INTERNATIONAL DESIGN CONFERENCE - DESIGN 2018
https://doi.org/10.21278/idc.2018.0459

STEERING A SHIP - INVESTIGATING AFFECTIVE STATE AND WORKLOAD IN SHIP SIMULATIONS

H. Dybvik, A. Wulvik and M. Steinert

Abstract

We present an experiment investigating concepts of affective state and workload in a large ship manoeuvring context. It is run on a consumer ship simulator software where student participants (N=31) perform two ecologically valid scenarios: sailing on open sea and in a harbour. Results from surveys show highly significant changes in terms of both affect and workload between the scenarios. Thus, one should consider varying affects and workloads from users in varying contexts, consequently demanding new design paradigms for product development, such as dynamically adaptive interfaces.

Keywords: human behaviour, emotional engineering, engineering design, empirical studies, ocean space

1. Introduction: The human element and the ship bridge

The ship bridge is where the captain and his crew controls the ship. Navigation, monitoring systems, and communicating with both internal and external personnel are important activities. Sea piloting, i.e. sailing on open sea normally consists of monitoring tasks and no active navigation at all. Harbour piloting, i.e. sailing in harbours, requires continuous adjustment of speed and course, monitoring ship systems, and communicating with both crew and external contacts. These scenarios range from the monotone to the highly complex (Norros, 2004; Nilsson et al., 2009).

Maritime accidents occur in either scenario (Nilsson et al., 2009), mostly as the result of human error. Research shows that 49 to 96 percent of all shipping incidents or marine causalities are caused by human error (Rothblum, 2000; Hetherington et al., 2006; Tzannatos, 2010). Given the large share of maritime accidents caused by human error, this paper aims to direct attention towards the human users and their mental state during ship operation with the goal of identifying opportunities for reducing accidents. The notion of human centred design (Woodson and Conover, 1970; Sanders and McCormick, 1987) has existed since the 1960s. When considering humans in engineering, they are usually represented by generic models based on certain boundary conditions (Balters and Steinert, 2017). Models often represent the "average" human, with a general and stable behaviour response. Kahneman and Tversky (1979, 1984) show that this is indeed not the case. They show that humans are not rational with stable behavioural responses to stimuli. Human behaviour is influenced by psychological, physiological and situational factors. This could be issues in personal life, lack of sleep, or suddenly demanding tasks that needs to be solved. Following the fact that humans are not static entities with known responses, but rather change over time and contexts, efforts should be made to gain insights about what might influence behaviour. Two potentially influential topics are the constructs of affect and workload. Knowledge about how affective state and workload influence operator performance could potentially aid engineers in their work to design and test new product solutions for the maritime industry. We believe that by taking these parameters into consideration, human error could be reduced by designing the system around the human, and not make the human adapt to the system.

The paper proposes and demonstrates an experimental setup to investigate differences in affective state and workload between two ecologically valid scenarios within the domain of large ship navigation. The goal of this paper is to show that there are measurable differences in affective state and workload between the two scenarios. This may influence new ship bridge designs. These tasks have been developed in cooperation with ship simulator instructors with extensive experience as ship navigators. The experiment is run in a consumer ship simulator software (N=31) where participants from a student population are tasked to steer a ship in the following scenarios: cruising on open sea and navigating a busy harbour. Data was collected through video, self-assessment surveys and physiology sensors. The paper is part of a larger study investigating the relationship between physiological data, affect, and workload. The foundation, description and analysis of the physiological data is not within the scope or aim of this paper, and will be discussed elsewhere.

The results from the self-assessments show highly significant differences in terms of both affect and workload for the two scenarios. Consequently, one will have to consider distinctly varying affects and workloads from users in varying contexts, which, if translated into GUI and UI design suggest new design paradigms such as dynamically adaptive interfaces.

2. Theoretical foundation

2.1. Affect

Psychology presents emotion or affect as a set of variables that may moderate behaviour (Balters and Steinert, 2017). There are two main schools of thought when describing affect. The first describe emotions as a set of discrete categories (Tomkins, 1962; Ekman and Friesen, 1971; Ekman, 1992). The second describe emotions as a combination of multiple dimensions (Thayer, 1967; Russell, 1980; Watson and Tellegen, 1985; Russell and Barrett, 1999). In this paper, we adopt the description of emotions or affect of Russell (1980), the Circumplex Model of Affect, later named the Affect Grid (Russell et al., 1989). Affect is described as a construct made up of the combination of the two dimensions, arousal-sleepiness and pleasure-displeasure, see Figure 1.

Several researchers have considered how stress might influence human performance (Westman and Eden, 1996; Healey and Picard, 2005; Balters and Steinert, 2017). Russell et al. (1989) describe the construct of stress as the combination of arousal and displeasure. This is also referred to as distress as opposed to eustress which is the combination of arousal and pleasure (Healey and Picard, 2005; Balters and Steinert, 2014, 2017). Baddeley (1972) shows that increased arousal seems to narrow attention, which in term increases performance on the task that is deemed most important, but decrease performance on all other tasks.

