International Journal of Human-Computer Studies 145 (2021) 102536

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

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

A low-cost predictive display for teleoperation: Investigating effects on human performance and workload

Henrikke Dybvik*, Martin Løland, Achim Gerstenberg, Kristoffer Bjørnerud Slåttsveen, Martin Steinert

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology (NTNU), Richard Birkelands vei 2B, 7491 Trondheim, Norway

ARTICLE INFO

Keywords:
Predictive display
Human operator
Performance
Subjective workload

ABSTRACT

Teleoperation in an environment with latency is difficult and highly stressful for human operators, resulting in high cognitive workload and decreased human performance. This work investigates if a simple predictive display can increase performance and lower subjective workload for the human operator when teleoperating a remotely operated vehicle (ROV). A predictive display based on image transformation was developed by applying positional and scale transformations to the video feed and tested. An experiment was designed, consisting of a simple navigational task (peg-in-hole game) with a ground ROV, in three distinct conditions: C1. Latency, C2. Latency with predictive display (PD) and C3. Baseline (no added latency). Findings from N = 57 participants show a statistically significant increase of 20% in human performance with the aid of the predictive display. Although differences in subjective workload was not statistically significant, both subjective performance and actual game performance did increase significantly by using the predictive display. In fact, the latter almost doubled for participants defining themselves as regular gamers. Lastly, A principle component analysis (PCA) was conducted investigating confounding factors with confirmatory results.

1. Introduction – Predictive technology can combat the detrimental effects of latency in teleoperation

Teleoperation, also called remote operation, is electronic remote control of machines or vehicles and it includes applications of remotely operated vehicles (ROVs) on ground, under water, subsea, aerial and in space (Draper et al., 1998). Teleoperation is a subclass of telepresence; "the perception of presence within a physically remote or simulated site" (Draper et al., 1998). Telepresence is generally viewed as being beneficial to mission performance and is furthermore hypothesized to improve efficiency and/or reduce operator workload (Draper et al., 1998). There are multiple challenges related to teleoperation, one of which is latency. In this work, we are interested in latency, also called time delay, which refers to the delay between operator input action (steering commands) and visible output response of the video feed (Chen et al., 2007). Teleoperation in an environment with latency, especially basic driving, is difficult and highly stressful for the human operator, resulting in high cognitive workload (Matheson et al., 2013) and decreased human performance (Chen et al., 2007), e.g. observed as an increase in task completion time or reduced accuracy (Lane et al., 2002). Approaches to overcome the detrimental effects of latency in teleoperation include increasing the level of automation (which excludes the human from the loop), provide information to increase the situational awareness of the human operator and predictive technology.

Predictive technology spans several approaches, either categorized as dynamic system models or free model approaches. Model free approaches include superimposed information models, 3D graphic models, and video manipulation. Superimposed information and 3D graphics models show promising results by greatly reducing task completion times, but require advanced algorithms, potentially expensive equipment and extensive information regarding the environment and the ROV. Video manipulation can increase performance of human operators' and it is simpler in comparison, as it alters the delayed video feed to mimic movements and environment in real time. Simple video manipulation can provide time efficient and inexpensive means to enhance performance of human operators' in settings where extensive information regarding the ROV and its environment is unavailable, or the opportunity to utilize expensive equipment or highly advanced algorithms is not a possibility.

With basis in existing video manipulation methods based on image transformation, we developed a simpler predictive display by applying image positional and scale transformations to the video feed. This

* Corresponding author.
E-mail address: henrikke.dybvik@ntnu.no (H. Dybvik).

https://doi.org/10.1016/j.ijhcs.2020.102536
Received 1 July 2019; Received in revised form 18 August 2020; Accepted 21 August 2020
Available online 25 August 2020

predictive display requires a few lines of code and can be applied to several ROV configurations. In this work, we are interested in human operators' performance and their subjectively experienced workload while using predictive technology. The aim of this article is to investigate if a simple predictive display can increase performance and lower subjective workload of human operators' during teleoperation. To do so an experiment was set up to investigate changes in human operator performance and workload when operating an ROV under three distinct conditions, each condition with a distinct display and latency. The participants were presented with a single, simple navigational task, framed as a peg-in-hole game using a ground ROV with a first-person camera view. The conditions were C1. Latency, 2. Latency with Predictive Display (w/PD) and 3. Baseline. Data collected included objective performance (task score), and subjective workload (RTLX), demographics and other variables. N = 57 participants were recruited and the hypotheses (task performance and subjective workload) tested using ANOVA. A post hoc Exploratory Data Analysis (EDA), specifically a principal component analysis (PCA) explore influencing factors.

Following the introduction, the article is structured as follows; the background second section covers challenges in teleoperation, focusing on latency and its detrimental effects on human performance and workload. Means to compensate for latency are discussed, emphasizing various predictive technologies. The third section describes development and implementation of a predictive display, and the experiment design, including stimuli, data collection, procedure, and data analysis. Section four presents the results of the statistical tests before providing the result from the EDA. A discussion of the presented results follows in section five, before the conclusion.

2. Background – Latency in teleoperation, human performance, and workload

This section describes challenges in telepresence, detailing latency and its detrimental effects, with a focus on human operator performance and subjectively experienced workload. Human operator performance decrease and workload increase as latency is introduced in teleoperation. Means to compensate for latency are discussed, predictive technologies in particular. Lastly, the section provides means to measure human operator performance objectively and workload subjectively.

2.1. Latency in teleoperation and its related challenges

2.1.1. Telepresence and related challenges

Draper et al. (1998) defines telepresence as "the perception of presence within a physically remote or simulated site". Teleoperation is one subclass of telepresence (Sheridan, 1995). Telepresence is beneficial to mission performance and is furthermore hypothesized to improve efficiency and/or reduce operator workload. Chen et al. (2007) reviewed 150 articles investigating factors in telepresence, and how they influence operator performance and challenges related. They found eight main factors; field-of-view (FOV), orientation, camera viewpoint, depth perception, video quality and frame rate, time delay (or latency) and motion.

2.1.2. Latency challenges in telepresence

In this work, we are interested in time delay, or latency, which will be used throughout this article, which refers to the delay between operator input action (steering commands) and visible output response of the video feed (Chen et al., 2007). Latency is usually a result of information having to be conveyed over a communication network (Chen et al., 2007). The total latency of the teleoperation system can further result from a combination of a number of reasons, such as software design, hardware design, physical limitations such as distance and obstacles, signal processing, etc. Thus, total latency can be both fixed and variable (Lane et al., 2002). There are important distinctions between the two, e.g. they influence performance differently (Davis et al., 2010; Neumeier et al., 2019; Oboe and Fiorini, 1998).

The causes of latency are not within the scope of this work, and we consider fixed delay only. We are interested in the total perceived latency; i.e. the time from when the human operator issues a command until they visually perceive a reaction in the vehicle in the video feedback.

2.1.3. Latency in teleoperation and its detrimental effects

Latency produces a mismatch between given input commands and visual feedback of vehicle reactions. This creates a conflict for human perception. To correct for this during operation the human operator must remember the inputs command given until they see the desired output produced by the vehicle in the video feed (Matheson et al., 2013). In addition, as new information is prompted on the video feedback this must be mentally connected with the commands issued previously (i.e. the vehicles previous state), and thereafter combined that with issuing new commands based on this conjunction of information (Ricks et al., 2004). Latencies as low as 10 - 20 ms can be detected by humans' visual perception (Chen et al., 2007). Taken together, this can degrade human performance (Chen et al., 2007) and can increase subjectively experienced workload (Ricks et al., 2004).

2.1.4. Latency in teleoperation degrades human performance

The detrimental effect of latency on human performance can be seen in Table 1, which includes relevant research investigating the effect of video feed latency on human performance in a given task. Human performance includes course completion time, task completion time, task score, accuracy, etc. This table describe the task and the related increase factor, where a 40% increase in task completion time corresponds to an increase factor of 1.40. For example a needle-driving task at 100 ms latency had an increase factor of 1.5 (Xu et al., 2014). The relationship between latency and task completion time is task dependent, notably it is similar for similar tasks. For example; a linear relationship between latency and task completion time was found in a mobile robot operating task (Ando et al., 1999) and a vehicle peg-in-hole task (Lane et al., 2002), whereas an exponential relationship was found in a telerobotic surgical technique task (Xu et al., 2014).

Table 1
Task completion time for a variety of tasks and latencies.

Author	Task	Participants	Latency [ms] and increase factor		
			100 – 300 ms	400 – 700 ms	800 – 1500 ms
(Fabrizio et al., 2000)	Pin transfer	N = 6	1.04 - 1.21*	1.17 - 1.41*	1.11 - 1.58*
(Xu et al., 2014)	Energy dissection	N = 16	1.4 - 1.8	2.7 - 4.3	
(Xu et al., 2014)	Needle-driving	N = 16	1.5 - 2.1	2.5 - 6.2	
(Perez et al., 2016)	Surgical simulator	N = 37	0.75	1.5	
(Lum et al., 2009)	Block transfer	N = 14	1.45	2.04	
(MacKenzie and Ware, 1993)	Target acquisition	N = 8	1.64		

* Estimated from graph.

2.1.5. Latency in teleoperation increase workload

The notion of workload or cognitive load is argued to be predictive of both performance in human-machine interactions as well as the mental state of the operator. Workload is described as a relation between the mental resources a task demands and the resources available from the human operator (Parasuraman et al., 2008). It is a multi-dimensional construct emerging from the interaction between task, context, operator capabilities, behavior, perceptions and (mental and physical) state (Hart and Staveland, 1988a; Parasuraman et al., 2008). This mental load posed on a human operator by latency in teleoperation negatively affects their ability to control a vehicle in an efficient manner (Ricks et al., 2004). The human operator's *subjectively* experienced workload is important (Hart and Staveland, 1988a), since this might alter behavior. Should an operator experience a situation as high workload, the operator might adopt strategies to mitigate workload. In the specific case of teleoperation human operators tend to perform steering commands correcting for the mismatch in given input and visually perceived output, causing the vehicle to oscillate and limiting top speed (Appelqvist et al., 2007). Teleoperation in an environment with time delay, in particular basic driving, is difficult and highly stressful for the human operator, resulting in high cognitive workload (Matheson et al., 2013). Extended exposure to such an environment can create cognitive overload leading to mental fatigue (Lim et al., 2010; Matheson et al., 2013).

2.1.6. Latency compensation

There are multiple approaches to reduce the detrimental effects of latency. First, increasing the level of automation (LOA) reduces the operator workload and improve safety (Dorais et al., 1999; ENDSLEY, 1999; Goodrich et al., 2001; Luck et al., 2006; Schutte, 2017). A second option is providing the human operator with information and/or previously given input commands, increasing situational awareness and leading to higher performance and/or decreasing subjective workload (Chen et al., 2007; Miller and Machulis, 2005; Nielsen et al., 2007). A third option is predictive technology, which can be displays, control algorithms and graphical models attempting to predict the state of the ROV based on the vehicles current state and commands issued by the operator. Chen et al. (2007) conclude it is the most promising solution if eliminating latency from the system is impossible, and highlight that predictive displays has been shown to reduce task performance time by up to 150%.

2.1.7. Predictive technology

A range of experiments where predictive technology has been used are shown in Table 2, illustrating a wide variety of experimental tasks, robot configurations and predictive method. Exact robot configuration can be known, including examples such as robot-arm manipulators fixed to a user defined reference frame, or not known, such as vehicles subjective to external forces or floating freely. The unknown robot configuration challenges the predictive technology as it must account for unknown and changing external factors. Common for the experiments in Table 2 is that they involve lateral movement in an alignment or aiming task, which are particularly exposed to detrimental effects of latency in video feedback. Correctional behavior commonly occurs, causing operators to overshoot a target or employ a wait-and-move strategy. This behavioral strategy increases task completion time and occurs around one second latency (Lane et al., 2002).

In general, predictive technology calculates a future predicted state of the robot based on different variables and methods. Methods can rely on dynamic system equations, such as Zhang and Li. (2016) who used a spacecraft's state equations and its dynamic properties to calculate the predicted state. An image of the predicted state is provided to the operator which can issue commands accordingly. In contrast, a model free approach, which excludes dynamics, is often used in contexts where accurate modeling of external forces isn't a possibility, such as in space applications. Predictive technology within model free approaches includes superimposed predictive information, 3D graphic models and video manipulation.

