scales (Overall, Responsiveness) the cumulative responses are consistent up to 180 ms of joystick delay, and begin deviating when a delay of 280 ms is included. This further suggests to consider delays up to 180 ms when searching for any learning effect.

We attempted to fit a non-linear model to the participant responses, with the general regression function "$y \sim b_1 + b_2 * x_2 - b_3 * x_1$", where y is the participant response, x_1 is the delay in milliseconds, x_2 is the time already spent in minutes, b_1 is an unknown base bias term, and b_2 and b_3 are unknown coefficients. The model was regressed using automated tools once for each response scale and once for all scales combined, using random seeds for b_1 to b_3 variables.

For most scales, the regression arrived at small values for the coefficients b_2 and b_3, giving a negligible effect of the delay and time-spent factors on the response. All regressed models had high prediction RMSE (from 0.825 for Immersion scale in the best case, to 1.56 for the Number of Logs scale). The p-values for the b_2 coefficient (corresponding to the time already spent variable) were consistently above 0.05.

Furthermore, automated tools were used to fit a reduced model from the starting function of "$y \sim 1 + x_1 + x_2 + x_1 * x_2 + {x_1}^2 + {x_2}^2 + {x_1}^2 * {x_2}^2$", allowing the fitting tools to estimate coefficients for each term and to drop any terms with no effect on the response y from the function. This regression again was attempted for each scale separately, and once for all scales together. Of these 7 estimated models, 4 removed the x_2 term entirely, indicating no impact from the time-spent factor. The remaining three models assigned small coefficients to the x_2 term (0.11 with p = 0.22, −0.04 with p = 0.02, and −0.08 with p = 0.06). Similar to before, the estimated models retained a large RMSE (1.54 in the worst case, 0.819 in the best case), which is excessive for reliably predicting responses on a scale of 1 to 5.

These results suggest that there is no significant effect from the time spent in the test to the participant responses or task performance in low-delay scenarios; in other words, no significant learning effect was observed in participants' results. As regards the other results, any existing learning effect that is this difficult to detect, would be counteracted in the experimental design by the randomization of test sequences, and therefore not affect the overall result in a serious way.

4.3.4. Time-in-test

To control for the effects from participant tiredness, we checked for possible effects from the Time-In-Test measurement on participant responses, without any joystick-delay based response filtering. The participant responses per Time-In-Test are shown in Fig. 26a–e, and their task performance is shown in Fig. 26f. The graph reveals that the mean responses per any particular joystick delay are erratic, however the mean response over all delays remains fairly flat regardless of the Time-In-Test.

A Repeated-Measures ANOVA test was performed on the data, with Time-In-Test and joystick delay as predictors, and per-scale response as the intercept (main factor). The joystick delay factor was shown as statistically significant ($F_{1,172}$ = 23.997, p = 2.2 $*$ 10^{-6}), which is consistent with the previously reported results. The Time-In-Test factor was not statistically significant ($F_{1,172}$ = 1.228, p = 0.269).

5. Discussion

5.1. Baseline experiment

The scale data indicates that the test subjects are not completely satisfied with Picture Quality (MOS = 3.6 i.e. between Fair and Good).

The Responsiveness is not problematic, and should not be, since the simulation is running on a sufficiently powerful PC, as evidenced by the RQ-score exceeding Good (MOS = 4.3).

For the Task Accomplishment Quality, the participant rating was between Fair and Good (MOS = 3.6) i.e. most people indicating a score somewhere in the middle. Our interpretation is that the test subjects did not have a strong opinion due to minimal experience in how a real system is works (as indicated in the pre-questionnaire).

Both the Immersive Quality (MOS = 4.3) and the Overall Quality (MOS = 4.1) were rated high i.e. exceeding good.

The SSQ indicates a very minor effect, although the disorientation symptom group showed a significant increase after the test, compared to the disorientation score before the test. A small fraction of people can show heightened sensitivity though.