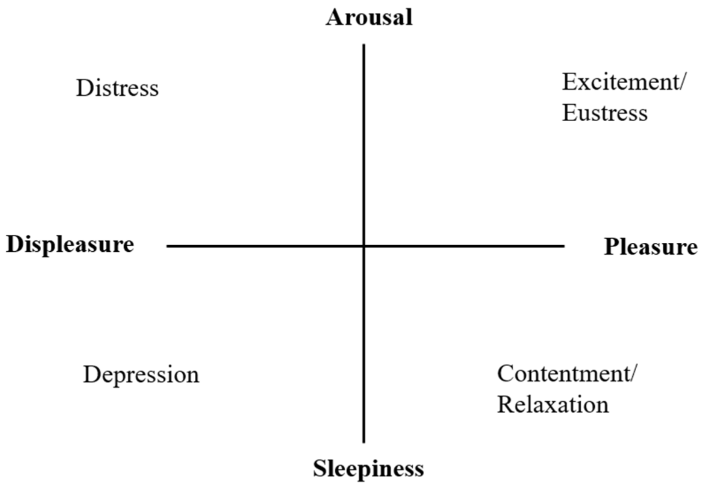

Figure 1. The Affect grid (adapted from Russell, 1980; and Russell et al., 1989)

2.1.1. Subjective measurements of affect

Assessing the subjective experience of affect is commonly done through self-report surveys. Affect can be evaluated through survey questions asking participants to evaluate levels of pleasantness and arousal (Russell, 1980), or through the single-item Affect grid (Russell et al., 1989). Positive and negative affect can be evaluated through the Positive and Negative Affect Schedule (PANAS) scales (Watson et al., 1988; Thompson, 2007). The Activation-Deactivation Adjective Check List (AD ACL) measures levels of activation (Thayer, 1967, 1986). Surveys provide a simple and low cost manner of gathering data of affective states. When using surveys in an experiment, they either interrupt participants, or must be used after tasks are finished. This might influence results, either because of the effect of an interruption, or that participants must recall how they felt during a task. Due to the subjective nature of surveys, there might be issues of self-filtering and different interpretations of questions.

2.1.2. Behavioural measurements of affect

Behavioural measurements of affect are typically concerned with measuring components of facial expression (Ekman and Friesen, 1978; Gottman and Krokoff, 1989), pitch of voice (Russell et al., 2003), and body posture and gesturing (Coulson, 2004; Wulvik et al., 2016). Advantages with behavioural measurements of affects is a very fine grained analysis of behaviour by trained experts, partially avoiding self-filtering of results, such as might be the case when answering surveys. Drawbacks are that these analyses are very labour- and time intensive, and that there might be issues of inter-coder reliability.

2.1.3. Physiological measurements of affect

The autonomic nervous system (ANS) is in charge of modulating peripheral functions of the body (Öhman et al., 2000; Mauss and Robinson, 2009). The ANS consists of the sympathetic and parasympathetic system. The sympathetic system is dominant during periods of activation, or “fight or flight”, while the parasympathetic system is dominant during resting periods of the body. Changes in affective state are linked to physiological responses through the ANS. These responses can be through heart rate, heart rate variability, breathing rate, pupil dilation, muscle tension, galvanic skin response, body temperature, blood pressure, and brain activity to mention some. Healey and Picard (2005) showed a relation between levels of stress and metrics derived from galvanic skin response and heart rate variability. Baltaci and Gokcay (2016) differentiates affective states from relaxation to stress through pupil dilation and facial temperature. For a more comprehensive overview we refer to Balters and Steinert (2017), Mauss and Robinson (2009) and Levenson (2014). Physiology sensors have the advantage of providing continuous data without interrupting the person being measured, as opposed to subjective measurements through surveys. One limitation is that human physiology is very complex, and it is difficult to control all influencing factors. Another challenge with physiology data is interpreting results. How does e.g. a change in measured voltage between two sensors placed on the chest translate into affect? We recommend reading Balters and Steinert (2017) for a more complete overview.

2.2. Workload

Workload or cognitive load refers to the mental effort imposed on working memory by a particular task. (Sweller, 1988; Paas and Van Merriënboer, 1994; Paas et al., 2003) Cognition is related to our perception, in that perceptual activity, such as thinking, deciding, calculation, remembering, looking, searching increases the perceptual load, thereby the workload (Hart and Staveland, 1988). As the working memory is limited, it can be overloaded by increasing the requirements for perceptual activity. Wierwille and Eggemeier (1993) provide an overview of methods to measure workload. These can be divided into Subjective, performance-based and physiological.

2.2.1. Subjective measurements of workload

The NASA Task Load Index (NASA-TLX), a multi-dimensional scale designed for obtaining workload estimates (Hart and Staveland, 1988; Hart, 2006). NASA-TLX consists of rating six sub-scales, mental demand, physical demand, temporal demand, performance, effort, and frustration, from low to high. Participants filling out the survey are also asked to pairwise compare the six dimensions in terms of how important they are for the performed task. An estimate of total workload is then calculated from the weighted average. Another multi-dimensional scale of subjective workload is the Subjective Workload Assessment Technique (SWAT) (Reid and Nygren, 1988). It uses three levels (low, medium, high) along three dimensions, time load, mental effort load, and psychological stress load, to assess workload.

Overall Workload (Vidulich and Tsang, 1987) is a single scale measurement of subjective workload, ranging from very low to very high. Vidulich and Tsang (1987) show that the single-dimension scale of Overall Workload has higher sensitivity than the multi-dimensional scale of NASA TLX. Hill et al. (1992) showed that both the single-dimension scale of Overall Workload and NASA TLX was superior to SWAT in terms of sensitivity.