The first category superimposes (or overlays) information on a delayed video feed, providing the operator with an estimate of the vehicles future state. Superimposed predictive information is often visualized as vector graphics where lines of dots follow a path. For example, Mathan et al. (1996) superimposed directional velocity information of a lunar rover on a video display. Further, airplane and helicopter displays have a *tunnel in the sky* showing where the aircraft should be going and a cross indicating the predicted trajectory (Grunwald et al., 1981). In cases with large amounts of lateral movement this approach might not be applicable as the predicted heading can come off screen.

3D graphics model (or virtual reality (VR) based predictive display) use sensor technology input such as Monocular Simultaneous Location and Mapping (SLAM), stereo imagery, vision-based structure from motion (SFM), light detection and ranging (LiDAR), or radio detection and ranging (radar), etc., to construct a three-dimensional world, wherein images from ordinary cameras are rendered on the surface of the virtual world.

Then, a virtual camera is placed inside the virtual world in the predicted position of the real camera and operators' are presented with the virtual video feed as virtual reality (VR) or augmented reality (AR). This method is particularly popular in combination with robot arm

Table 2
Predictive technology with task completion time reduction.

Author	Robot system Task	Predictive technology Camera	Participants Latency	Reduction in task completion time
(Lu et al., 2018)	Car simulator Driving	Model-free framework Simulated human	N = 12 Not reported	8%
(Hu et al., 2016)	2-6 DOF manipulator Camera alignment	Simulated 3D Virtual	N = 15 300 ms, 500 ms, 1000 ms	33%, 58%, 65%*
(Zheng et al., 2016)	Car simulator Driving	Model-free framework Simulated human	N = 5 900 ms	35%
(Lovi et al., 2010)	Robot arm on Segway Object alignment	Vision-based monocular modelling At end effector	N = 5 300 ms	33%*
(Matheson et al., 2013)	Rover Driving	Projected field of view Fixed to car	N = 12 3000 ms	48% - 64%*
(Rachmielowski et al., 2010)	Virtual with Phantom OMNI Alignment	Reconstructed 3D environment At end effector	N = 12 300 ms	29% - 30%*
(Mathan et al., 1996)	Lunar vehicle Manoeuvring	Superimposed directional information Fixed to car	N = 8 5000 ms	24% - 30%
(Bejczy et al., 1990)	6DOF PUMA robot Tapping	Superimposed phantom robot Fixed	N = 2 1000 ms, 4000 ms	13% - 34%, 40% - 56%

* Estimated from graph.

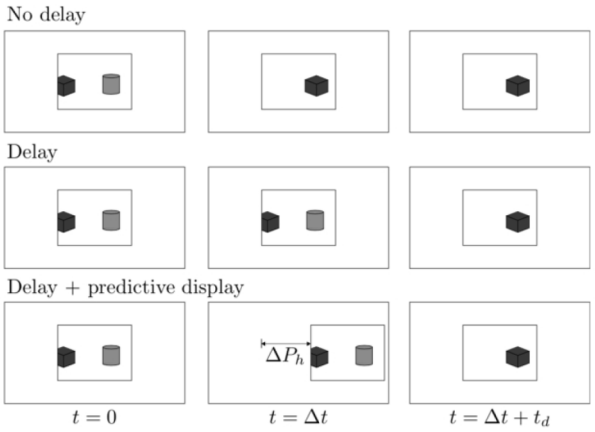

Fig. 1. Monitor for the human operator. The outer box is total screen size, whereas the inner box is the video feed.

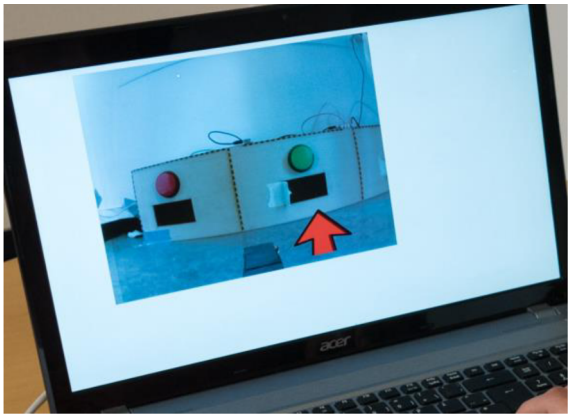

Fig. 2. Predictive display visualization. The operator has recently turned the ROV to the right, and as a result the video has moved to the left. The red arrow has not moved and works as an indication of where the ROV will be heading when the video feed has caught up with the time delay.

manipulators. The 3D environment can be constructed a priori, and exact location of the robot arm is known (Ricks et al., 2004). A limitation arises when tasks are performed in unknown and unstructured areas, and since environment geometry is unknown real time mapping and rendering can be difficult. Additional hardware may be required and calculations can become computationally intensive. Moreover, additional challenges, such as oscillopsia occur when latency is introduced in VR head-mounted displays (Allison et al., 2001).

Video manipulation does not require 3D information about the environment. It alters the delayed video feed to mimic movements and environment in real time. A simple example would be to zoom into the image if the robot is moving forward. Matheson et al. (2013) halved task completion time at a latency of three seconds in an ROV experiment using this method, by cropping and projecting the image. A similar result is obtained by capturing a wide FOV video, possibly 360 degrees, and then only displaying a section of that image to the operator. The section can be moved around in the video as a response to steering commands and thus provide fluid and seemingly real time feedback (Baldwin et al., 1999). Advantageous to video manipulation techniques are low cost, ease of implementation and not requiring a structured environment. Furthermore, prediction error propagation cannot occur since the presented video feed consists only of alterations to the latest image. However, it cannot recreate parallax movement (such as passing an object or corner) which 3D graphics models can achieve.

3. Method - Experiment investigating a predictive display under three conditions

An experiment was set up to investigate changes in human operator performance and workload when operating an ROV under three distinct conditions, each condition with a distinct display and latency. The participants were presented with a single, simple navigational task, framed as a peg-in-hole game which was the same for all three conditions. The conditions were C1. Latency, C2. Latency with Predictive Display (w/PD) and C3. Baseline, and they are described in detail in this section. First, this section describes development and implementation of a predictive display Then, the experiment design follow, which includes research objective, hypotheses, stimuli (description of task and conditions), data collection (objective performance and subjective workload), setup, experimental procedure, and data analysis.

3.1. Predictive display development

Predictive technology that reconstructs a 3D environment based on sensory data requires advanced algorithms, potentially expensive equipment, and extensive information regarding environment and ROV. In cases where this is not a possibility video manipulation provides simple and inexpensive means to increase human operators' performance.

The projected display by Matheson et al. (2013) is the simpler video manipulation method of the ones considered in Table 2, while retaining a great increase in human operator performance. However, information on the vehicles' ground trajectory is required to calculate changes in perspective. By disregarding the effects of change in perspective and applying positional and scale transformations to the video feed we obtain an even simpler approach. As such, by applying positional and scale transformations to the video feed we developed a predictive display based on image transformation. The predictive display can be applied to several robot configurations though it was developed for ROVs initially; It is appropriate only for screen-based systems and other alternatives are needed for predictive head-mounted display systems.

3.1.1. Predictive display implementation in detail

The developed predictive display repositions the delayed video feed on the monitor so objects in the video feed appear in correct size and position on the screen as if there was no latency (see Figs. 1 and 2). It uses user input (i.e. steering commands) and predefined ROV speed to predict how the FOV would move in the scene, repositioning and scaling the video feed accordingly.

The positional transformation can be explained by considering an ROV with an onboard camera rotating about its center of mass, turning with an angular velocity of ω°/s. The camera FOV is φ°, with horizontal resolution R_h pixels. A counterclockwise rotation for Δt s moves the ROV $\Delta\theta^\circ$. Objects in the video feed moves $(R_h \cdot \omega/\varphi) \cdot \Delta t = \eta \cdot \Delta t = \Delta P_h$ pixels to the right. η = *pixel turn rate*, which depends on screen resolution, angular velocity and camera FOV. The pixel turn rate, user input and total system delay t_d is used to create the predictive display. The video feeds' position on the monitor is calculated at a set interval *dt* (preferably at a minimum video frame rate (FPS)). If the ROV moves to the left, time since last update *dt* multiplies with pixel turn rate to find change in horizontal video position ΔP_h. The video feed then moves ΔP_h to the right on the monitor. When a time t_d has passed (system delay has caught up), the video feed is moved back.

For backward and forward translation, similarly as for *pixel turn rate*, a *pixel scale rate* can be found and used to scale the video feed. For backwards and forwards ROV translation, scaling of objects depends on how close they are to the camera. An average distance is used as an

approximation. The video feed scale transformation works as the aforementioned positional transformation.

Finally, the predictive display uses a red arrow to visualize the future position as illustrated in Fig. 2.

3.2. Research objective and hypotheses

Research objective: Investigate if such a simple predictive display still increase human operators' performance and reduce workload.

Based on the research objective, we sought to test the following hypotheses:

- A simple predictive display significantly increases human operators' performance (objectively measured by task score performance, - i.e. the number of hits achieved in 90s by the participant).
- A simple predictive display significantly decreases human operators' subjective workload (subjectively measured by RTLX's six dimensions, mental demand, physical demand, temporal demand, performance, effort and frustration, evaluated on eleven-point scales (Hart, 2006)).

3.3. Stimuli – Peg-in-hole-game under three conditions

The experiment encompassed a single, navigational task, in which we measured operator performance by means of an achieved score over a fixed time period.

3.3.1. Rationale behind task selection

Chen et al. (2007) reports benefits of predictive technology to be very task dependent. A peg in hole task was selected due to its applicability in teleoperation (Lane et al., 2002). A task encompassing as much lateral navigation as possible was selected, as this is where the predictive display can provide the most help, in contrast to for example navigational tasks with longer stretches of forward motion (and the maximal velocity of the ROV would create a ceiling effect). A short timeframe of 90 seconds was chosen to reduce any learning effect that might accompany a longer maneuvering course. A fixed time period made total experiment length predictable, participants used 10 min and 56 s on average (SD 1 min and 12 s). This aided in recruiting new participants. Furthermore, time pressure in combination with score achievement made participants fully devoted to the task at hand, and we argue this led to participants performing close to the best of their ability. We further argue that a single, simple task will minimize the effect of other factors on performance, e.g. trouble understanding the task, or being highly experienced in related tasks such as gaming, driving, or other navigational tasks.

3.3.2. Task

Participants were given a modified 'peg-in-hole' task. The peg was mounted on a remotely controlled ground vehicle, and there were three rectangular holes in three rectangular boxes with accompanying LEDs. One LED would light up at a time, in random order, to which the participant was instructed to perform as many 'hits' as possible by inserting the peg in the hole within the given timeframe. Task and time given (90s) was the same for three distinct conditions. During the task, a red timer indicating remaining time was constantly visible in the screens' upper right corner.

3.3.3. Three conditions

All participants repeated the task three times, under three distinct conditions. The display provided to the participants would differ in each condition. The conditions, latency and displays were as follows:

Condition 1. Latency: 700 ms delay (250 ms inherent system delay + 450 added delay). No predictive display.
Condition 2. Latency with Predictive Display (PD): 700 ms delay (250 ms inherent system delay + 450 added delay). With predictive display.
Condition 3. Baseline: 250 ms inherent system delay.[1] No predictive display.

Throughout the paper we refer to the conditions as:

C1. Latency
C2. Latency w/PD
C3. Baseline

3.3.4. 3 × 3 Latin Square Design

The sequence of the conditions was randomized according to a 3 × 3 Latin Square Design to avoid potential order and/or learning effects. All six combinations were used. Each participant was automatically assigned to one of the combinations, ensuring equal group sizes across conditions as far as possible. Due to the number of participants recruited; three of the combinations had 10 participants, and three combinations had 9 participants.

3.4. Data collection – Performance measured objectively and workload measured subjectively

N = 58 participants were recruited to test the predictive display. We collected objective measures of human performance and subjective measures of workload. Demographic data were also collected.