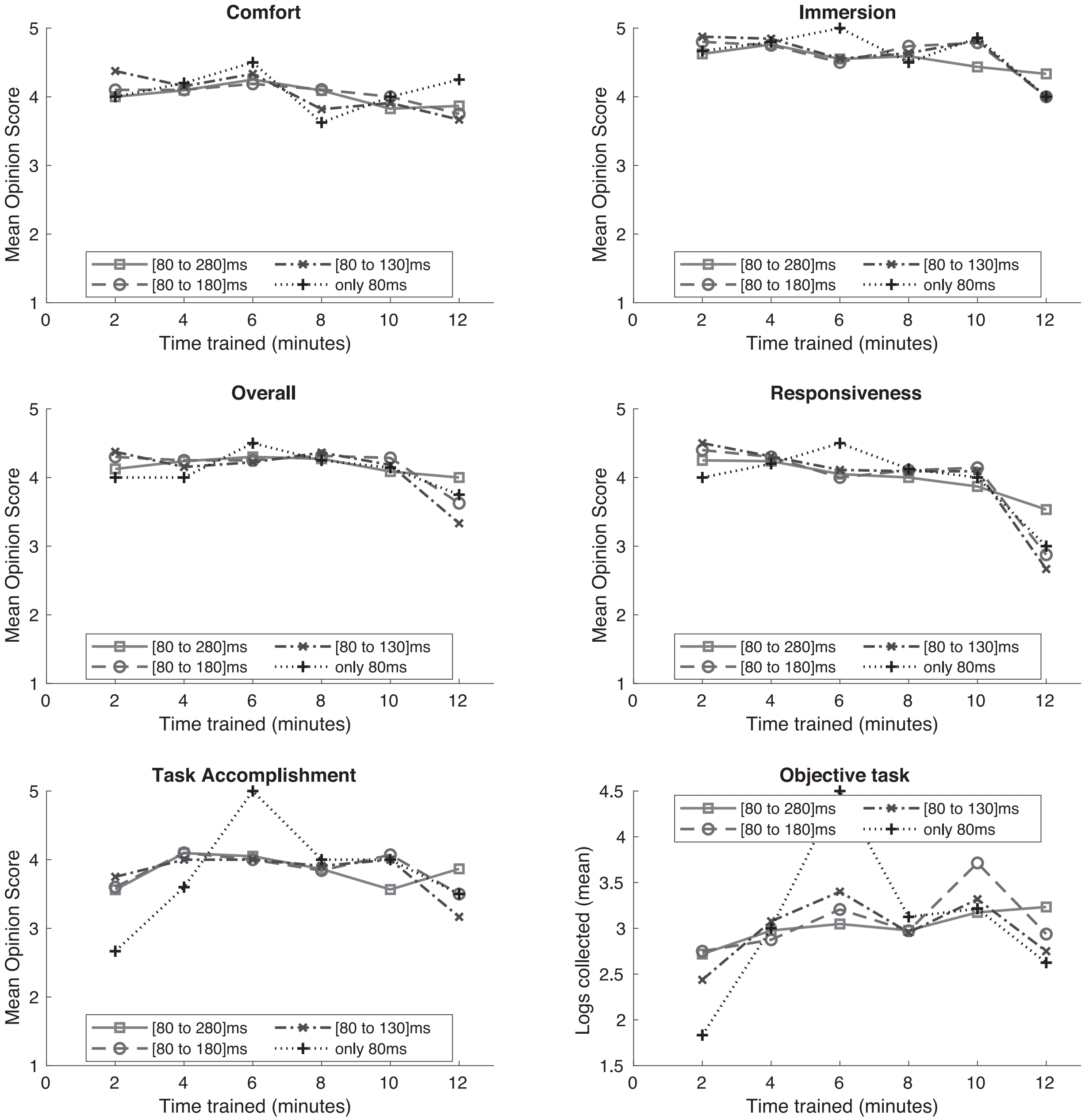


Fig. 25. (a–f): Mean participant responses for test instances with low joystick delay. Each line corresponds to a different definition of "low joystick delay", and lists the cumulative mean responses from cases within the given delay range.

5.2. Display and joystick delay experiment

In the display and joystick delay experiment, we discovered some impact on lower quality for higher delay, but the effect is relatively small and we only found significant effects on the highest level on added Display delay (30 ms) for Comfort Quality and Immersion Quality. One explanation for this seemingly small effect is that the scale analysis includes very few data samples from test subjects that did not finish the test. A reasonable assumption is that these test subjects would have rated the quality as being lower.

Another explanation is that the task was not sensitive enough to delays in the range in the current study. Earlier studies have shown that impact of delay on task performance is very task dependent, see e.g. [25,26]. Furthermore, test subjects may not always clearly identify the delay as being the source of the problem, as has been shown in telemeeting applications [37]. It can be noted that in the ratings from the test subjects, several inversions exist, i.e. that a test subject has rated lower quality of case with shorter delay compared to the case with longer delay.