2.2.2. Performance based measurements of workload

Performance is expected to decrease with increases in workload through reduction in speed and accuracy (Wierwille and Eggemeier, 1993). Two strategies of evaluating workload through performance are common, primary and secondary task performance. Primary task performance, e.g. steering a ship might be insensitive to variations in workload, due to the operator recruiting extra resources to maintain performance (Hart and Wickens, 1990). Secondary task performance can both be assessed through external tasks and embedded tasks. External tasks are not part of the system being tested, e.g. calculating arbitrary arithmetic, while embedded tasks have a logical connection to the primary task, e.g. communicating via radio on a ship.

2.2.3. Physiological measurements of workload

Changes in physiological states has been shown to correspond with changes in workload (Galy et al., 2012). Common physiological measurements to evaluate workload are heart rate variability (HRV) (McDuff et al., 2014), electroencephalography (EEG) (Wilson and Russell, 2003), pupillary response (Iqbal and Bailey, 2005), and galvanic skin response (GSR) (Nourbakhsh et al., 2012).

3. Ship navigation experiment

An experiment was created to investigate concepts of affect and workload in two different ecologically valid scenarios in the context of large ship navigation. One task concerned steering a large ship on open water. The other task concerned steering a large ship through a busy harbour. These tasks can be described as low and high activity respectively. The aim of the experiment was to identify potential differences in affective state and workload in the different scenarios. The implication of different affective states and levels of workload for the various scenarios is that users could have changing capabilities, and that this should be addressed through the design of systems in the future. For the experiment, we formulate the following research question:

Is there a measurable difference in affective state and workload between low and high activity scenarios in the context of large ship navigation?

3.1. Scenarios

Two ecologically valid scenarios were created in the commercial ship simulator software Ship Simulator Extremes (*Ship Simulator Extremes*, 2010), replicating two typical situations in large ship navigation. Ecologically validity is obtained by the nature of the primary and the secondary task and the nature of the environmental stimuli i.e. sounds. Scenarios describe common activities on board large ships in daily operation. The scenarios and stimuli were developed in cooperation with several ship navigators with long experience as professional navigators.

3.1.1. Ship navigation on open sea – low level of activity

The first scenario was designed to recreate a low-activity situation where the task was to navigate on open sea. This is typically an uneventful task with long periods of time spent monitoring systems. The environment was set to *Dover*, and ship set to *Pride of Rotterdam,* a 215-meter long car ferry. The ship was placed close to the exit of Dover harbour with the front of the ship pointing towards the English Channel. Participants were instructed to steer the ship straight ahead towards Calais, France. The task lasted for 15 minutes, but the duration was unknown to participants. The monotonous sound of a ship engine was added to create a realistic backdrop.

3.1.2. Ship navigation in a busy harbour – high level of activity

The second scenario was designed to simulate a high-activity situation where the task was to navigate a busy harbour under a time constraint with additional secondary in the form of radio communication. The environment was set to *Rotterdam*, and *Pride of Rotterdam* was again used at ship. The participants were instructed to steer through narrow channels to a designated berth for docking. Upon leaving the starting position, a ten-minute timer would start and be displayed in the top left corner of the screen, instructing participants to reach their destination within this time limit. At regular intervals throughout the ten minutes, participants were prompted to answer eight pre-recorded questions via radio from immigration, customs and the ship's main office. These questions were voiced by three different people unfamiliar to the participants. Answers to the questions could be found in two lists provided to the participants, a cargo manifest and a crew list. These were consciously designed to be hard to read, with small letters and lots of superfluous information. Questions were repeated after 90 seconds if no answer had been given, or upon request of the participants. If participants reached the designated berth, a new destination was given. The task was designed in such a way that the final destination would be next to impossible to reach in the available ten minutes.

3.2. Physical environment

The aim of the physical environment was a controlled, static, physical space for conducting the abovementioned ecologically valid scenarios in the context of large ship navigation. A honeycomb cardboard cubicle was built (similar to the one made by Leikanger et al. (2016), equipped with a 27" computer screen mimicking the window view. A keyboard had the numerical pad marked with stickers indicating what ship functionality they controlled, e.g. rudder, thruster, etc. Today, a ship bridge control interface consists of button arrays resembling a keyboard. Additionally, much of monitoring tasks are conducted using information conveyed on a computer screen. Headphones eliminated external noise, ensuring exposure to the sound introduced by the experimenters only, i.e. ship engine noise, radio chatter and the task-specific questions. Effects from changes in external light was controlled by obscuring ambient light and illuminating the cubicle artificially with an LED strip and normal ceiling lights. Additional equipment included a mouse for answering the surveys, two web cameras for recording and monitoring the participant, a Bluetooth antenna hidden close to the devices, lists with information regarding the questions in the second scenario and marking tape indicating the area for placing the left hand. Figure 2 shows the experiment environment.