3.4.1. Participants

Participants were voluntary selected from NTNU, Department of Mechanical and Industrial Engineering. Our aim was to recruit as many participants as possible within the time constraint we were working with. A total of 58 participants performed the experiment, one participant was excluded in the analysis due to incomplete information. The remaining N = 57 participants received the same information and were included in the analysis. Age ranged from 23 to 30 years (24.7 ± 1.5). There were 19 female and 38 men. We gathered level of education, how often they played video games, how often they use a computer and eye health information, which can be found in Table 3.

3.4.2. Objective performance measurements

Two performance measurements are common among experiments on predictive technology: course completion time and task score (Lu et al., 2018; Mathan et al., 1996; Matheson et al., 2013; Zhang and Li, 2016; Zheng et al., 2016). In the former, the task is to navigate through a predefined pathway with the vehicle and measuring the time necessary to complete the course. In the latter, the task typically involves aligning or aiming at a given target, assigning a score to the number of times the target was met. Using a task score as a performance measure enables a fixed time for experiments, which was desirable for us to be able to recruit more participants. The number of hits made by participants in each of the 90 s test period was used as a performance measure.

Additional objective data collected included total number of hits made in all three test periods, and number of key presses in each of the test periods.

3.4.3. Subjective workload measurements

NASA Task Load Index (TLX) is common and highly accepted for remote operation and ROV applications (Hart, 2006; Hill et al., 1992; Hu et al., 2016; Ma and Kaber, 2006; Zhang and Li, 2016), and was initially developed for experimental tasks that include cognitive and

[1] The variability of inherent system delay was repeatedly quantified (10 times) to 1 – 5 ms difference each time. The average of those 10 measurements was used as inherent system delay.

Table 3
Participant data.

Variable	Options	Frequency	Percent
Gaming	Daily	2	3.5
	Weekly	15	26.3
	Monthly	8	14.0
	Yearly	17	29.8
	Never	15	26.3
Education	Nursery school	1	1.8
	Some college credit, no degree	38	66.7
	Bachelor's degree	10	17.5
	Master's degree	8	14.0
Eye health	No visual aid	32	56.1
	Spectacles	4	7.0
	Contact lenses	10	17.5
	Both spectacle and contact lenses	11	19.3

manual control tasks, and supervisory control tasks (Hart and Staveland, 1988b). TLX is multidimensional, provide good diagnostic properties for assessing underlying mechanisms of subjective workload, and has been shown to have high sensitivity (Hart, 2006; Hendy et al., 1993; Hill et al., 1992; Vidulich and Tsang, 1987). A modified version of TLX, Raw TLX (RTLX) was chosen to assess workload. The six dimensions (mental demand, physical demand, temporal demand, performance, effort and frustration) were rated on eleven-point scales. The weighting process in TLX consists of pairwise comparison of all six dimensions. It was not conducted, since we are not interested in the subjective importance of each dimension in a specific task, rather we're interested in comparing the subjective workload of different tasks (the three conditions). Furthermore, this weighing process consumes time, and, in this context, it was deemed more important to have a short survey, leaving more time for recruiting participants and conducting experimental runs. This modification is what is referred to as RTLX. One additional modification was made to the survey, as a pilot study of the experiment showed that a participant found it more intuitive to rate good performance with a high number. In the original survey a low value corresponds to good performance. Therefore, this metric and the corresponding description was reversed, such that a high value corresponded to good performance. After data collection, this value was reversed back for conventional analysis and reporting.

Furthermore, a question of perceived delay time was added to the survey, to investigate participants' subjective experienced latency in each individual condition and to compare the individual conditions, the latter in hopes of providing a measure of effectiveness for the predictive display in reducing the subjectively perceived latency of the system.

3.4.4. Data collection procedure

Both survey data and experiment data were recorded with the ROV computer using an SQLite database.

3.5. Setup

A 17" laptop running a 2.3GHz Intel Core i7-3610QM CPU and Windows 10 was used. The laptop screen served as monitor and the keyboard's arrow keys were used to steer the ROV. The keyboard and a remote mouse were used to answer the surveys. The ROV was running a Raspberry Pi 3 Model B+, and equipped with a forward facing Raspberry Pi Camera V2 and a wide angle lens with horizontal FOV of 76.5°. The robot was constructed using three wheels, two of them connected to a DC motor and the third a caster wheel for support (see Fig. 4). A wooden box with three holes and LEDs were used to register task performance. The distance between the holes (center to center) was D = 30 cm while the holes itself has a width of W = 10cm. This translates to a Fitts's index of difficulty of Id = log2 (2D/W) = 2.58 bits (Fitts, 1954). The robot ran eduROV[2] software, which provided an interface to control the robot, handling control commands, adding desired latency to the communication, and logging data.

3.6. Experimental procedure

After entering the experiment room, participants were shown the setup to ensure that they understood the situation and what they were tasked to do. The participant was placed in a chair at a desk with a laptop, with their back to the game (see Fig. 3). The participant would have no visual perception of the physical setup during the experiment. To ensure there was no auditory perception of the ROV, participants wore an ear protection headset. Information was given in writing on the computer screen. After giving consent to participate in the experiment, participants filled out a demographic survey. Information describing the experiment was provided; How to steer the vehicle, the task and performance measure, and the following procedure of the experiment. Each participant was automatically assigned to one of the groups corresponding to the 3 × 3 Latin Square Design. The participant would then conduct a 30s practice period followed by a 90s test period. After each block of practice and test period the participant filled out a survey of mental workload and perceived delay time. The starting position (indicated by the black mark in Figs. 3 and 4) was identical for all periods. The third block concluded the experiment and the participants were escorted out.

The participant was not informed of the fact that one of the conditions would have a predictive display, nor how it worked. To be able to take advantage of the predictive display is therefore dependent on the individual participants ability to intuitively understand the display. It was assumed that the practice period before each test would suffice in giving the participant the needed training in the display for the game. However, the questionnaire included a question of time delay, which could have influenced participant's attention to delay in the next conditions.

3.7. Analysis – Classical statistics and exploratory multivariate analysis

3.7.1. Classical statistics – Analysis of variance (ANOVA)

Subjective measurements used for analysis were collected after each condition and performance measurements were collected continuously during each condition. An analysis of variance (ANOVA) was conducted to investigate the effects of the predictive display on both subjective and objective measurements, i.e. this statistical test investigated the predetermined hypotheses. The characteristics of the data was inspected and in the case of violations of assumptions, the non-parametric alternative to one-way repeated measures ANOVA, the Friedman test was conducted. Data distribution was visually inspected using Normal Q-Q Plots for all variables and conditions. ANOVA F-test is found to be insensitive or robust (Krishnaiah, 1980; Schmider et al., 2010) to general nonnormality, and can for equal group sizes be used with confidence in most practical situations. We consider the sample size of 57 to be high, and we have continued with the analysis and when possible conducted a Friedman test for comparison purposes. Mauchly's test evaluates sphericity, an assumption which is considered difficult not to violate in practice (Weinfurt, 2000), over-detecting deviations from sphericity in large samples (Kesselman et al., 1980). Maxwell and Delaney (2003), recommend using an adjusted test, interpreting the result of using a Greenhouse-Geisser correction and thus ignoring the result of Mauchly's test. This was done here, calculating epsilon according to Greenhouse & Geisser (1959), and using it to correct the one-way repeated measures ANOVA. The Bonferroni post hoc test (Maxwell, 1980, Maxwell &Delaney 2004) was used to test all possible pairwise combinations of conditions. Statistical tests were performed using SPSS Statistics (*IBM SPSS Statistics 25*, 2017).

[2] https://github.com/trolllabs/eduROV/.

Fig. 3. Experiment setup. The participant can only see the robot through the display provided on the laptop screen, which is a first-person camera view.

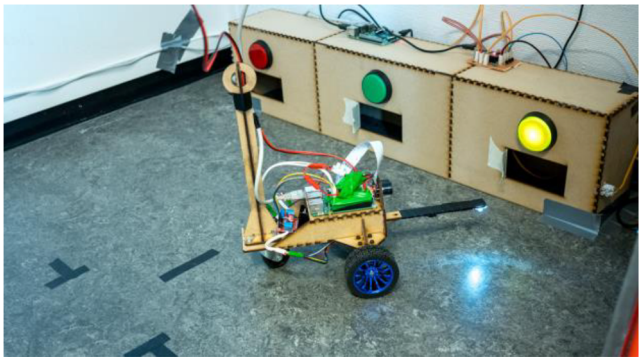

Fig. 4. Experiment setup. The three wheeled ROV with the peg mounted and the wooden box.

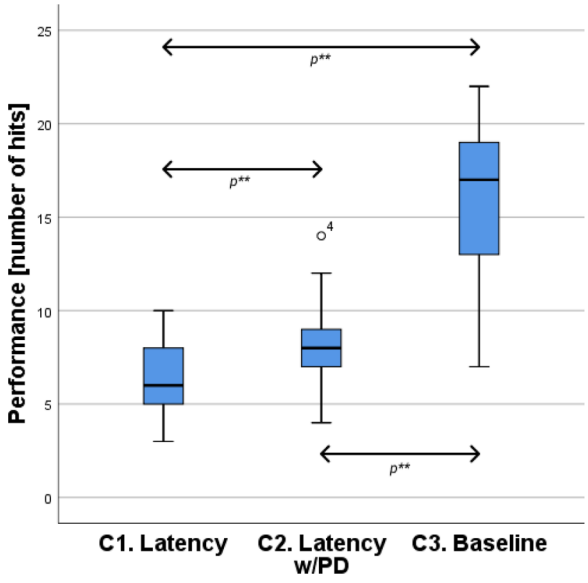

Fig. 5. Descriptive statistics of performance (objective). Original data reported. Statistically significant differences at $p < 0.01$ are indicated by p^{**}.

3.7.2. Exploratory data analysis

Furthermore, we wanted to explore and hypothesize regarding other potential relationships between the variables. Therefore, an exploratory data analysis (EDA) (Tukey, 1977), specifically a principal component analysis (PCA) was conducted to explore whether there were any interesting patters or observations in the data collected. Here, no hypothesis was determined, and all effects described emerged post-hoc. The PCA was conducted using scikit-learn (Pedregosa et al., 2011) and Jupyter Lab Notebook (Kluyver et al., 2016).

Table 4
One-way repeated measured ANOVA F-test for performance (objective).

Variable Performance	N	Outliers	Normality	Sphericity	Epsilon (ε)	F-statistic	Sig.	Sample effect size	Population effect size [c]
Performance [number of hits]	57	Yes (1)[a]	Approx. normal[b]	, $\chi2(2) = 25.7$, $p < 0.0001$	0.728	$F(1.46, 81.54) = 316.34$	$p < .0001^{**}$	$\eta^2 = 0.850$	$\omega^2 = 0.787$
	56[d]	No	Approx. normal[b]	$\chi2(2) = 26.4$, $p < 0.0001$	0.721	$F(1.44, 79.32) = 308.69$	$p < .0001^{**}$	$\eta^2 = 0.849$	$\omega^2 = 0.786$

*: $p < 0.05$, **: $p < 0.01$.
a) There was one outlier in C2. Latency w/ PD, as assessed by visual inspection of a boxplot. SPSS Statistics defines outliers as values greater than 1.5 box-plots from the edge of the plot. This value (14 hits) is genuinely unusual, we know from the experiment and the data that this participant performed above average in all three conditions. We ran the analysis both with and without outliers, reporting both results.
b) Visual inspection of Normal Q-Q Plots and histograms for all three conditions.
c) Calculated according to Wickens and Keppel (2004).
d) outlier removed.

Table 5
Pairwise comparisons of performance (objective).

Variable	C1. Latency – C2. Latency w/PD			C1. Latency – C3. Baseline			C2. Latency w/PD – C3. Baseline		
	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]
Performance [number of hits]	-1.298	0.264	$p < 0.0001^{**}$	-9.754	0.491	$p < 0.0001^{**}$	-8.456	.471	$p < 0.0001^{**}$

b: Adjustment for multiple comparisons: Bonferroni.
*: $p < 0.05$, **: $p < 0.01$.

Table 6
Pairwise comparisons subjective variables.