The SSQ demonstrates a significant increase of symptoms. This is most likely connected to the Display delay, since an analysis of when test subjects stopped the experiment revealed that it was for the highest added Display delay. Furthermore, the 30 ms added Display delay had a statistically significant lower comfort quality. The SSQ score included all participants, even those that stopped, but the CQ was with a few exceptions based on the test subjects completing the test.

There was very little impact by the added Joystick delay. We can see tendencies to lower MOS on longer delays. However, no significant effects were found for the scales and as such, we attributed the significant effects on symptoms of SSQ to the Display delay. The Joystick delay had less impact, although we cannot identify the relative contributions of the two different delays.

It is known from the operation of the real crane system that the crane operators are normally very good at compensating for a bit of

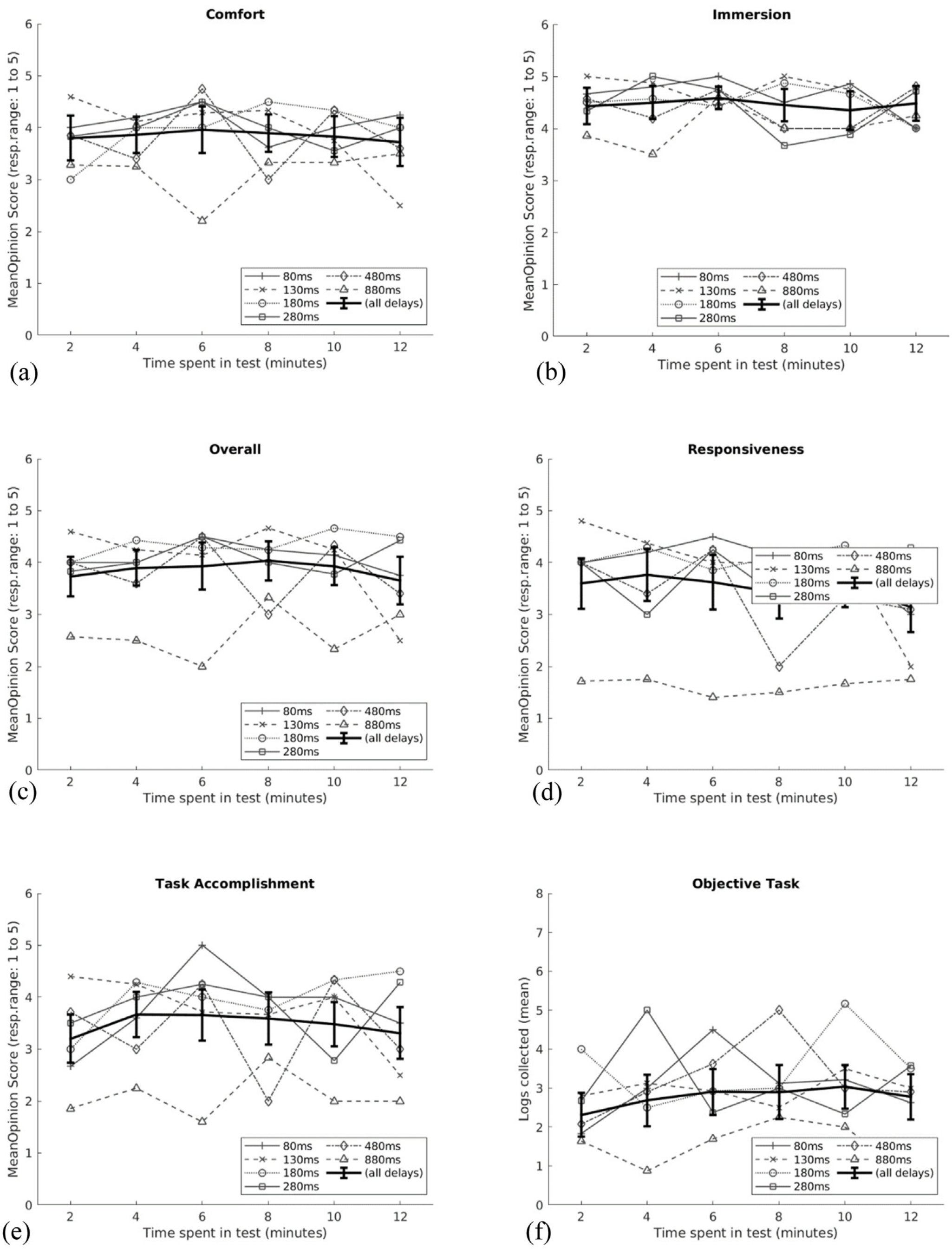


Fig. 26. Mean participant responses, sorted by time spent in the test at the response-giving point. The vertical bars for mean-of-all-delays lines indicate the 95% confidence interval.