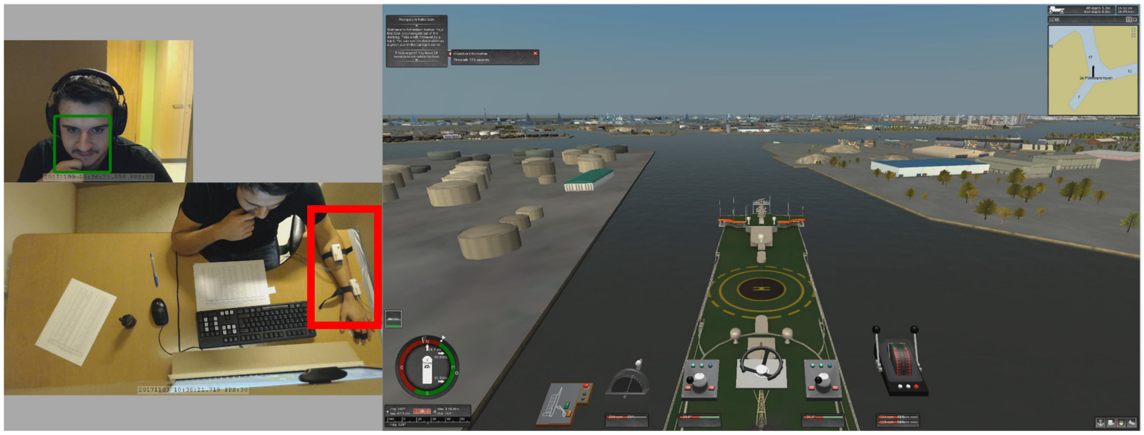

Figure 2. Experiment environment, both physical and virtual. ECG (top) and GSR (bottom) sensors highlighted in red rectangle

3.3. Collecting data from participants

A combination of self-report surveys and physiology sensors were used for data collection in the experiment. In addition, video was recorded to allow in-depth analysis of collected data.

3.3.1. Self-report surveys

To evaluate subjectively experienced affect, participants were asked to evaluate their state of arousal, awakeness, alertness, pleasantness, and stress on scales from 0 to 10. Arousal and pleasantness was taken directly from the Circumplex Model of Affect. Awakeness and alertness were added after pilot studies uncovered that participants had trouble understanding the meaning of arousal to triangulate their meaning. A question of stress was included in the surveys to capture the participants' notion of stress directly, and not only as a combination of arousal and pleasantness.

For self-assessment of workload, the single-dimension Overall Workload scale (Vidulich and Tsang, 1987) and NASA TLX (Hart and Staveland, 1988) was used. Overall Workload was evaluated on a scale from 0 to 10, and the six dimensions of the NASA TLX survey was evaluated on scales from 1 to 7, as well as 15 pairwise comparisons. All survey answers were collected through Google Forms.

3.3.2. Physiology sensors

Two types of physiology data were collected in this experiment, electrocardiography (ECG) and galvanic skin response (GSR). Electrocardiography measures electric potentials over the heart through sensors placed on the skin. The Shimmer3 ECG unit (Shimmersense, 2017a) was used in this experiment, with a sampling rate of 512 Hz. Five sensors were placed on the skin of participants per the instructions provided by Shimmer, with the V_x lead placed on position six. Data collected through ECG is measured in millivolts [mV], and can be translated into variables such as heart rate and heart rate variability. Galvanic skin response (GSR) is a measurement of conductance over the skin. The Shimmer3 GSR+ Unit (Shimmersense, 2017b) was used to measure skin conductivity. Two sensors were connected to the underside of the medial phalanx on the index and middle finger of the left hand. Sampling rate was set to 128 Hz.

3.3.3. Organising stimuli and synchronizing data

iMotions 6.4 (*iMotions*, 2017), a software platform for biometric research was as framework for presenting stimuli and synchronizing data. The sequence of instructions, surveys and simulator tasks were pre-defined in iMotions. Physiology data and video were given a common timestamp from iMotions, syncing data for future analysis.

3.4. Stating the hypotheses

The analysis of results in this paper concerns the change of self-reported affective state and workload. We operationalise the research question stated above into testable hypotheses.

Is there a measurable difference in affective state and workload between high and low activity scenarios in the context of large ship navigation?

3.4.1. Affect hypotheses

Affect is measured by asking participants to evaluate their level of arousal, awakeness, alertness, pleasantness, and stress. This leads to the following five hypotheses:

- **Affect H1:** *There is a significant change in self-reported arousal between low and high activity tasks.*
- **Affect H2:** *There is a significant change in self-reported awakeness between low and high activity tasks.*
- **Affect H3:** *There is a significant change in self-reported alertness between low and high activity tasks.*
- **Affect H4:** *There is a significant change in self-reported pleasantness between low and high activity tasks.*

- **Affect H5:** *There is a significant change in self-reported stress between low and high activity tasks.*

3.4.2. Workload hypotheses

Workload has been evaluated by participants assessing their overall workload on a single scale and through filling out the NASA TLX survey. This leads to the two following hypotheses:

- **Workload H1:** *There is a significant change in overall workload between low and high activity tasks.*
- **Workload H2:** *There is a significant change in TLX workload between low and high activity tasks.*

3.5. Running the experiment

This section aims to display how the experiment was run, and give a detailed description of the data foundation.

3.5.1. Participants

Participants in this experiment came from an engineering background (N=31). Age ranged from 19 to 33 years (24.0 ± 2.74). Out of 31 participants, 18 were male and 13 female. In addition, there were eleven participants were excluded from the analysis due to technical errors and failure to follow instructions. In the invitation to the experiment, participants were asked to participate in a study concerning "Ship Manoeuvring Behaviour". They were asked to wear a loose top for convenient connection of physiology sensors.