Variable	C1. Latency – C2. Latency w/PD			C1. Latency – C3. Baseline			C2. Latency w/PD –C3. Baseline		
	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]
Subjective Overall Workload	0.336	0.163	$p = 0.133$	1.775	0.141	$p = 0.000^{**}$	1.439*	0.177	$p = 0.000^{**}$
Mental Demand 0-10	0.158	0.235	$p = 1.000$	2.105	0.226	$p = 0.000$	1.947*	0.306	$p = 0.000^{**}$
Physical Demand 0-10[f]	0.035	0.221	$p = 1.000$	0.702	0.227	$p = 0.009$	.667*	0.211	$p = 0.008^{**}$
Temporal Demand 0-10	0.175	0.221	$p = 1.000$	0.456	0.236	$p = 0.176$	.281	0.288	$p = 1.00$
Subjective Performance 0-10	0.789	0.244	$p = 0.006^{*}$	2.825	0.240	$p = 0.000$	2.035	0.212	$p = 0.000^{**}$
Effort 0-10[f]	0.246	0.234	$p = 0.894$	1.351	0.213	$p = 0.000$	1.105	0.241	$p = 0.000^{**}$
Frustration 0-10	0.679	0.283	$p = 0.059$	3.179	0.304	$p = 0.000$	2.500	0.306	$p = 0.000^{**}$

b) Host hoc Pairwise comparisons were adjusted for Bonferroni.
f) A Friedman test with pairwise comparisons using a Bonferroni correction for multiple comparisons was carried out for comparison purposes. Results were corroborated.
*: $p < 0.05$, **: $p < 0.01$.

4. Results

The following section presents the results of the statistical tests before providing the result from the EDA.

4.1. Performance (objective)

A one-way repeated measures ANOVA was conducted to determine whether differences in human performance (the number of hits, an objective measure) between the three conditions were statistically significant. Descriptive statistics of performance data are illustrated in Fig. 5, and Table 5 shows all pairwise comparisons of the conditions. Table 4 contains the ANOVA F-test statistic, data characteristics and pretests.

Performance was statistically significant different in the three conditions, with a performance increase of 20% from C1. Latency to C2. Latency w/PD. Performance increased from $M = 6.2$ *hits* in C1. Latency, to $M = 7.5$ *hits* in C2. Latency w/PD, to $M = 16$ *hits* in C3. Baseline. There was a statistically significant increase in performance of $M = 1.3$ hits ($SD = 0.26$) from C1. Latency to C2. Latency w/PD. In summary, there was a statistically significant difference between means and, therefore, we accept the alternative hypothesis; The predictive display significantly increases performance of the human operator.

4.2. Subjective workload

This section presents the results from statistical analysis of subjective workload measures. Overall Subjective Workload is presented first, before also presenting the individual workload dimensions.

Since we conducted RTLX, the values of the individual workload dimensions (mental, physical, temporal, performance, effort and frustration) were averaged to obtain an estimate of the overall workload (Hart, 2006). This averaged score is addressed as Subjective Overall Workload in the following. Separate one-way repeated measures ANOVA was conducted for overall workload and the six individual workload dimensions to determine the effects of the predictive display on lowering subjective workload in the three conditions. The results from the ANOVA F-test, including pretests for all variables can be found in Table A2, Appendix A, whereas descriptive statistics and pairwise comparisons can be found in Table A1, Appendix A, and Table 6, respectively. The following paragraphs describe individual results before providing an overall explanation.

4.2.1. Subjective overall workload

Subjective Overall Workload was statistically significant different under the three conditions. There was a decreased subjective workload from $M = 5.3$ $(SD = 0.2)$ in C1. Latency, to $M = 4.9$ $(SD = 0.2)$ in C2. Latency w/PD, to $M = 3.5$ $(SD = 0.2)$ in C3. Baseline. Pairwise comparisons of the three conditions was carried out using the Bonferroni post hoc test, which revealed that the mean decrease in subjective workload from C1. Latency to C2. Latency w/PD was not statistically significant ($M = 0.35$, $SD = 0.16$, $p = 0.133$). There was a statistically significant mean decrease in subjective workload from C2. Latency w/PD to C3. Baseline ($M = 1.44$, $SD = 0.18$, $p < 0.001$), and from C1. Latency to C3. Baseline ($M = 1.775$, $SD = 0.14$ $p < 0.001$). A Friedman test produced corroborating results. Therefore, we cannot reject the null hypothesis and cannot accept the alternative hypothesis. The predictive display does not decrease human operators' subjective overall workload.

4.2.2. Mental demand (individual workload dimension)

Mental demand was statistically significantly different in the three conditions, however, post hoc analysis with a Bonferroni adjustment revealed that mental demand did not significantly decrease from C1. Latency to C2. Latency w/PD. There was a statistically significant decrease in mental demand from C1. Latency to C3. Baseline and from C2. Latency w/PD to C3. Baseline. The predictive display did not reduce participants' mental demand.

4.2.3. Physical demand (individual workload dimension)

Physical demand was statistically significantly different in the three conditions, however, post hoc analysis with a Bonferroni adjustment revealed that physical demand did not significantly decrease from C1. Latency to C2. Latency w/PD. A Friedman test with pairwise comparisons using a Bonferroni correction for multiple comparisons was carried out for comparison purposes, which gave the same result. There was a statistically significant decrease in physical demand from C1.

Table 7
One-way repeated measured ANOVA F-test subjective latency.

Variable	N	Outliers	Normality	Sphericity	Epsilon (ε)	F-statistic	Sig.	Sample effect size	Population effect size[c]
Subjective latency	57	Yes (13)[a] (10 unique)	Approx.[c]	Yes $\chi2(2) = 5.575$, p = 0.062	-	F(2,112) = 45.734	$p < 0.001$**	$\eta^2 = 0.450$	$\omega^2 = 0.343$
	52[b]	4 unique	Approx.[c]	Yes $\chi2(2) = 3.617$, p = 0.164	-	F(2,102) = 44.684	$p < 0.001$**	$\eta^2 = 0.467$	$\omega^2 = 0.359$

*: $p < 0.05$, **: $p < 0.01$.
a) Number of outliers in parentheses. There was no reason to exclude any outliers and so they were kept in first analysis. In the second, 5 extreme outliers were excluded, which yielded a dataset with 4 unique outliers. Further reduction did not yield a dataset without outliers. Both results are reported here.
b) 5 extreme outliers removed.
c) Visual inspection of Normal Q-Q Plots and histograms for all three conditions.

Latency to C3. Baseline, and from C2. Latency w/PD to C3. Baseline. The predictive display did not reduce participants' physical demand.

4.2.4. Temporal demand (individual workload dimension)

Temporal demand was not statistically significantly different in the three conditions, according to both ANOVA and Friedman test. The predictive display did not reduce participants' temporal demand.

4.2.5. Subjective performance (individual workload dimension)

Subjective Performance was statistically significantly different in the three conditions. Subjective Performance was evaluated at $M = 5.53$ in C1. Latency, $M = 4.74$ in C2. Latency w/PD, and $M = 2.70$ in C3. Baseline, with a low value corresponding to a performance closer to perfect. There was a statistically significant decrease of $M = 0.79$ $(SD = 0.24, p = 0.006)$ between C1. Latency and C2. Latency w/PD, a statistically significant decrease of $M = 2.83$ $(SD = 0.24, p < 0.001)$ between C1. Latency to C3. Baseline, and a statistically significant decrease of $M = 2.04$ $(SD = 0.21, p < 0.001)$ between C2. Latency w/PD to C3. Baseline. A Friedman test with a Bonferroni correction for multiple comparisons was carried out for comparison purposes, corroborating result at $p < 0.001$. The median of Subjective Performance was statistically significant different between C1. Latency ($Mdn = 5$) and C3. Baseline ($Mdn = 2$) ($p < 0.001$), statistically significant between C2. Latency w/PD ($Mdn = 5$) and C3. Baseline ($p < 0.001$), but not statistically significant different between C1. Latency condition and C2. Latency w/PD ($p < 0.132$). In addition to a statistically significant decrease from both latency conditions (C1. Latency and C2. Latency w/PD) to C3. Baseline, it is also noteworthy that the mean decrease towards C3. Baseline is greater from C1. Latency than the decrease from C2. Latency w/PD; Which means participants thought they performed better with the predictive display than without it, given different latencies, and given equal latency. In summary, the predictive display increased participants subjective performance, i.e. participants thought their performance was better with the predictive display.

4.2.6. Effort (individual workload dimension)

Effort was statistically significant different in the three conditions, however, post hoc tests with a Bonferroni adjustment revealed that there was not a statistically significant difference between C1. Latency and C2. Latency w/PD. There was a statistically significant decrease from C3. Baseline to the two latency conditions (C1. Latency and C2. Latency w/PD). A Freidman test with a Bonferroni correction for multiple comparisons corroborated these results. The predictive display did not reduce participants' effort.

4.2.7. Frustration (individual workload dimension)

Frustration was statistically significantly different in the three conditions and post hoc tests with a Bonferroni adjustment revealed that there was a statistically significant decrease in frustration from C1. Latency to C3. Baseline, as well as from C2. Latency w/PD to C3. Baseline, though not from C1. Latency to C2. Latency w/PD. The predictive display did not reduce participants' frustration.

4.2.8. Overall result for workload

The analysis of Subjective Overall Workload and the individual workload dimensions did not show a statistically significant difference between C1. Latency and C2. Latency w/PD, with the exception of the individual variable Subjective Performance, in which participants reported a statistically significant mean increase of 0.789.

The predictive display does not reduce participants' mental demand, physical demand, temporal demand, effort, nor frustration; However, the predictive display increased participants' subjective performance, i.e. participants' thought they performed better with the predictive display. In summary for subjective workload, we cannot reject the null hypothesis, i.e. the predictive display does not reduce participants subjective workload.

Table 8
Pairwise comparisons subjective latency.

Variable	C1. Latency – C2. Latency w/PD			C1. Latency – C3. Baseline			C2. Latency w/PD – C3. Baseline		
	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]
Subjective Latency [ms]	78.33	68.42	$p = 0.771$	582.21	73.14	$p = 0.000$**	503.88	55.37	$p = 0.000$**

b) Host hoc Pairwise comparisons were adjusted for Bonferroni.
*$p < 0.05$, **: $p < 0.01$.

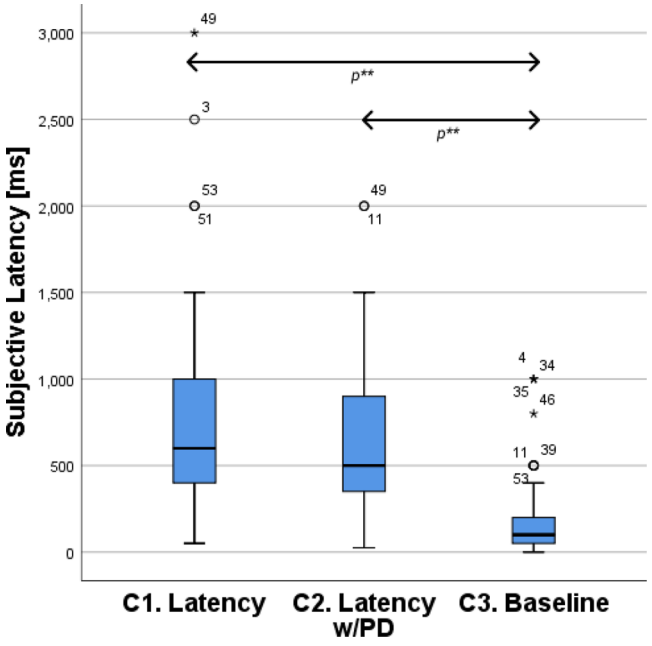

Fig. 6. Descriptive Statistics Subjective Latency. Original data reported. Statistically significant differences at $p < 0.01$ are indicated by p**.

4.3. Subjective latency

A one-way repeated measures ANOVA was conducted on Subjective Latency to determine if there was a statistically significant reduction from C1. Latency – to C2. Latency w/PD condition. Subjective latency (evaluated in ms by the participants) was statistically significantly different in the three conditions. Post hoc tests with a Bonferroni adjustment revealed that there was a statistically significant decrease in Subjective Latency from C1. Latency to C3. Baseline condition as well as from C2. Latency w/PD to C3. Baseline condition, but not from C1. Latency to C2. Latency w/PD. A Friedman test corroborated these results, (Tables 7 and 8).