delay in the crane controls, which is the Joysticks in this study. It is therefore reasonable to assume that also novice operators can manage to compensate for some delay when operating the crane. Furthermore, the baseline delay is fairly long (80 ms), so the shorter added Joystick delays are relatively small and could get unnoticed just because of that.

The actual undisturbed log loading time became shorter than we anticipated when planning and testing the experiment, as most test subjects needed almost 1 min to record their ratings, which is longer than for instance when giving scores on e.g. paper or a computer interface. It may have contributed to giving less influence on the experienced delay. However, one minute is still enough time to get quite a good understanding of the environment and the test subjects were fully immersed during the rating period and continuing performing their task, so we believe it had a minor influence, but intend to investigate this further.

5.3. Joystick delay experiment

The joystick delay experiment was motivated by the small statistically non-significant effects observed in the display and joystick delay experiment, for the impact on the added Joystick delay There, several inversions could be noticed where test subjects sporadically rated a lower quality of a case with shorter delay than of a case with longer delay, e.g. the zero added delay case had sometimes been rated as low as the 800 ms case.

The results presented here corroborate previous findings about insensitivities of hand controller latency. For a Telesurgery application, Rayman et al. [38] found that a set of robotic laparoscopic tasks could be performed at latencies up to 400 ms without significantly affecting task performance, and that simple tasks were possible to perform with a high level of accuracy with delays as high as 800 ms. Kim and Ryu [39] studied a haptic teleoperation task at different visual and haptic delays. Their results show that task performances were largely unaffected for delays below 200 ms, and that lower haptic delays were preferable to synchronized visual and haptic delays. In our log loading task, for all scales and the number of logs loaded, it is only for the longest delay that there is a clear effect i.e. 800 ms. This may be largely dependent on this particular task. However, if humans were very sensitive to this type and easily disturbed by it, larger effects should have been observed at shorter added delays, which is in line with the observation that both the experienced and the inexperienced test subjects react in a similar way.

The SSQ showed a significant increase of symptoms. The SSQ score included all participants, even those that stopped. Overall, this test did not seem to be too strenuous for most people to complete. However, there are some that are very sensitive, in this case a total 2 out of 30 participants.

5.4. Experimental procedure

The experiments are not precisely based on standardized procedures, as there are none yet currently available. However, the design is very much in line with current established QoE methodological approaches. Very briefly we can summarize the approach as follows: A technical parameter, in this case the delay is varied in an experiment where a trial is performed with a specific value of this parameter, then the test subjects rate their experience after each trial. This works well in the case where a specific understanding of the impact of such a parameter is of interest, but may not lead to any deeper understanding of the overall user experience or how that could be improved in the design of the system. Most of the standardized procedures e.g. ITU-R Rec. BT.500 [8] are not task based, although ITU-T Rec. P.912 [40] is a step in that direction with the recognition task. The recommendation ITU-T Rec. P.1301 is covering tele-meetings and therefore includes interaction, but the task is targeting conversation and is therefore different from the task in this study. We think that the current study shows that the methodological approach taken here has given useful results in understanding the impact of delay on task performance of this kind. However, we believe that further research is needed to fully understand how to bring UX and QoE closer together. Here it will be important to incorporate the qualitative methods from UX in a better way, especially when the test subjects with specialized backgrounds such as crane operator, truck drivers and wheel loader drivers cannot be used in large numbers, and quantitative statistics become uncertain.

For these experiments, verbal responses from the test subjects were used to get the rating on each of the individual scales, since the test subjects wore a VR-headset throughout the experiments and there was no implemented interface for responses in the simulator. This puts a high demand on the test leader to be consistent in the communication with the test subjects, not to influence them differently with the way scales are asked for and answers are received. A potential improvement would be to have the verbal questions pre-recorded and played to the test subjects after each trial, with the test leader writing down the responses. Even though the test leader communicated the questions to test participants verbally, and recorded the participants' verbal answers, the test leader also had a reference form of the questions printed. Therefore, even if the verbal inflections may have varied between repetitions, the structure and grammar of the questions posed to participants were consistent between the experiment sessions.