3.5.2. Experimenters

Two researchers conducted the experiment. The first experimenter would greet, brief, and attach sensors to participants. All interactions were scripted in advance to ensure that every participant was exposed to the same stimuli. The experimenter read all instructions from a manuscript, wore similar clothing (black jeans, light coloured dress shirt, hair pulled back in pony-tail, and no make-up). The second experimenter would sit behind a wall controlling the stimuli. After the experiment finished, the first experimenter debriefed the participant and removed sensors.

3.5.3. Protocol

Participants were greeted, introduced to the experiment, and informed about what kind of data that would be recorded. A consent form was signed by the participant, agreeing to have video, physiology data (electrocardiography and galvanic skin response) and survey answers recorded. Physiology sensors were attached by the experimenter. Participants were then instructed to sit down in front of a computer screen, and place their left hand on the table, making sure their arm was resting comfortably. They were told that instructions may be given both on-screen and through audio. In the case of audio instructions, answers should be given through a radio handset. Usage of the radio handset was explained and demonstrated. Participants were instructed to keep their left hand still throughout the experiment to ensure the quality of GSR data recorded. After instructions were given, the experimenter left the room and joined the second experimenter behind a wall. The computer screen showed a black image with white crosshairs in the middle when participants entered the room. When participants were ready to start, the second experimenter manually started the sequence of stimuli in iMotions. Participants were first presented with neutral stimuli. Participants then filled out a survey on their affective state to serve as a reference baseline. Information about the experiment was given in writing with a white background. They were informed that they would be controlling the ship *Pride of Rotterdam* and execute various tasks. Participants were informed that there were two printed lists, a crew list and cargo manifest, to their right side. These lists should not be used before instructed to do so. Following the initial brief, participants were shown a video giving instructions for how to control the ship with the keyboard. All keys to be used on the keyboard were physically labelled with a short explanatory name. After receiving instructions, participants were informed that the first task would begin and the computer screen switched to the simulator software for the low activity task. The second experimenter manually unpaused the software and gave over control to the participants. After 15 minutes from leaving the harbour in Dover the software would display a loading screen, initialising the high activity task. The second experimenter would manually change the view to the second survey, concerning affective state and workload. After completing the survey, the view was manually switched back to the simulator software, starting the high activity task. The ten-minute timer would start after the ship had started moving, and pre-recorded radio questions were manually played at pre-defined intervals by the second experimenter. After the ten minutes passed, a screen telling participants that they failed their mission (no participants were able to complete the mission, as expected). The third survey was presented to participants, asking about affective state and workload. When completed, participants were prompted to answer background questions, e.g. age, gender, occupation, in a fourth survey. After completing the final survey, they were informed that the experiment was finished, and were thanked for their participation. The first experimenter would walk back to debriefing the participants, thank them for their contribution, remove the sensors and ask them not to share content or details about the experiment to others. All equipment was cleaned and printed lists were replaced after each participant.

4. Survey results

Survey results from the 31 participants completing the experiment was analysed in SPSS Statistics (IBM, 2016) to investigate potential statistical differences in affective states and workload. A total of seven variables were tested for statistically significant change in values on an 11-point scale, between low and high activity tasks. Statistical tests were selected based on the properties of recorded data, i.e. outliers, normal-, and symmetric distributions. Difference in values between the two scenarios is the foundation for the tests. Paired samples t-test was used for normally distributed date without significant outliers. For data violating the assumptions of normal distribution or no significant outliers, the Wilcoxon signed-rank test was used if the data was symmetrically distributed. For non-symmetric distributions, the Sign test was used. The Wilcoxon signed-rank test and the Sign test evaluates median differences as opposed to mean differences in the paired samples t-test. Outliers are defined by SPSS statistics as values more than 1.5 box-lengths from the edge of a box in a box plot. Shapiro-Wilk's test for normal distribution was used to assess whether values were normally distributed, where significance values larger than 0.05 indicates normally distributer variables. Symmetricity of distribution was evaluated visually using histograms. Data are mean ± standard deviation, unless otherwise stated. Table 1 contains descriptive statistics, and Table 2 contains metrics associated with assumptions that decide which statistical tests to use along with the corresponding results. As shown in Table 2, all seven variables are significantly different in the two scenarios.

Table 1. Descriptive statistics

Variable	S1	S1 Median	S2	S2 Median	Diff.	Diff. Median
Arousal	5.61 ± 2.38	6	6.68 ± 2.70	8	1.06 ± 1.75	1
Awakeness	6.48 ± 2.05	7	7.61 ± 2.04	8	1.13 ± 1.09	1
Alertness	6.26 ± 1.95	7	7.42 ± 1.86	8	1.16 ± 1.61	1
Pleasantness	6.35 ± 1.62	7	4.68 ± 1.80	4	-1.68 ± 1.54	-2
Stress	3.23 ± 1.94	3	6.39 ± 2.14	7	3.16 ± 1.88	3
Overall Workload	2.03 ± 1.70	2	8.03 ± 1.78	8	6.00 ± 2.93	7
TLX Workload	2.86 ± 1.23	2.73	5.60 ± 0.74	5.80	2.73 ± 1.37	2.86