There were multiple outliers for this measure, and we reran the analysis with the extreme outliers removed, which only slightly increased effect size, however it did not change the overall result. The predictive display did not reduce participants' estimation of latency when comparing C1. Latency and C2. Latency w/PD, see Fig. 6.

4.4. Gamers

Those who play games weekly or more often were defined as gamers. The potential increased performance gain, measured objectively, by gamers is investigated here. A two-way mixed ANOVA was conducted, comparing the mean differences of performance (objective) between two independent groups, Gamers and Non-Gamers, under the three conditions. Descriptive statistics can be found in Table 9 and

Table 9
Descriptive statistics gamer vs non-gamer. Original data reported.

Variable	Group	N	C1. Latency		C2. Latency w/PD		C3. Baseline	
			Mean	SD	Mean	SD	Mean	SD
Performance [number of hits]	Gamer	17	6.47	0.42	8.41	0.46	17.71	0.92
	Non-Gamer	40	6.08	0.27	7.10	0.30	15.20	0.60

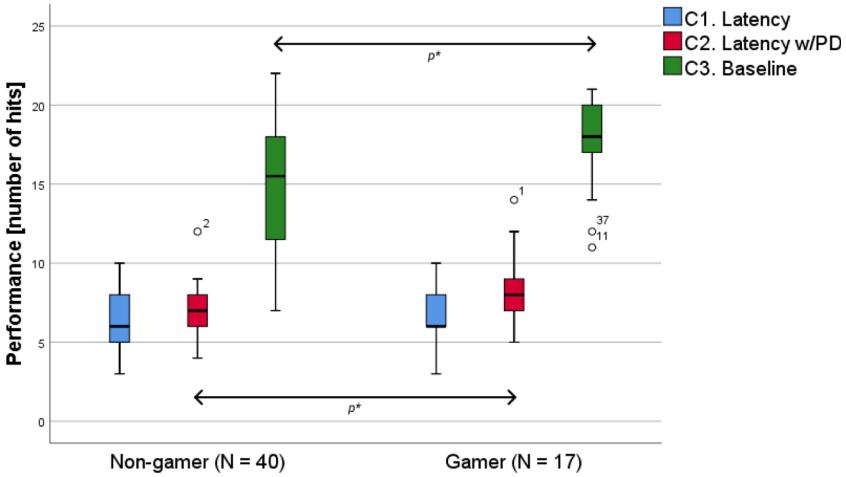

Fig. 7. Performance of Gamers vs. Non-gamer. Original data reported. Statistically significant differences at $p < 0.01$ are indicated by p**.

Table 10
Two-way mixed ANOVA F-test on performance (objective) for gamers vs. non-gamer.

Variable	N	Outliers	Normality	Homogeneity	Sphericity	Epsilon	F-statistic	Sig.	Sample effect size	Population effect size [g]
Performance[a] [number of hits]	57	Yes (4)[bc]	Yes[d]	Yes[e]	No $\chi^2(2) = 24.895$, p = .000	$\varepsilon = 0.728$	F(1.46, 80.32) = 2.72[f]	$p = 0.088$	$\eta^2 = 0.047$	$\omega^2 = 0.02$
	53[h]	No	Yes[d]	Yes[e]	No $\chi^2(2) = 21.406$, p = .000	$\varepsilon = 0.742$	F(1.48, 75.65) = 5.769[f]	$p = 0.009$**	$\eta^2 = 0.102$	$\omega^2 = 0.057$

*: $p < 0.05$, **: $p < 0.01$.
a) Performance was separated for the two independent groups.
b) Assessed by visual inspection of a boxplot. SPSS Statistics defines outliers as values greater than 1.5 box-plots from the edge of the plot. The 4 outliers were kept in the first analysis as there was no reason to exclude them. In the second, they were excluded. Both results are reported here.
c) By examination of studentized residuals for values greater than ± 3, one outlier was found with a studentized residual value of 3.04. The outlier was kept in the subsequent analysis since its value is close to the threshold and as there was no reason to exclude it.
d) Visual inspection of Normal Q-Q Plots of the distribution and the distribution of studentized residuals for all three conditions.
e) Levene's test assessed homogeneity of variance, and Box's test evaluated homogeneity of covariances.
f) A Greenhause Geisser correction was applied.
g) Calculated according to Wickens and Keppel (2004).
h) outliers removed.

Fig. 7. The ANOVA F-test can be found in Table 10.

The interaction between gaming experience and conditions on performance had a level of significance of $p = 0.088$. Univariate post hoc tests indicated that there was not a statistically significant difference between gamers ($M = 6.5$ *hits*) and non-gamers ($M = 6.1$ *hits*) in C1. Latency ($F(1,55) = 0.622$, $p = 0.43$, sample effect size $\eta^2 = 0.01$). However, there was a significant increase in performance for gamers in C2. Latency w/PD ($F(1,55) = 5.71$, $p = 0.02$, sample effect size $\eta^2 = 0.094$), in which gamers had $M = 8.4$ *hits*, whereas non-gamers had $M = 7.1$ *hits*. Furthermore, there was a significant increase in performance for gamers in C3. Baseline ($F(1,55) = 5.203$, $p = 0.026$, sample effect size $\eta^2 = 0.086$), in which gamers had $M = 17.7$ *hits*, whereas non-gamers had $M = 15.2$ *hits, (Tables 10 and 11)*.

When considering the two independent groups (Gamer, Non-Gamer), there was a significant main effect of gaming ($F(1,55) = 6.311$, $p = 0.015$, sample size effect $\eta^2 = 0.103$), with gamers performing better than non-gamers. Gamers performed on average $M = 10.9$ *hits*, which is $M = 1.4$ *hits* ($SD = 0.6$) above the performance of non-games with $M = 9.5$ *hits*.

The analysis was also conducted without outliers (see Table 11), which yielded more than a doubling of effect size (sample effect size $\eta^2 = 0.102$ and population effect size $\omega^2 = 0.057$), and a lower p value (p = 0.009), which means that the interaction between gaming experience and conditions on objective performance reached statistical significance. Univariate post hoc tests (on the pruned dataset) indicated a statistically insignificant difference between gamers ($M = 6.3$ *hits*) and non-gamers ($M = 6.0$ *hits*) in C1. Latency ($F(1,51) = 0.240$, $p = 0.63$, sample effect size $\eta^2 = 0.005$), and a statistically insignificant difference increase in performance for gamers in C2. Latency w/PD ($F(1,51) = 3.52$, $p = 0.066$, sample effect size $\eta^2 = 0.065$), in which gamers had a $M = 7.9$ *hits*, whereas non-gamers had $M = 6.0$ *hits*. There was a significant increase in performance for gamers in C3. Baseline ($F(1,51) = 8.249$, $p = 0.006$, sample effect size $\eta^2 = 0.1.39$), in which gamers achieved $M = 18.4$ *hits*, whereas non-gamers had $M = 15.1$ *hits*. When considering the two independent groups (Gamer, Non-Gamer), there was a significant main effect of gaming ($F(1,51) = 6.929$, $p = 0.011$, sample size effect $\eta^2 = 0.12$), with gamers performing better than non-gamers. Gamers performed on average $M = 10.9$ hits, which is $M = 1.5$ *hits* ($SD = 0.6$) above the performance of non-games with $M = 9.4$ *hits*.

Gamers performed better than non-gamers on average.

4.5. Exploratory data analysis - PCA

A principal component analysis (PCA) was conducted to explore whether there were any interesting patters or observations in the data collected. We had no predetermined hypothesis, and all effects described in this section emerged post-hoc.

A total of 35 variables collected during the experiment were standardized (removing the mean and scaling to unit variance) and used in the PCA. Fig. 8 shows a Scree plot of the Principal Components (PCs) eigenvalues. The first 10 eigenvalues are larger than 1, the first 5 have an eigenvalue above 2, the first two are greater than 4 and the first eigenvalue is greater than 7. Fig. 9 shows the cumulative sum of explained variance, which did not have a clear 'elbow-shape', however the first 7-10 PCs retains 72.1% – 82.8%[3] of the variance of the original data.

A Score plot, and a Loading plot of the two first Principal Components (PCs) can be found in Figs. 10 and 11, and 12 respectively. The first two PCs explains 22.6% and 13.6% of the total variance. The following result emerged post-hoc, and thus interpretations made accordingly.

4.5.1. Interpreting the Score plot

Fig. 11. Score Plot of the first two principal components with gaming

[3] Accurate percentage of explained variance retrieved from data, and not estimated from graph.

Table 11
Pairwise comparisons of differences in performance (objective) for gamer vs. non-gamer.

Variable	C1. Latency			C2. Latency w/PD			C3. Baseline			
Performance [number of hits]	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]	Mean Diff.	SD	Sig.[b]	
Gamer – Non-Gamer (outliers removed)	57	.396	.501	*p = 0.434*	1.312	.549	*p = 0.020**	2.506	1.099	*p = 0.026**
	53	0.260	0.53	*p = 0.626*	0.954	0.509	*p = 0.066*	3.300	1.149	*p = 0.006***

* $p < 0.05$, **: $p < 0.01$.
a) outliers removed.
b) Host hoc Pairwise comparisons were adjusted for Bonferroni.

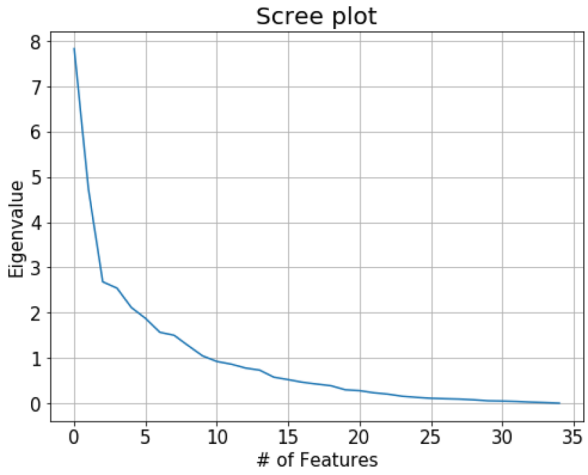

Fig. 8. Scree plot.

experience.Fig. 10. shows the Score Plot of PC1 and PC2 with legends indicating gender, from which we can see that PC2 tend to separate woman and men quite accurately. The woman cluster in the positive range of PC2 and the men in the negative range, with only a few datapoints crossing zero. Fig. 11 show the Score plot with gaming experience, in which we observe subtle trends in the scatterplot based on gaming experience; Those who never gamed predominantly resides in the positive range of PC2; Furthermore those who games more often tended to cluster in the negative range of PC2. When viewing Figs. 10 and 11 simultaneously we observe that the women in this experiment typically gamed yearly or never, with two exceptions of woman gaming on a monthly basis. Men gamed most often, typically yearly, monthly, weekly and two participants daily.

4.5.2. Interpreting the Loading plot

From the loading plot in Fig. 11, we see that the total hits in each of the conditions seem to be correlated as they cluster together. Total hits in each condition (Total hits C1 – C3) and total hits for all conditions combined

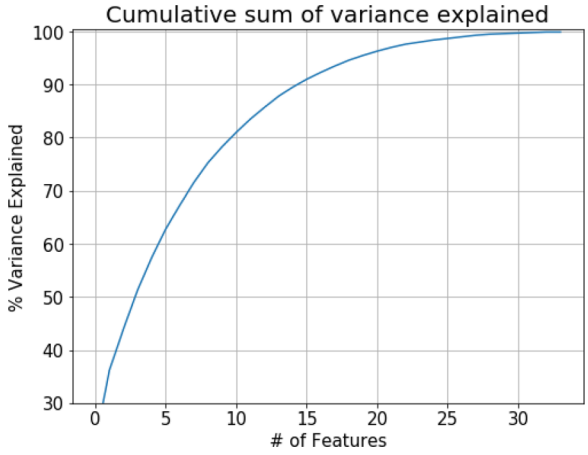

Fig. 9. Cumulative sum of variance explained.