The test participants did not take breaks from VR between each trial. The virtual environment was reset, however the HMD remained on from experiment start to end (or cancellation). This lack of breaks may have contributed to an increased susceptibility to Simulator Sickness, due to the participants being immersed in an unreliable, changing virtual environment over the course of all trials.

The number of test subjects used could not be based on pre-planning in the same way as suggested by Brunnström and Barkowsky [41], since the variance in this type of experiments is less known than in traditional QoE experiments targeting video and audio. We could see in the Baseline study that the average standard deviation is 0.7, which is similar to what could be expected in well controlled subjective video quality experiments. However, when adding delay in the other experiments we saw an increased standard deviation, in the range of 0.9 to 1.1 for the scales and very high (1.5) for the performance metric of number of logs. This indicates that more test subjects are needed in these types of experiments. The obtained standard deviations can now be used for future planning of experiments and can be easily done by using the web tool "VQEGNumSubjTool" [42], based on [41]. In our experiments within this study, we have used 18 participants for BEXP, 25 for DJEXP (started with 35), and 29 for JEXP (started with 31), which seem to be appropriate numbers based on the observed standard deviation.

The length of the experiment per each test subject, the length of all trials, and the number of conditions are a test design balance question. Optimally, these variables should be balanced such that useful results are obtained without excessively tiring the test subjects. With our test design, we targeted a session time of 20 min or less per participant. For DJEXP we had 9 conditions, which were switched every 2 min and during the last half minute the ratings were collected. It turned out that half a minute was a bit short, so in JEXP, we altered the rating to be between the 2 min trials to keep this constant. We then reduced the number of conditions to 6, in order to keep the total time in the VR-headset about 20 min.

The 2 min period seems to have been appropriate for the test subjects, as we could observe that the subjects were able to operate the crane at least a couple of times. Each crane operation involved moving from a pile of logs to the truck, back and forth, giving the participants enough time to sense the influence of the simulated delay.

6. Conclusion

The baseline study shows that most people are more or less happy with the VR-system and that it does not have a strong effect on any symptoms as listed in the SSQ. There is some room for improvement since all scales were not above Good (> 4). For instance, the Picture Quality only had a MOS of 3.6.

In the display and joystick delay study, we found significant effects on Comfort Quality and Immersion Quality for higher Display delay (30 ms), but very small impact of Joystick delay. Furthermore, the Display delay had strong influence on the symptoms in the SSQ, as well as causing test subjects to decide not to continue to the end with the experiments, and this was also found to be connected to the longer added Display delays (≥20 ms).

In the joystick delay study, we found no significant effects of delays on the task performance (number of logs loaded) or on any scales up to 200 ms. Very weak effects were found for 400 ms, it was only found significantly lower in responsiveness quality. A strong significant effect was found for 800 ms added delay, being significantly lower for the number of logs and for all scales against all the other delays (with the exception one case). It seems as if the delays need to become at least about half a second to be clearly noticeable and disturbing for this type of task. Although the group of experienced log lifting test subjects are relatively small, it supports the findings of the inexperienced group to be just applicable to the inexperienced test subjects but seems to apply more generally. The symptoms reported in the Simulator Sickness Questionnaire were significantly higher for all the symptom groups, but most reported just slight symptoms, a few also moderate and just one, a severe symptom. Also, for the SSQ the results were very similar for both the inexperienced and experienced group. Two out of thirty test persons stopped the test prematurely due to their symptoms. Thus, most test persons were fine using the VR-simulator, but a few seem to be very sensitive.

The overall conclusion is that latency in the display update has a severe impact and should be avoided or limited to very short latency, i.e. less than 30–35 ms. For latency in the Joysticks or hand controllers, much longer latencies can be tolerated and if it is kept below 0.5 s, it has a very limited impact on the test task of this study.