Table 2. Testing for statistical difference change in variables between low and high activity scenarios

Variable	Outliers	Shapiro-Wilk's test	Symmetric	95% CI Lower	95% CI Upper	Sig. (2-tailed)
Arousal	Yes	0.063	Yes	0.42	1.71	< 0.01 [a]
Awakeness	No	< 0.01	No	0.73	1.53	< 0.01[c]
Alertness	Yes	0.014	No	0.57	1.75	< 0.01[c]
Pleasantness	No	0.214	Yes	-2.24	-1.11	< 0.01 [a]
Stress	No	0.112	Yes	2.47	3.85	< 0.01 [a]
Overall Workload	Yes	< 0.01	No	4.93	7.07	< 0.01[c]
TLX Workload	No	0.48	Yes	2.23	3.24	< 0.01[a]

a: Paired samples t-test, b: Wilcoxon signed-rank test, c: Sign test

5. Discussion: Interpreting the results and the way forward

Results in the above tests show that there are significant changes in all seven variables. TLX Workload is a weighted sum of the six dimensions: Mental demand, physical demand, temporal demand, performance, effort, and frustration level. With the exception of performance (p=0.69), all dimensions had significantly changes. This might be due to difficulties related to comparing performance in two very different and unfamiliar scenarios. Changes are quite small for the variables of arousal, awakeness and alertness, with mean changes of around one on an eleven-point scale. Variables of pleasantness, stress, overall workload and TLX workload have a larger change (see Table 1). We are not sure whether the differences in magnitude of change is due to real differences, or due to how participants interpret the survey questions. One can speculate e.g. that participants did not have a clear understanding of the concepts of arousal, awakeness, and alertness, or at least had difficulties evaluating them. Pleasantness, stress and workload might be more intuitively understandable for the participants, which might be the reason for the difference in magnitude of change. This finding is interesting, as it contrasts with the fact that Russell (1980) defines stress as a combination of arousal and pleasantness. We know from literature that there is supposed to be a link between physiological data and arousal, e.g. heart rate variability and skin conductance. Further work will include analysing physiological data and comparing results with subjective assessment of affective state and workload, investigating the relationship between the two. One limitation of our study is that participants were sampled from a student population. Results might have been influenced by this fact, due to being unfamiliar with the situation of ship piloting. We believe that the findings that show a difference in affective state and workload between the two scenarios are valid for the context of ship navigation, although the effect size should be verified through testing with professional navigators in more realistic contexts, i.e. professional ship simulators or real ships.

The results nevertheless show that there is a clear difference in affective state and workload in the two scenarios tested in this experiment. Consequently, one should consider distinctly varying affects and workloads from users in varying contexts. This, if translated into product development, GUI, and UI design suggest new design paradigms such as dynamically adaptive interfaces.

References

Baddeley, A.D. (1972), “Selective attention and performance in dangerous environments”, *British Journal of Psychology*, Vol. 63 No. 4, pp. 537–546. https://doi.org/10.1111/j.2044-8295.1972.tb01304.x

Baltaci, S. and Gokcay, D. (2016), “Stress Detection in Human–Computer Interaction: Fusion of Pupil Dilation and Facial Temperature Features”, *International Journal of Human–Computer Interaction*, Vol. 32 No. 12, pp. 956–966. https://doi.org/10.1080/10447318.2016.1220069

Balters, S. and Steinert, M. (2014), “Decision-making in engineering-a call for affective engineering dimensions in applied engineering design and design sciences”, *Proceedings of the 2014 International Conference On*

Innovative Design and Manufacturing (ICIDM 2014), August 13-15, 2014, Montreal, Canada, IEEE, pp. 11–15. https://doi.org/10.1109/IDAM.2014.6912663

Balters, S. and Steinert, M. (2017), "Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices", *Journal of Intelligent Manufacturing*, Vol. 28 No. 7, pp. 1585 - 1607. https://doi.org/10.1007/s10845-015-1145-2

Coulson, M. (2004), "Attributing Emotion to Static Body Postures: Recognition Accuracy, Confusions, and Viewpoint Dependence", *Journal of Nonverbal Behavior*, Vol. 28 No. 2, pp. 117–139. https://doi.org/10.1023/B:JONB.0000023655.25550.be

Ekman, P. (1992), "An argument for basic emotions", *Cognition and Emotion*, Vol. 6 No. 3-4, pp. 169–200. https://doi.org/10.1080/02699939208411068

Ekman, P. and Friesen, W.V. (1971), "Constants across cultures in the face and emotion", *Journal of Personality and Social Psychology*, Vol. 17 No. 2, pp. 124-129. https://doi.org/10.1037/h0030377

Ekman, P. and Friesen, W.V. (1978), *Facial Action Coding System: A Technique for the Measurement of Facial Movement*, Consulting Psychologists Press, Palo Alto, California.

Galy, E., Cariou, M. and Mélan, C. (2012), "What is the relationship between mental workload factors and cognitive load types?", *International Journal of Psychophysiology*, Vol. 83 No. 3, pp. 269–275. https://doi.org/10.1016/j.ijpsycho.2011.09.023

Gottman, J.M. and Krokoff, L.J. (1989), "Marital interaction and satisfaction: a longitudinal view", *Journal of Consulting and Clinical Psychology*, Vol. 57 No. 1, pp. 47-52.