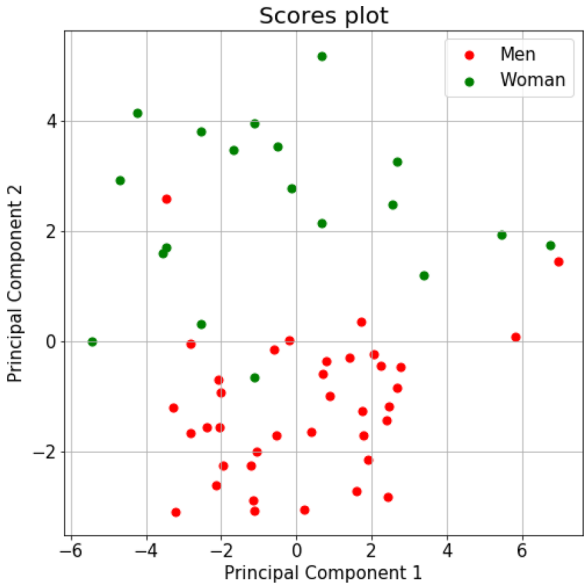

Fig. 10. Score Plot of the first two principal components gender.

(Total hits C1 + C2 + C3) cluster together, as does subjective performance (Performance C1 – C3). We observe that eye health and computer usage have a loading close to zero, thus not contributing to the definition of the principal components, and unimportant for defining the direction of some underlying latent variable. All participants used a computer daily, thus this variable had the same value across the participant population. The eye health, level of education, key strokes in C1. Latency (Key strokes C1) and C2. Latency w/PD (Key strokes C2), and age also have a loading

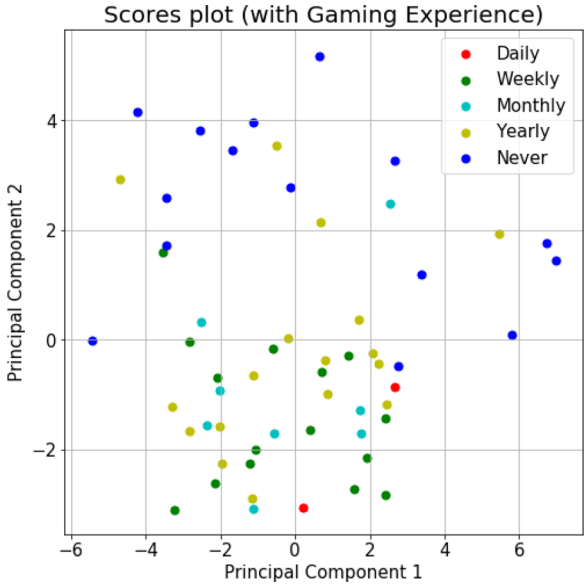

Fig. 11. Score Plot of the first two principal components with gaming experience.

close to zero and are less important for the model.

The subjective performance is negatively correlated with the other TLX dimensions, especially noteworthy is the opposite positions of subjective performance (Performance C1 – C3) and frustration (Frustration C1 – C3). The hits in each condition, and the total hits for all three conditions are clustered together and are therefore correlated. Furthermore, we see that gender and gaming experience have a high loading on PC2, thus contributing greatly to defining PC2 (that gender contributed to PC2 we also knew from the scores plot), and that they are positioned quite close together in comparison to the other variables and therefore are correlated. Among these participants men tended to game more than woman, which is also reflected when looking at the raw data.

5. Discussion

5.1. Performance

The results show that there is a statistical difference in performance when controlling the ROV without and with the help of the predictive display. Subjects performed on average 20% better with a sample effect size $\eta^2 = 0.850$, and population effect size of $\omega^2 = 0.787$. This can be categorized as a medium to large effect (Kirk, 2013), especially when considering the simplicity and low cost of implementing the predictive display. Previous research describes a wide range (8% to 65%) of task time reduction from predictive technology. A direct comparison to any specific

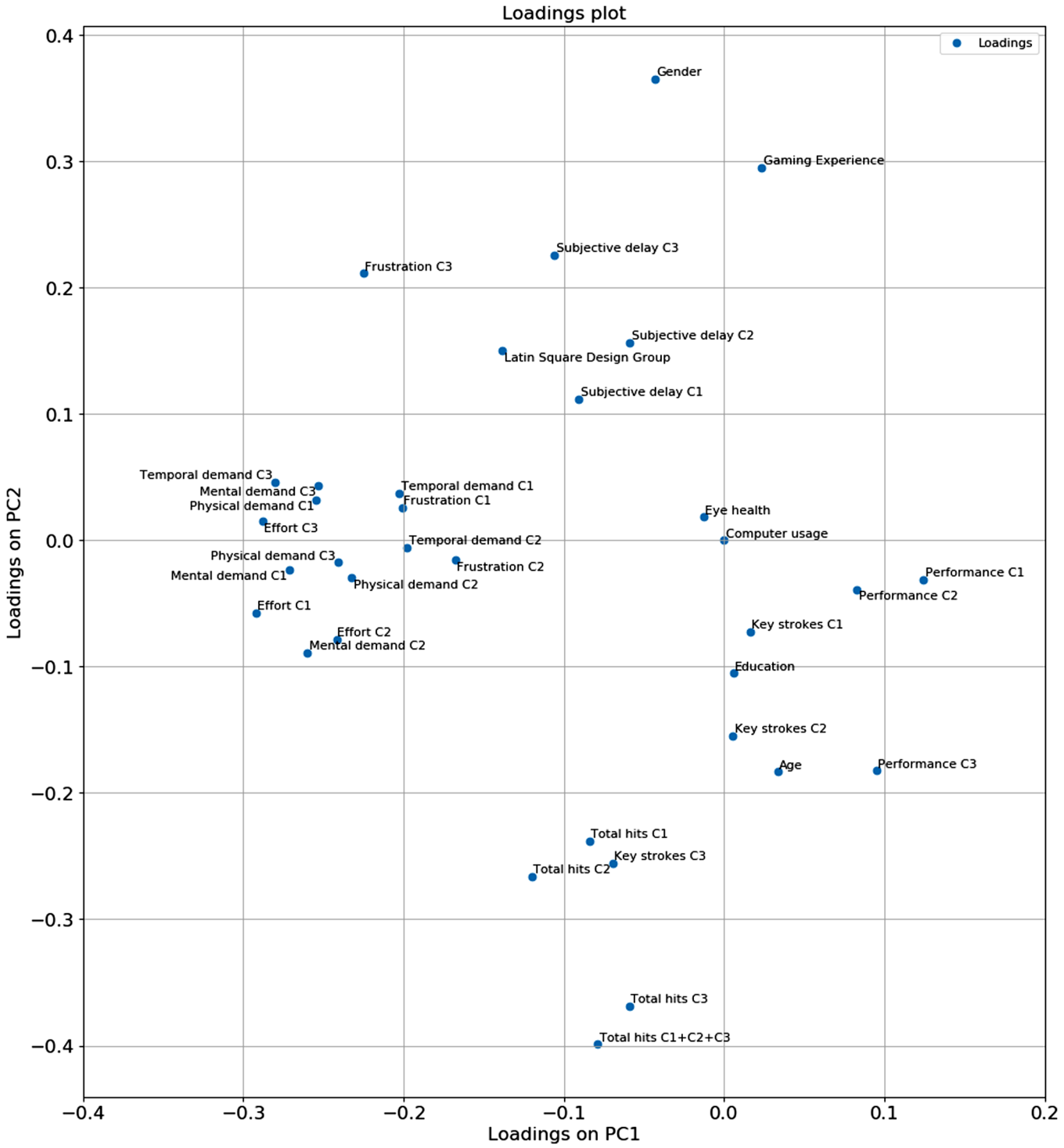

Fig. 12. Loadings plot.

experiment is challenging, however a performance increase of 20% in this experiment is probably in the lower range of what was found in the other experiments in Table 1. The task time reduction measure is considered to be comparable to the performance gain measured in this experiment. However, the predictive method used here is the simplest solution to implement at the lowest cost. Moreover, the participants only had 30 seconds to intuitively learn and train in using the predictive display, since none of the participants were told that there would be a predictive display nor how it worked. Some immediately understood what the predictive display was trying to tell them, others did not understand that there had been a predictive display until the experiment was over. The ones who tried to use the predictive display the way it was intended typically performed better than those who did not use it. It may be that performance could have been improved more if participants were informed about the predictive display's functionality.

As expected, participants performed significantly better in the baseline condition, in which there was only 250 ms latency. This latency is well above what human perception is able to pick up on, which most participants did. As discussed under 5.4, most participant underestimated the latency in the third condition reporting barely above 0 ms.

5.2. Subjective workload

Subjects reported minimal differences between C3. Baseline, C1. Latency and C2. Latency w/PD. There was a statistically significant difference in subjective overall workload between the three conditions, however Bonferroni post hoc tests revealed that differences between C1. Latency and C2. Latency w/PD was not statistically significant. Therefore, we cannot say that the predictive display reduces subjective workload. The only significant difference was found in subjective performance, in which participants felt that they on average performed 14% better when using the predictive display. The actual performance increase was 20%. They also reported that they felt 11% less frustrated using the PD, though this is not statistically significant. Participants also stated that C3. Baseline was better in all metrics, with an exception of temporal demand where the difference was not significant.

Participants reported no significant difference in mental, physical and temporal demand between C1. Latency and C2. Latency w/PD. We consider these three metrics to be a good description of the total subjective workload in this experiment setup. Some participants, especially those who did not understand what the predictive display was trying to tell them, even reported it as distracting. Due to the predictive display's functionality, the video feed is constantly moving around and scaling up and down. This can understandably be distracting. Some participants immediately understood how the predictive display worked, and they typically reported the predictive display as helpful. To the experimenter they also seemed to be more relaxed, however there are no recorded data illustrating this. During the task, a red timer indicating the remaining time was constantly visible for participants to see in the upper right corner. In addition, the ROV had rapid acceleration and was able move fast if the operator managed to do so. Overall, this made for a hectic and exiting experience for the subjects. This may explain why there is no significant change in the temporal demand, even compared to C3. Baseline. The fact that the participants reported a better value (smaller) in the other five metrics for the no delay condition, is as expected. The experimenter also observed a tendency of participants performing correcting steering commands, causing the ROV to oscillate greatly before hitting or missing the target, which corroborates prior research (Appelqvist et al., 2007). This was particularly prominent in the C1. Latency condition, again illustrating the detrimental effect of latency on both human performance and behavior. These findings support earlier research describing how video latency negatively affects the user experience in teleoperation.

5.3. Gaming

The gamers performed 30% better with the predictive display, while non-gamers performed 17% better. Interestingly the gamers increased their score almost twice as much as non-gamers when shifting from C1. Latency to C2. Latency w/PD, though the exact reason for this is unclear. The arrow in the predictive display acts as an aiming device, which could be a more familiar concept for gamers. This finding could also indicate that gamers are more used to having to adapt to unfamiliar setting and interfaces in a computer competing context. Furthermore, when comparing the scores of gamers and non-gamers, it is interesting to note that gamers only performed better than non-gamers in C2. Latency w/PD and C3. Baseline, but not in C1. Latency. This could indicate that the amount of experience may not be crucial for obtaining a high score (equal to high performance) in a situation with considerable latency. Thought post hoc tests on the pruned dataset were not statistically significant at $p<0.05$ in C2. Latency w/PD, the level of significance $p = 0.066$ was close to that threshold. A level of statistical significance may have been achieved with additional participants conducting the experiment, and equal group sizes, as both may have a large effect on p-values (Krishnaiah, 1980). In both analysis, there was a significant main effect of gaming, meaning gamers performed better on average. More interesting is the population effect size, which increased from $\omega^2 = 0.02$, a small association to $\omega^2 = 0.057$, a medium association (Kirk, 2013), which means that the effect of gaming, and the ability gamers had to take advantage of the PD, reaches some practical significance. Taken together, we interpret this to mean gamers were better able to take advantage of the predictive display to increase objective performance.