CRediT authorship contribution statement

Kjell Brunnström: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Project administration, Funding acquisition. **Elijs Dima:** Formal analysis, Data curation, Writing - original draft, Writing - review & editing. **Tahir Qureshi:** Writing - original draft, Writing - review & editing. **Mathias Johanson:** Writing - review & editing, Funding acquisition. **Mattias Andersson:** Writing - review & editing. **Mårten Sjöström:** Writing - review & editing, Investigation, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The economic support from the Knowledge Foundation, Sweden (grant nr 20160194) and VINNOVA (Sweden's innovation agency, grant nr 2018-00735) is hereby gratefully acknowledged. We would also like to thank Joakim Edlund and Sidra Muneer for helping out performing experiments as well as Jade Lumley for the thorough native language review.

References

[1] ITU-T, Vocabulary for Performance, Quality of Service and Quality of Experience (ITU-T Rec. P.10/G.100), International Telecommunication Union (ITU), ITU Telecommunication Standardization Sector: Place des Nations, CH-1211 Geneva 20, 2017.

[2] P. Le Callet, S. Möller, and A. Perkis (Eds.), 2012. Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003). Lausanne, Switzerland, 2012.

[3] K. Brunnström, M. Sjöström, M. Imran, M. Pettersson, M. Johanson, Quality of experience for a virtual reality simulator, in: Human Vision and Electronic Imaging 2018, Society for Imaging Science and Technology, Burlingame, California USA, 2018, 28 Jan. - 2 Feb, 2018.

[4] K. Brunnström, E. Dima, M. Andersson, M. Sjöström, T. Qureshi, M. Johanson, Quality of experience of hand controller latency in a virtual reality simulator, in: Human Vision and Electronic Imaging 2019, Society for Imaging Science and Technology, Burlingame, California USA, 2019, pp. HVEI–218, 13-17 Jan 2019.

[5] F. Okura, M. Kanbara, N. Yokoya, Augmented telepresence using autopilot airship and omni-directional camera, in: IEEE International Symposium on Mixed and Augmented Reality 2010, IEEE Xplore, Seoul, Korea, 2010, pp. 259–260.

[6] V.V. Saxena, T. Feldt, M. Goel, Augmented telepresence as a tool for immersive simulated dancing in experience and learning, in: The India HCI 2014 Conference on Human Computer Interaction, ACM New York, NY, USA, 2014, pp. 86–89, http://dx.doi.org/10.1145/2676702.2676708.

[7] S. Möller, A. Raake, Quality of Experience - Advanced Concepts, Applications and Methods, in: T-Labs Series in Telecommunication Services, Springer International Publishing, Switzerland, 2014.

[8] ITU-R, Methodology for the Subjective Assessment of the Quality of Television Pictures (ITU-R Rec. BT.500-14), International Telecommunication Union, Radiocommunication Sector, 2019.

[9] ITU-T, Subjective Video Quality Assessment Methods for Multimedia Applications (ITU-T Rec. P.910), International Telecommunication Union, Telecommunication standardization sector, 1999.

[10] K. De Moor, M. Fiedler, P. Reichl, M. Varela, Quality of Experience: From Assessment to Application (Dagstuhl Seminar 15022), DROPS (Dagstuhl Online Publication Service), 2015, http://dx.doi.org/10.4230/DagRep.5.1.57, (http://drops.dagstuhl.de/opus/volltexte/2015/5036/).

[11] ITU-T, 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, 2014.

[12] ITU-T, Display Requirements for 3D Video Quality Assesment (ITU-T Rec. P.914), International Telecommunication Union, 2016.

[13] ITU-T, Information and Guidelines for Assessing and Minimizing Visual Discomfort and Visual Fatigue from 3D Video (ITU-T Rec. P.916), International Telecommunication Union, 2016.

[14] ITU-T, Subjective Assessment Methods for 3D Video Quality (ITU-T Rec. P.915), International Telecommunication Union, 2016.

[15] J. Puig, A. Perkis, F. Lindseth, T. Ebrahimi, Towards an efficient methodology for evaluation of quality of experience in augmented reality, in: Fourth International Workshop on Quality of Multimedia Experience (QoMEX 2012), IEEE Xplore, Melbourne, Australia, 2012, pp. 188–193.

[16] P. Tripicchio, E. Ruffaldi, P. Gasparello, S. Eguchi, J. Kusuno, K. Kitano, M. Yamada, A. Argiolas, M. Niccolini, M. Ragaglia, C.A. Avizzano, A stereo-panoramic telepresence system for construction machines, Procedia Manuf. 11 (2017) 1552–1559, http://dx.doi.org/10.1016/j.promfg.2017.07.292.