Hart, S.G. (2006), "Nasa-Task Load Index (NASA-TLX); 20 Years Later", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* Vol. 50 No. 8, pp. 904–908. https://doi.org/10.1177/154193120605000909

Hart, S.G. and Staveland, L.E. (1988), "Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research", *Advances in Psychology*, Vol. 52, pp. 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hart, S.G. and Wickens, C.D. (1990), "Workload Assessment and Prediction", In: Booher, H.R. (Ed.), *Manprint,* Springer, Dordrecht, pp. 257–296. https://doi.org/10.1007/978-94-009-0437-8_9

Healey, J.A. and Picard, R.W. (2005), "Detecting stress during real-world driving tasks using physiological sensors", *IEEE Transactions on Intelligent Transportation Systems*, Vol. 6 No. 2, pp. 156–166. https://doi.org/10.1109/TITS.2005.848368

Hetherington, C., Flin, R. and Mearns, K. (2006), "Safety in shipping: The human element", *Journal of Safety Research*, Vol. 37 No. 4, pp. 401–411. https://doi.org/10.1016/j.jsr.2006.04.007

Hill, S.G., Iavecchia, H.P., Byers, J.C., Bittner, A.C., Zaklade, A.L. and Christ, R.E. (1992), "Comparison of Four Subjective Workload Rating Scales", *Human Factors*, Vol. 34 No. 4, pp. 429–439. https://doi.org/10.1177/001872089203400405

IBM (2016), *IBM SPSS Statistics for Mac.* [online] IBM, New York, USA. Available at: https://www.ibm.com/products/spss-statistics

iMotions (2017), *iMotions biometric research platform*. [online] iMotions. Available at: https://imotions.com

Iqbal, S.T. and Bailey, B.P. (2005), "Investigating the Effectiveness of Mental Workload As a Predictor of Opportune Moments for Interruption", *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05), Portland, USA, April 2-7, 2005*, ACM, New York, USA, pp. 1489–1492. https://doi.org/10.1145/1056808.1056948

Kahneman, D. and Tversky, A. (1979), "Prospect Theory: An Analysis of Decision under Risk", *Econometrica,* Vol. 47 No. 2, pp. 263–292.

Kahneman, D. and Tversky, A. (1984), "Choices, values, and frames", *American Psychologist*, Vol. 39 No. 4, pp. 341-350. https://doi.org/10.1037/0003-066X.39.4.341

Leikanger, K.K., Balters, S. and Steinert, M. (2016), "Introducing the Wayfaring Approach for the Development of Human Experiments in Interaction Design and Engineering Design Science", *Proceedings of the DESIGN 2016 / 14th International Design Conference, Dubrovnik, Croatia, May 16-19, 2016*, The Design Society, Glasgow, pp. 1751–1762.

Levenson, R.W. (2014), "The Autonomic Nervous System and Emotion", *Emotional Review*, Vol. 6 No. 2, pp. 100–112. https://doi.org/10.1177/1754073913512003

Mauss, I.B. and Robinson, M.D. (2009), "Measures of emotion: A review", *Cognition and Emotion,* Vol. 23 No. 2, pp. 209–237. https://doi.org/10.1080/02699930802204677

McDuff, D., Gontarek, S. and Picard, R. (2014), "Remote measurement of cognitive stress via heart rate variability", *Proceedings of the 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2014), Chicago, Illinois, August 26-30, 2014*, IEEE, pp. 2957–2960. https://doi.org/10.1109/EMBC.2014.6944243

Nilsson, R., Gärling, T. and Lützhöft, M. (2009), "An experimental simulation study of advanced decision support system for ship navigation", *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 12 No. 3, pp. 188–197. https://doi.org/10.1016/j.trf.2008.12.005

Norros, L. (2004), Acting under uncertainty: The core-task analysis in ecological study of work, VTT Technical Research Centre of Finland, Espoo, Finland.

Nourbakhsh, N., Wang, Y., Chen, F. and Calvo, R.A. (2012), "Using Galvanic Skin Response for Cognitive Load Measurement in Arithmetic and Reading Tasks", *Proceedings of the 24th Australian Computer-Human Interaction Conference (OzCHI '12), Melbourne, Australia, November 26-30, 2012,* ACM, New York, USA, pp. 420–423. https://doi.org/10.1145/2414536.2414602

Öhman, A., Hamm, A. and Hugdahl, K. (2000), "Cognition and the autonomic nervous system: orienting, anticipation, and conditioning", In: Cacioppo, J.T., Tassinary, L.G. and Berntson, G.G. (Eds.), *Handbook of psychophysiology*, 2nd ed., Cambridge University Press, New York, pp. 533–575.

Paas, F., Tuovinen, J.E., Tabbers, H. and Gerven, P.W.M.V. (2003), "Cognitive Load Measurement as a Means to Advance Cognitive Load Theory", *Educational Psychologist*, Vol. 38 No. 1, pp. 63–71. https://doi.org/10.1207/S15326985EP3801_8

Paas, F.G. and Van Merriënboer, J.G. (1994), "Instructional control of cognitive load in the training of complex cognitive tasks", *Educational Psychology Review,* Vol. 6 No. 4, pp. 351–371. https://doi.org/10.1007/BF02213420

Reid, G.B. and Nygren, T.E. (1988), "The subjective workload assessment technique: A scaling procedure for measuring mental workload", *Advances in Psychology*, Vol. 52, pp. 185–218. https://doi.org/10.1016/S0166-4115(08)62387-0

Rothblum, A.M. (2000), "Human error and marine safety", *National Safety Council Congress and Expo, Orlando, Florida, October 16-18, 2000*.