We observe that the combination of predictive display and related training (in the form of playing similar games at least once a week) results in twice a performance gain compared to only predicative display. In this experiment participants were not informed of the predictive display's functionality, which leads us to consider what the performance gain might have been if participants' were aware of the functionality a priori and if they received training in using the predictive display. Simultaneously considering an increased effect size when removing outliers, i.e. a stronger result, leads us to believe that a greater performance gain might have been the result of specialized training prior to the experiment. Therefore, we hypothesize that the combination of predictive display and extensive training produces a greater increase in performance. Research corroborates this; A priori gaming experience have been found to relate to performance in desktop and immersive virtual environments (Richardson et al., 2011), and video gaming suggested a s a training regimen to increase processing speed, which contributes to increased cognitive performance (Dye et al., 2009). Moreover, studies investigating causality supports action video gaming as a training method (Dye et al., 2009; Green & Bavelier, 2003; Richardson et al., 2011). Generally, we hypothesize that assistive technology in combination with (potentially minimal) training produces high performance gain (output). When compared to the necessary implementation of technology and training (input), we consider this a good trade-off between input and output.

5.4. Subjective latency

About 75% of the participants underestimated the latency in the third condition. Many of them barely reported over 0 ms, but the actual latency was 250 ms. These findings support previous research, which states that smaller latencies closer to zero is difficult to differentiate from no latency. Questioning participants about latency could have influenced their attention to latency in the forthcoming conditions. However, the randomized Latin Square Design of conditions should account for any order effects caused by this question. Furthermore, this question was primarily included to investigate whether participants experienced lower latency with the aid of the predictive display when comparing conditions with equal latencies, which was not the case. The predictive display did not decrease the subjectively experienced latency for participants in this experiment.

5.5. Exploratory data analysis discussion

Effects discussed here emerged post hoc; Thus, is interesting to see

effects of gender and gaming experience show up in the PCA, since there are known effects of both. From the scores plots (Figs. 10 and 11) we see that PC2 separates women and men quite accurately with a few exceptions. Furthermore, PC2 tends to separate participants by their gaming experience, and by combining the loadings plot (Fig. 11) and scores plot (Figs. 10 and 11) we observe that the male participants, the exceptions in the upper regions of PC2, never gamed. When further investigating the loadings plot (Fig. 11) and scores plots (Figs. 10 and 11) simultaneously we see that gender and gaming both had high loadings on PC2, thus contributing to PC2. In the loadings plot (Fig. 11) we see participants objective performance (Total hits C1 – C3) having a high negative loading, which means it also contributes to the definition to PC2. Males are generally more experienced in gaming (Richardson et al., 2011), and in both studies investigated by Richardson et al. (2011) high gaming experience was related to higher task performance. Video gaming involves several spatial and cognitive abilities, and studies investigating causation show that gaming experience can improve mental rotation and visual attention (Moffat et al., 1998; Richardson et al., 2011). For instance, performance in visual search tasks, visual attention, visual memory, contrast sensitivity, and judging relative velocity have all been shown to improve with gaming experience (Dye et al., 2009; Moffat et al., 1998; Richardson et al., 2011). Performance in dynamic spatial tasks that required reasoning about moving stimuli (e.g. tracking objects) also improved (Richardson et al., 2011); And all those abilities are important for a high objective performance (Total hits C1 – C3) in this experiment. When specifically considering spatial abilities, there are known gender differences, including visuospatial abilities such as spatial orientation and spatial visualization (Moffat et al., 1998). Males outperform females in spatial performance tasks; In particular when it involves mental rotations, whether that task is paper-and-pencil (manipulations and transformations of geometric figures and forms) or in a virtual environment (Moffat et al., 1998; Richardson et al., 2011). Since males generally have more gaming experience than females and video game experience influence visuospatial processes, this might further contribute to gender differences in spatial tasks (Richardson et al., 2011), and moreover the objective performance (Total hits C1 – C3) in this experiment. In fact, females and males with similar levels of gaming experience did not differ in dynamic spatial ability, and gender differences were eliminated when gaming experience was included as a covariate (Richardson et al., 2011). Since the females in our experiment generally had less gaming experience, and those who did tended to cluster towards the male gamers, and since non-gaming males tended to cluster towards the females, we therefore identify an effect of gaming experience. We do recognize the high collinearity between gender and gaming experience, both had a high loading on PC2 (Fig. 11); However, further analysis is needed to examine what exactly separates the data here. Still, PC2 consists mainly of objective measures, e.g. gender, gaming experience, and objective performance (Total hits C1 – C3). For PC1, we have high loadings on individual workload dimensions (which are subjective), in which all are correlated except for subjective performance, and so they contribute to the definition of PC1. In summary, PC1 consists mainly of subjective variables from surveys, whereas PC2 consists mainly of objective variables collected in the experiment.

6. Conclusion – An increase in human performance

This work investigated human operators' performance and their subjectively experienced workload in a teleoperation context when using a predictive display. Human operator performance decrease and workload increase as latency is introduced in teleoperation, but there exist several approaches to combat these detrimental effects; One of which is predictive technology. A predictive display based on image transformation was developed by applying positional and scale transformations to the video feed and tested experimentally. An experiment was set up to test the predictive display and investigate changes in human operator performance and workload when operating an ROV. N = 57 participants conducted a simple navigational task (peg-in-hole game), under three conditions: C1. Latency, C2. Latency with predictive display and C3. Baseline. ANOVAs showed a statistically significant increase of 20% in human performance with the aid of the predictive display. Differences in overall subjective workload was not statistically significant, except for with subjective performance where participants felt they performed better with the predictive display. Gaming experience was advantageous, in fact gamers increased their score with almost twice as much as non-gamers. An exploratory data analysis (EDA) investigated confounding factors with confirmatory results.

CRediT authorship contribution statement

Henrikke Dybvik: Writing - original draft, Writing - review & editing, Visualization, Formal analysis, Validation. **Martin Løland:** Software, Investigation, Data curation, Formal analysis, Conceptualization, Methodology. **Achim Gerstenberg:** Supervision. **Kristoffer Bjørnerud Slåttsveen:** Supervision. **Martin Steinert:** Supervision.

Declaration of Competing Interest

None.

Acknowledgements

This research is supported by strategic funds from the department for Mechanical and Industrial Engineering (MTP) at NTNU.

Appendix A

Tables A1, A2.

Table A1
Descriptive statistics subjective variables.

Variable	C1. Latency		C2. Latency w/PD		C3. Baseline	
	Mean	SD	Mean	SD	Mean	SD
Subjective Overall Workload	5.263	0.197	4.927	0.192	3.488	0.193
Mental Demand 0-10	5.667	0.273	5.509	0.301	3.561	0.271
Physical Demand 0-10	2.877	0.285	2.842	0.293	2.175	0.245
Temporal Demand 0-10	5.842	0.277	5.667	0.280	5.386	0.307
Subjective Performance 0-10[1)]	5.526	0.307	4.737	0.274	2.702	0.214
Effort 0-10	6.018	0.260	5.772	0.266	4.667	0.278
Frustration 0-10	5.625	0.319	4.946	0.275	2.446	0.243

Table A2
One-way repeated measured ANOVA F-test.

Variable	N	Outliers[a, e]	Normality	Sphericity	Epsilon (ε)	F-statistic	Sig.	Sample effect size	Population effect size[c]
Subjective Overall Workload	57	Yes (1)	Yes	Yes $\chi^2(2) = 3.787$, p = 0.151	-	$F(2, 112) = 68.322$	p< 0.001**	$\eta^2 = 0.55$	$\omega^2 = 0.441$
	56[i]	No	Yes	Yes $\chi^2(2) = 3.999$, p = 0.135	-	$F(2, 110) = 68.311$	p< 0.001**	$\eta^2 = 0.55$	$\omega^2 = 0.445$
Mental demand	57	No	Yes[d]	No $\chi^2(2) = 9.962198$, p = 0.007	ε = 0.858[b]	F(1.716, 96.082) = 41.286	p<0.001**	0.424	$\omega^2 = 0.32$
Physical demand	57	Yes (1) [f]	Yes[d]	Yes χ2(2) = 0.357, p = 0.837.	-	F(2, 112) = 6.474	p = 0.002**	$\eta^2 = 0.104$	$\omega^2 = 0.060$
	56[i]	No	Yes[d]	Yes χ2(2) = 1.198, p = 0.549.	-	F(2, 110) = 5.601	p = 0.005**	$\eta^2 = 0.092$	$\omega^2 = 0.052$
Temporal demand	57	Yes (2) [f]	Yes[d]	No χ2(2) = 6.498, p = 0.039.	ε = 0.900[b]	F(1.799, 100.771) = 1.690	p = 0.192	$\eta^2 = 0.029$	$\omega^2 = 0.008$
	55[i]	No	Yes[d]	No χ2(2) = 6.504, p = 0.039.	ε = 0.896	F(1.793, 96.819) = 1.686	p = 0.193	$\eta^2 = 0.030$	$\omega^2 = 0.008$
Subjective performance	57	Yes (2) [f]	Yes[d]	Yes χ2(2) = 1.552, p = 0.460.	-	F(2, 112) = 78.578	$p < 0.001$**	$\eta^2 = 0.584$	$\omega^2 = 0.476$
	55[i]	Yes (3)[g]	Yes[d]	Yes χ2(2) = 1.972, p = 0.373.	-	F(2, 108) = 81.030	$p < 0.001$**	$\eta^2 = 0.600$	$\omega^2 = 0.492$
Effort	57	Yes (1) [f]	Yes[d]	Yes χ2(2) = 1.143, p = 0.565	-	$F(2, 112) = 19.641$	$p < 0.001$**	$\eta^2 = 0.260$	$\omega^2 = 0.179$
	56[i]	No	Yes[d]	Yes χ2(2) = 1.277, p = 0.528	-	$F(2, 110) = 18.627$	$p < 0.001$**	$\eta^2 = 0.253$	$\omega^2 = 0.173$
Frustration	56	No[h]	Yes[d]	Yes χ2(2) = 0.519, p = 0.771	-	F(2,112) = 63.275	$p < 0.001$**	$\eta^2 = 0.535$	$\omega^2 = 0.426$

*:$p < 0.05$, **: $p < 0.01$.
a) Number of outliers in parentheses.
b) A Greenhause Geisser correction was applied.
c) Calculated according to Wickens and Keppel (2004).
d) Visual inspection of Normal Q-Q Plots and histograms for all three conditions.
e) Visual inspection of a boxplot.
f) Outliers was kept in the first ANOVA as there was no reason for excluding them and a Friedman test with pairwise comparisons using a Bonferroni correction was carried out for comparison purposes, as this test is less affected by outliers. Results were corroborated. We also reran the analysis with outliers excluded, which resulted in somewhat higher effect size. The overall result was the same.
g) Excluding initial outliers did not yield a dataset without outliers. Further outlier removal was not conducted to avoid constructing a highly reduced, and thus unrepresentative dataset.
h) There were three outliers in the sample with N = 57. One outlier in C2. Latency w/PD, reported the highest frustration while feeling like their performed the worst with PD. We assume this was due to not understanding what PD was trying to do and thus we removed this participant from the analysis of Frustration. The two outliers in C3. Baseline reported high frustration in all three conditions and were therefore kept. When excluding the abovementioned participant and rerunning the analysis, there were not outliers in the data. N = 56 data points were used for this specific analysis.
i) outlier removed.

References

Allison, R.S., Harris, L.R., Jenkin, M., Jasiobedzka, U., Zacher, J.E., 2001. Tolerance of temporal delay in virtual environments. Proc. IEEE Virtual Reality 2001, 247–254. https://doi.org/10.1109/VR.2001.913793.

Ando, N., Lee, J.-H., Hashimoto, H., 1999. A study on influence of time delay in tele-operation—quantitative evaluation on time perception and operability of human operator. In: IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics. 5. pp. 1111–1116. https://doi.org/10.1109/ICSMC.1999.815712. Cat. No.99CH37028), 5.

Appelqvist, P., Knuuttila, J., Ahtiainen, J., 2007. Development of an Unmanned Ground Vehicle for task-oriented operation—considerations on teleoperation and delay. In: 2007 IEEE/ASME International Conference on Advanced Intelligent Mechatronics, pp. 1–6. https://doi.org/10.1109/AIM.2007.4412567.