[17] L. Lachs, Multi-Modal Perception. R. Biswas-Diener & E. Diener (Eds.), Noba Textbook Series: Psychology. Champaign. 2020. Available from: http://noba.to/cezw4qyn, Access Date: 12 Aug 2020.

[18] A. Smuts, What is interactivity? J. Aesthet. Educ. 43 (4) (2009) 53–73.

[19] L. Janowski, P. Kozłowski, R. Baran, P. Romaniak, A. Glowacz, T. Rusc, Quality assessment for a visual and automatic license plate recognition, Multimedia Tools Appl. 68 (1) (2014) 23–40, http://dx.doi.org/10.1007/s11042-012-1199-5.

[20] N. Dużmańska, P. Strojny, A. Strojny, Can simulator sickness be avoided? A review on temporal aspects of simulator sickness, Front. Psychol. 9 (2018) 2132, http://dx.doi.org/10.3389/fpsyg.2018.02132.

[21] K. Debattista, T. Bashford-Rogers, C. Harvey, B. Waterfield, A. Chalmers, Subjective evaluation of high-fidelity virtual environments for driving simulations, IEEE Trans. Hum.-Mach. Syst. 48 (1) (2018) 30–40, http://dx.doi.org/10.1109/THMS.2017.2762632.

[22] T. Ni, H. Zhang, C. Yu, D. Zhao, S. Liu, Design of highly realistic virtual environment for excavator simulator, Comput. Electr. Eng. 39 (7) (2013) 2112–2123, http://dx.doi.org/10.1016/j.compeleceng.2013.06.010.

[23] G. Strazdins, B.S. Pedersen, H. Zhang, P. Major, Virtual reality using gesture recognition for deck operation training, in: OCEANS 2017 - Aberdeen, 2017.

[24] M. Suznjevic, M. Mandurov, M. Matijasevic, Performance and qoe assessment of HTC vive and oculus rift for pick-and-place tasks in VR, in: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), 2017.

[25] C. Jay, M. Glencross, R. Hubbold, Modeling the effects of delayed haptic and visual feedback in a collaborative virtual environment, ACM Trans. Comput.-Hum. Interact. 14 (2) (2007) 8, http://dx.doi.org/10.1145/1275511.1275514.

[26] C. Jay, R. Hubbold, Delayed visual and haptic feedback in a reciprocal tapping task, in: First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. World Haptics Conference, 2005.

[27] B. Knörlein, M.D. Luca, M. Harders, Influence of visual and haptic delays on stiffness perception in augmented reality, in: 2009 8th IEEE International Symposium on Mixed and Augmented Reality, 2009.

[28] Q. Qian, Y. Ishibashi, P. Huang, Y. Tateiwa, H. Watanabe, K. Psannis, QoE assessment of object softness in remote robot system with haptics: Comparison of stabilization control, 2018.

[29] K. Desai, S. Raghuraman, R. Jin, B. Prabhakaran, QoE studies on interactive 3D tele-immersion, in: 2017 IEEE International Symposium on Multimedia (ISM), 2017, 2017.

[30] A. Tatematsu, Y. Ishibashi, N. Fukushima, S. Sugawara, QoE assessment in haptic media, sound and video transmission: Influences of network latency, in: 2010 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR 2010), 2010, 2010.

[31] R.S. Kennedy, N.E. Lane, K.S. Berbaum, M.G. Lilienthal, Simulator sickness questionnaire: An enhanced method of quantifying simulator sickness, Int. J. Aviat. Psychol. 3 (3) (1993) 203–220.

[32] K. Brunnström, K. Wang, S. Tavakoli, B. Andrén, Symptoms analysis of 3D TV viewing based on simulator sickness questionnaires, Quality User Exp. 2 (1) (2017) 1–15, http://dx.doi.org/10.1007/s41233-016-0003-0.

[33] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (3/4) (1965) 591–611, http://dx.doi.org/10.2307/2333709.

[34] S.E. Maxwell, H.D. Delaney, Designing Experiments and Analyzing Data : A Model Comparison Perspective, second ed., Lawrence Erlbaum Associates, Inc, Mahwah, New Jersey, USA, 2003.

[35] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (6) (1945) 80–83, http://dx.doi.org/10.2307/3001968.