Russell, J.A. (1980), "A circumplex model of affect", *Journal of Personality and Social Psychology*, Vol. 39 No. 6, pp. 1161–1178. https://doi.org/10.1037/h0077714

Russell, J.A. and Barrett, L.F. (1999), "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant", *Journal of Personality and Social Psychology*, Vol. 76 No. 5, pp. 805-819. https://doi.org/10.1037/0022-3514.76.5.805

Russell, J.A., Bachorowski, J.-A. and Fernández-Dols, J.-M. (2003), "Facial and Vocal Expressions of Emotion", *Annual Review of Psychology*, Vol. 54, pp. 329–349. https://doi.org/10.1146/annurev.psych.54.101601.145102

Russell, J.A., Weiss, A. and Mendelsohn, G.A. (1989), "Affect grid: A single-item scale of pleasure and arousal", *Journal of Personality and Social Psychology,* Vol. 57, pp. 493–502. https://doi.org/10.1037/0022-3514.57.3.493

Sanders, M.S. and McCormick, E.J. (1987), *Human factors in engineering and design*, McGraw-Hill.

Shimmersense (2017a), Shimmer3 ECG/EMG Unit, Available at: http://www.shimmersensing.com/products/shimmer3-ecg-sensor

Shimmersense (2017b), Shimmer3 GSR+ Unit, Available at: http://www.shimmersensing.com/products/shimmer3-wireless-gsr-sensor

Ship Simulator Extremes (2010), *Ship Simulator Extremes.* [online] ShipSim.com. Available at: https://www.shipsim.com/products/shipsimulatorextremes

Sweller, J. (1988), "Cognitive load during problem solving: Effects on learning", *Cognitive Science*, Vol. 12 No. 2, pp. 257–285. https://doi.org/10.1207/s15516709cog1202_4

Thayer, R.E. (1967), "Measurement of Activation through Self-Report", *Psychological Reports*, Vol. 20 No. 2, pp. 663–678. https://doi.org/10.2466/pr0.1967.20.2.663

Thayer, R.E. (1986), "Activation-Deactivation Adjective Check List: Current Overview and Structural Analysis", *Psychological Reports,* Vol. 58 No. 2, pp. 607–614. https://doi.org/10.2466/pr0.1986.58.2.607

Thompson, E.R. (2007), "Development and Validation of an Internationally Reliable Short-Form of the Positive and Negative Affect Schedule (PANAS)", *Journal of Cross-Cultural Psychology*, Vol. 38 No. 2, pp. 227–242. https://doi.org/10.1177/0022022106297301

Tomkins, S. (1962), *Affect imagery consciousness: Volume 1: The positive affects,* Springer Publishing Company.

Tzannatos, E. (2010), "Human Element and Accidents in Greek Shipping", *The Journal of Navigation*, Vol. 63, pp. 119–127. https://doi.org/10.1017/S0373463309990312

Vidulich, M.A. and Tsang, P.S. (1987), "Absolute Magnitude Estimation and Relative Judgement Approaches to Subjective Workload Assessment", *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 31 No. 9, pp. 1057–1061. https://doi.org/10.1177/154193128703100930

Watson, D. and Tellegen, A. (1985), "Toward a consensual structure of mood", *Psychological Bulletin*, Vol. 98 No. 2, pp. 219-235. https://doi.org/10.1037/0033-2909.98.2.219

Watson, D., Clark, L.A. and Tellegen, A. (1988), “Development and validation of brief measures of positive and negative affect: The PANAS scales”, *Journal of Personality and Social Psychology*, Vol. 54 No. 6, pp. 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Westman, M. and Eden, D. (1996), “The inverted-U relationship between stress and performance: A field study”, *Work Stress*, Vol. 10 No. 2, pp. 165–173. https://doi.org/10.1080/02678379608256795

Wierwille, W.W. and Eggemeier, F.T. (1993), “Recommendations for Mental Workload Measurement in a Test and Evaluation Environment”, *Human Factors*, Vol. 35 No. 2, pp. 263–281. https://doi.org/10.1177/001872089303500205

Wilson, G.F. and Russell, C.A. (2003), “Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks”, *Human Factors*, Vol. 45 No. 4, pp. 635–644. https://doi.org/10.1518/hfes.45.4.635.27088

Woodson, W.E. and Conover, D.W. (1970), *Human engineering guide for equipment designers*, University of California Press, California.

Wulvik, A., Erichsen, J. and Steinert, M. (2016), “Capturing Body Language in Engineering Design – Tools and Technologies”, *Proceedings of the NordDesign 2016, Trondheim, Norway, August 10-12, 2016*, The Design Society, Bristol, pp. 165-174.

Andreas Simskar Wulvik, PhD Student
Norwegian University of Science and Technology - NTNU, Department of Mechanical and Industrial Engineering
Richard Birkelands Veg 2B, 7034 Trondheim, Norway
Email: andreas.wulvik@ntnu.no