Baldwin, J., Basu, A., Zhang, H., 1999. Panoramic video with predictive windows for telepresence applications. In: Proceedings 1999 IEEE International Conference on Robotics and Automation. 3. pp. 1922–1927. https://doi.org/10.1109/ROBOT.1999.770389. Cat. No. 99CH36288C), 3.

Bejczy, A.K., Kim, W.S., Venema, S.C., 1990. The phantom robot: predictive displays for teleoperation with time delay. In: IEEE International Conference on Robotics and Automation Proceedings. 1. pp. 546–551. https://doi.org/10.1109/ROBOT.1990.126037.

Chen, J.Y.C., Haas, E.C., Barnes, M.J., 2007. Human performance issues and user interface design for teleoperated robots. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) 37 (6), 1231–1245. https://doi.org/10.1109/TSMCC.2007.905819.

Davis, J., Smyth, C., McDowell, K., 2010. The effects of time lag on driving performance and a possible mitigation. IEEE Trans. Rob. 26 (3), 590–593. https://doi.org/10.1109/TRO.2010.2046695.

Dorais, G., Bonasso, R. P., Kortenkamp, D., Pell, B., & Schreckenghost, D. (1999). Adjustable autonomy for human-centered autonomous systems. 16–35.

Draper, J.V., Kaber, D.B., Usher, J.M., 1998. Telepresence. Hum. Factors 40 (3), 354–375. https://doi.org/10.1518/001872098779591386.

Dye, M.W.G., Green, C.S., Bavelier, D., 2009. Increasing speed of processing with action video games. Curr. Dir. Psychol. Sci. 18 (6), 321–326. https://doi.org/10.1111/j.1467-8721.2009.01660.x.

ENDSLEY, M.R., 1999. Level of automation effects on performance, situation awareness and workload in a dynamic control task. Ergonomics 42 (3), 462–492. https://doi.org/10.1080/001401399185595.

Fabrizio, M.D., Lee, B.R., Chan, D.Y., Stoianovici, D., Jarrett, T.W., Yang, C., Kavoussi, L.R., 2000. Effect of time delay on surgical performance during telesurgical manipulation. J. Endourol. 14 (2), 133–138. https://doi.org/10.1089/end.2000.14.133.

Fitts, P.M., 1954. The information capacity of the human motor system in controlling the amplitude of movement. J. Exp. Psychol. 47 (6), 381. https://doi.org/10.1037/h0055392.

Goodrich, M.A., Olsen, D.R., Crandall, J.W., Palmer, T.J., 2001. Experiments in adjustable autonomy. In: Proceedings of IJCAI Workshop on Autonomy, Delegation and Control: Interacting with Intelligent Agents, pp. 1624–1629.

Green, C.S., Bavelier, D., 2003. Action video game modifies visual selective attention. Nature 423 (6939), 534–537. https://doi.org/10.1038/nature01647.

Hart, S.G., 2006. Nasa-task load index (NASA-TLX); 20 years later. Proc. Hum. Factors Ergonomics Soc. Ann. Meet. 50 (9), 904–908. https://doi.org/10.1177/154193120605000909.

Hart, S.G., Staveland, L.E., 1988a. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183.

Hart, S.G., Staveland, L.E., 1988b. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183.

Hendy, K.C., Hamilton, K.M., Landry, L.N., 1993. Measuring subjective workload: when is one scale better than many? Hum. Factors 35 (4), 579–601.

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklade, A.L., Christ, R.E., 1992. Comparison of four subjective workload rating scales. Hum. Factors 34 (4), 429–439. https://doi.org/10.1177/001872089203400405.

Hu, H., Perez, C., Sun, H., Jagersand, M., 2016. Performance of predictive display tele-operation under different delays with different degree of freedoms. In: 2016 International Conference on Information System and Artificial Intelligence (ISAI), pp. 380–384. https://doi.org/10.1109/ISAI.2016.0087.

IBM SPSS Statistics 25. (2017). IBM Corp.

Kirk, R., 2013. Experimental Design: Procedures for the Behavioral Sciences. SAGE Publications, Inc. https://doi.org/10.4135/9781483384733.

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., Willing, C., 2016. Jupyter Notebooks – a publishing format for reproducible computational workflows. In: Loizides, F., Schmidt, B. (Eds.), Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press, pp. 87–90. https://doi.org/10.3233/978-1-61499-649-1-87.

Krishnaiah, P.R., 1980. Analysis of *Variance* 1 North-Holland.

Lane, J.C., Carignan, C.R., Sullivan, B.R., Akin, D.L., Hunt, T., Cohen, R., 2002. Effects of time delay on telerobotic control of neutral buoyancy vehicles. In: Proceedings 2002 IEEE International Conference on Robotics and Automation. 3. pp. 2874–2879. (Cat. No. 02CH37292), 3. https://doi.org/10.1109/ROBOT.2002.1013668.

Lim, J., Wu, W., Wang, J., Detre, J.A., Dinges, D.F., Rao, H., 2010. Imaging brain fatigue from sustained mental workload: an ASL perfusion study of the time-on-task effect. Neuroimage 49 (4), 3426–3435. https://doi.org/10.1016/j.neuroimage.2009.11.020.

Lovi, D., Birkbeck, N., Herdocia, A.H., Rachmielowski, A., Jägersand, M., Cobzaş, D., 2010. Predictive display for mobile manipulators in unknown environments using online vision-based monocular modeling and localization. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5792–5798. https://doi.org/10.1109/IROS.2010.5649522.

Lu, S., Zhang, M.Y., Ersal, T., Yang, X.J., 2018. Effects of a delay compensation aid on teleoperation of unmanned ground vehicles. In: Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, pp. 179–180. https://doi.org/10.1145/3173386.3177064.

Luck, J.P., McDermott, P.L., Allender, L., Russell, D.C., 2006. An investigation of real world control of robotic assets under communication latency. In: Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction, pp. 202–209. https://doi.org/10.1145/1121241.1121277.

Lum, M.J.H., Rosen, J., King, H., Friedman, D.C.W., Lendvay, T.S., Wright, A.S., Sinanan, M.N., Hannaford, B., 2009. Teleoperation in surgical robotics – network latency effects on surgical performance. In: 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6860–6863. https://doi.org/10.1109/IEMBS.2009.5333120.

Ma, R., Kaber, D.B., 2006. Presence, workload and performance effects of synthetic environment design factors. Int. J. Hum. Comput. Stud. 64 (6), 541–552. https://doi.org/10.1016/j.ijhcs.2005.12.003.

MacKenzie, I.S., Ware, C., 1993. Lag As a determinant of human performance in interactive systems. In: Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems, pp. 488–493. https://doi.org/10.1145/169059.169431.

Mathan, S., Hyndman, A., Fischer, K., Blatz, J., Brams, D., 1996. Efficacy of a predictive display, steering device, and vehicle body representation in the operation of a lunar vehicle. In: Conference Companion on Human Factors in Computing Systems, pp. 71–72. https://doi.org/10.1145/257089.257147.

Matheson, A., Donmez, B., Rehmatullah, F., Jasiobedzki, P., Ng, H.-K., Panwar, V., Li, M., 2013. The effects of predictive displays on performance in driving tasks with multi-second latency: aiding tele-operation of lunar rovers. Proc. Hum. Factors Ergonomics Soc. Ann. Meet. 57 (1), 21–25. https://doi.org/10.1177/1541931213571007.

Maxwell, S.E., Delaney, H.D., 2003. Designing Experiments and Analyzing Data: a Model Comparison Perspective. Routledge.

Miller, D.P., Machulis, K., 2005. Visual aids for lunar rover tele-operation. In: Battrick, R. (Ed.), Proceedings of 8th International Symposium on Artificial Intelligence, Robotics and Automation in Space. ESA Publishing, Noordwijk, Netherlands.

Moffat, S.D., Hampson, E., Hatzipantelis, M., 1998. Navigation in a "virtual" maze: sex differences and correlation with psychometric measures of spatial ability in humans. Evol. Hum. Behav. 19 (2), 73–87. https://doi.org/10.1016/S1090-5138(97)00104-9.

Neumeier, S., Wintersberger, P., Frison, A.-K., Becher, A., Facchi, C., Riener, A., 2019. Teleoperation: the holy grail to solve problems of automated driving? sure, but latency matters. In: Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 186–197. https://doi.org/10.1145/3342197.3344534.

Nielsen, C.W., Goodrich, M.A., Ricks, R.W., 2007. Ecological interfaces for improving mobile robot teleoperation. IEEE Trans. Rob. 23 (5), 927–941. https://doi.org/10.1109/TRO.2007.907479.

Oboe, R., Fiorini, P., 1998. A design and control environment for internet-based telerobotics. Int. J. Robot. Res. 17 (4), 433–449. https://doi.org/10.1177/027836499801700408.

Parasuraman, R., Sheridan, T.B., Wickens, C.D., 2008. Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. J. Cogn. Eng. Decis. Mak. 2 (2), 140–160. https://doi.org/10.1518/155534308X284417.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Perez, M., Xu, S., Chauhan, S., Tanaka, A., Simpson, K., Abdul-Muhsin, H., Smith, R., 2016. Impact of delay on telesurgical performance: study on the robotic simulator dV-Trainer. Int. J. Comput. Assist. Radiol. Surg. 11 (4), 581–587. https://doi.org/10.1007/s11548-015-1306-y.

Rachmielowski, A., Birkbeck, N., Jägersand, M., 2010. Performance evaluation of monocular predictive display. In: 2010 IEEE International Conference on Robotics and Automation, pp. 5309–5314. https://doi.org/10.1109/ROBOT.2010.5509652.

Richardson, A.E., Powers, M.E., Bousquet, L.G., 2011. Video game experience predicts virtual, but not real navigation performance. Comput. Hum. Behav. 27 (1), 552–560. https://doi.org/10.1016/j.chb.2010.10.003.

Ricks, B., Nielsen, C.W., Goodrich, M.A., 2004. Ecological displays for robot interaction: a new perspective. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems. 3. pp. 2855–2860. IROS) (IEEE Cat. No.04CH37566), 3. https://doi.org/10.1109/IROS.2004.1389842.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., Bühner, M., 2010. Is it really robust. Methodology 6 (4), 147–151. https://doi.org/10.1027/1614-2241/a000016.

Schutte, P.C., 2017. How to make the most of your human: design considerations for human–machine interactions. Cogn. Technol. Work 19 (2), 233–249. https://doi.org/10.1007/s10111-017-0418-2.

Sheridan, T.B., 1995. Teleoperation, telerobotics and telepresence: a progress report. Control Eng. Pract. 3 (2), 205–214. https://doi.org/10.1016/0967-0661(94)00078-U.

Tukey, J.W., 1977. Exploratory Data Analysis 2 Reading, Mass.

Vidulich, M.A., Tsang, P.S., 1987. Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. Proc. Hum. Factors Soc. Ann. Meet. 31 (9), 1057–1061. https://doi.org/10.1177/154193128703100930.

Wickens, T.D., Keppel, G., 2004. Design and Analysis: a Researcher's Handbook. Pearson Prentice-Hall.

Xu, S., Perez, M., Yang, K., Perrenot, C., Felblinger, J., Hubert, J., 2014. Determination of the latency effects on surgical performance and the acceptable latency levels in telesurgery using the dV-Trainer® simulator. Surg. Endosc. 28 (9), 2569–2576. https://doi.org/10.1007/s00464-014-3504-z.

Zhang, Y., Li, H., 2016. Handling qualities evaluation of predictive display model for rendezvous and docking in lunar orbit with large time delay. In: 2016 IEEE Chinese Guidance, Navigation and Control Conference (CGNCC), pp. 742–747. https://doi.org/10.1109/CGNCC.2016.7828878.

Zheng, Y., Brudnak, M.J., Jayakumar, P., Stein, J.L., Ersal, T., 2016. An Experimental Evaluation of a Model-Free Predictor Framework in Teleoperated Vehicles**This work was supported by the Automotive Research Center (ARC) in accordance with Cooperative Agreement W56HZV-14-2-0001 U.S. Army Tank Automotive Research, Development and Engineering Center (TARDEC) Warren, MI. UNCLASSIFIED: Distribution Statement A. Approved for public release. #27479. IFAC-Papers OnLine 49 (10), 157–164. https://doi.org/10.1016/j.ifacol.2016.07.513.