[36] L.L. Havlicek, N.L. Peterson, Robustness of the T test: A guide for researchers on effect of violations of assumptions, Psychol. Rep. 34 (3_suppl) (1974) 1095–1114, http://dx.doi.org/10.2466/pr0.1974.34.3c.1095.

[37] ITU-T, Telemeeting Assessment - Effect of Delays on Telemeeting Quality (ITU-T Rec. P.1305), International Telecommunication Union, Telecommunication standardization sector, 2016.

[38] R. Rayman, S. Primak, R. Patel, M. Moallem, R. Morady, M. Tavakoli, V. Subotic, N. Galbraith, A. Van Wynsberghe, K. Croome, Effects of Latency on Telesurgery: An Experimental Study, Springer Berlin Heidelberg, 2005, pp. 57–64, http://dx.doi.org/10.1007/11566489_8.

[39] Y. Kim, J. Ryu, Performance analysis of teleoperation systems with different haptic and video time-delay, in: 2009 ICCAS-SICE, 2009, 2009.

[40] ITU-T, Subjective Video Quality Assessment Methods for Recognition Tasks (Rec. ITU-T P.912) (Rec. ITU-T P.912), 2016.

[41] K. Brunnström, M. Barkowsky, Statistical quality of experience analysis: on planning the sample size and statistical significance testing, J. Electron. Imaging 27 (5) (2018) 11, http://dx.doi.org/10.1117/1.JEI.27.5.053013.

[42] W. Robitza, K. Brunnström, VQEGNumSubjTool - Calculating Number of Subjects, 2019, Access Date: Access Date|, Available rom: https://slhck.shinyapps.io/number-of-subjects/.

Kjell Brunnström Ph.D., is a Senior Scientist at RISE Research Institutes of Sweden AB and Adjunct Professor at Mid Sweden University. He is an expert in image processing, computer vision, image and video quality assessment having worked in the area for more than 25 years. Currently, he is leading standardization activities for video quality measurements as Co-chair of the Video Quality Experts Group (VQEG). His current research interests are in Quality of Experience for visual media in particular video quality assessment both for 2D and 3D, including AR, VR and remote operation of machines, as well as display quality related to the TCO Certified.

Elijs Dima received his B.Sc. and M.Sc. degrees in Computer Engineering from Mid Sweden University, Sweden, in 2013 and 2015, and his Lic. degree in 2018. Since 2015, he has been a researcher, lab supervisor, teaching assistant and Ph.D. student in the Realistic 3D research group at Mid Sweden University. His research interests include 360-degree video and light field capture, rendering and streaming, parallel data processing, and the synchronization, calibration, modeling, and development of Virtual Reality, Augmented Reality and multi-camera systems.

Tahir Qureshi got his Ph.D. from the Royal Institute of Technology 2012 researching model-based development as the main research area. He worked on architecture centric approaches for developing automotive embedded systems. From Jan 2013 Dr Qureshi has been working at HIAB AB as a research manager on embedded systems.

Mathias Johanson, Ph.D., is R&D manager at Alkit Communications AB and an expert on video-mediated communication and distributed collaborative environments. His research interests also include automotive telematics and e-health systems and services.

Mattias Andersson received the M.Sc. in Applied Physics and Electrical Engineering in 1998, the Licentiate of Technology in Media Technology in 2004 and the PhD in Media Technology, all from Linköping University in 2006. He then worked in commercial and industrial life as with computer vision, multispectral imaging and color appearance. He joined Mid Sweden University in 2011 and is currently working as a research engineer. His current research interests include multidimensional imaging and augmented reality.

Mårten Sjöström received the M.Sc. degree in electrical engineering and applied physics from Linköping University, Sweden, in 1992, the Licentiate of Technology degree in signal processing from KTH, Stockholm, Sweden, in 1998, and the Ph.D. degree in modeling of nonlinear systems from EPFL, Lausanne, Switzerland, in 2001. He was an Electrical Engineer with ABB, Sweden, from 1993 to 1994, was a fellow with CERN from 1994 to 1996. He joined Mid Sweden University in 2001, and was appointed an Associate Professor and a Full Professor in Signal Processing in 2008 and 2013, respectively. He has been the Head of the Computer and System Sciences with Mid Sweden University since 2013. He founded the Realistic 3-D Research Group in 2007. His current research interests are within multidimensional signal processing and imaging, as well as system modeling and identification.