MENTAL WORKLOAD (MWL) MEASUREMENT OF OFFICERS IN SIMULATED SHIP NAVIGATION: DETERMINING THE REDLINES OF PERFORMANCE

Reference NO. IJME727, DOI No: 10.5750/ijme.v164i1.727

B Özsever, Maritime Faculty, Piri Reis University, Turkey and **L Tavacıoğlu,** Maritime Faculty, Istanbul Technical University, Turkey.

KEY DATES: Submitted: 04/05/21; Final acceptance: 07/03/22; Published 15/06/22

SUMMARY

The main aim of this study is to measure the mental workload of the operators according to the increasing workload during simulated ship navigation and it is aimed to contribute to the clarification of upper redline of task demands. Eye responses and performance results of twelve participants were recorded during the measurements carried out in bridge simulator. In addition, a specific tool (NASA-TLX) was used to assess twelve participants at the end of each step of the scenarios. The results showed that mental workload of the participants increased as the task load increased and their performance decreased. It was observed that the developed Artificial Neural Network model can predict operator mental workload based on eye response indices (accuracy: 79.2%). This study is considered to contribute to the literature by defining an upper redline of task demands for an operator and monitoring near real-time mental workload indicators based on the physiological data of operators in the presence of autonomous ships and in navigational conditions where the automation level of ships gradually increases.

Keywords: mental workload, ship navigation, eye response, navigation performance**,** human factors**.**

NOMENCLATURE

ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under the Curve
CPA	Closest Point Approach
CSSI	Cognitive Seafarer-Ship Interface
F.I.	Fix Interval (minutes)
IMO	International Maritime Organization
M	Mean
MSE	Mean Squared Error
MWL	Mental Workload
p	Probability value
ROC	Receiver Operating Characteristic
SD	Standard Deviation
Sig.	Significance
t	t value
TCPA	Time to Closest Point of Approach
XTE	Cross Track Error

1. INTRODUCTION

Recent technological developments introduced autonomous ship concept that requires less seafarers on board. However, having duty persons on board, no matter how small in number, still makes human element an important subject for autonomous ships of the future. Within the four autonomous ship categories projected by IMO, only the fully autonomous ships will be operating with no seafarers on board. All the other three categories will require seafarers to be present either on board or ashore for remote controlling (IMO, 2018). This implies even with autonomous ships; human element will still be a major concern though the number of seafarers on board will be significantly reduced. This is because the remaining crew on board will have to continuously monitor and if necessary, intervene the operation of the ship. This obligation will require the seafarers to maintain their high cognitive states and optimal behaviours at all times when they are on duty. While human error is the primary contributor of accidents where about 85% of all accidents were caused by human error (Kurt et al., 2016), it was stated that 16% of collisions, 30% of groundings was related to mental fatigue of watchkeeping officers (Akhtar and Bouwer Utne, 2015) in furtherance the determination that technology and automation have reduced the number of crew and increased the workload of officers (Horberry et al., 2008; Louie and Doolen, 2007). This clearly indicates that human element related issues will continue to be one of the major issues in marine transportation assets (Özsever and Tavacıoğlu, 2019).

Workload is defined simplistically as a demand placed upon humans. Demand is determined by the aim to be achieved by the task performance. So, workload can be defined the effect of the demand on the individual in terms of the phases used in information processing and energetics. More specifically, workload is the amount of information processing capacity used for task performance (De Waard, 1996). Performance change is based on the balance between resource supply that is to say information processing capacity, and demand (Embrey et al., 2006). When resource demands exceed available supply, performance is assumed to be decreased. According to Kahneman (1973), the cognitive system has a single pool of limited capacity. Large amounts of resources are required for difficult tasks, especially when these tasks are coupled with concurrent tasks. On the contrary, easy and automated tasks require less resource with time sharing efficiency. However, mental workload (MWL) is directly neither performance nor task demand. Practice, experience, operator's state can affect the performance. Similarly, increasing level of skill can make individual need less mental effort (Sheridan and Simpson, 1979).

Mental workload, the effect of demand on operator, is an interaction between operator and task structure. Complexity and difficulty are the main characteristics of demand. Complexity is the number of stages of processing and difficulty is processing effort and it is related to amount of resources (De Waard, 1996). Mental workload, in terms of demand / resource balance, is a product of the resources available to meet the task demands (Young et al., 2015). Demand is determined by the aim to be achieved by the task performance and cannot be linked directly to workload. Assessment of workload is combined with task difficulty as experienced by the operator since the operator can give several reactions to the task demands such as adaptation or giving up (Gopher & Donchin, 1986 stated in De Waard (1996)). Although task performance cannot alone indicate any change in workload, suboptimal workload leads to errors and incidents. Suboptimal workload can be described either overload or underload. They stated the relationship between performance, task demand and resource supply in Figure 1. The left region of the red lines is called the 'reserve capacity' (underload) and right region is called the 'overload' region (Figure 1). In underload region task demands could be misperceived by operator and it could lead to performance decrement. Alternatively, in overload region when task demands exceed the resource supply, performance could be decreased (Young et al., 2015).

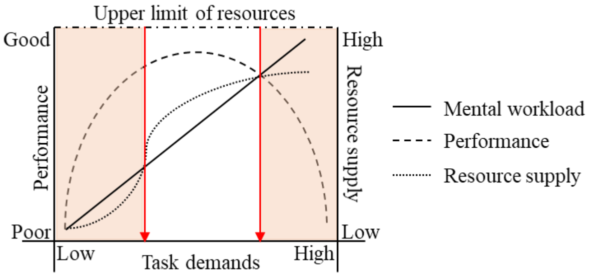


Figure 1. The relationship between task demand and resource supply associated with mental workload and performance (adapted from Young et al. (2015))

The maximum capacity of an operator has been limited to task circumstance. If the task is low demand task, operator cannot cope with any critical situation when he/she has suddenly faced with increased demand. Malleable Attentional Resource Theory (MART) clarifies why mental underload can lead to performance deterioration (Young and Stanton, 2002). This theory is more acceptable in navigational aspects because of that contains automation systems. Watchkeeping officer may not cope with the situation in case of any failure in automation systems or being exposed to unexpected danger when his/her attention decreases in non-traffic area with auto-pilot.

Mental workload measurement is relatively unknown in maritime domain, compared to other industries such as aviation, rail way, car driving etc. (Özsever and Tavacıoğlu, 2018). In maritime human factor research, there are several data collection methods related to mental workload or fatigue. These are physiological, physical (eye movement etc.), environmental measures, performance analysis in simulator environment, interviews, questionnaires, observations and log books, accident / incident analysis and computer-aided design / evaluations (Horberry et al., 2008). Commonly, physiological-physical, subjective and performance measures, which are defined as the components of triangulated measurement strategy (Wierwille and Eggemeier, 1993), have been used in workload measure studies (Embrey et al., 2006). However, acceptable level of workload still cannot be defined in maritime domain (Orlandi and Brooks, 2018). The studies conducted in recent years, have focused the MWL measurements in some maritime-specific tasks. Wu et al. (2017) associated the EEG and the HRV data, obtained from 10 participants in engine control room simulator, with MWL as task difficulty increased. Orlandi and Brooks (2018) applied similar method to ship pilots and reached similar results. Yan et al. (2019) used eye response measurement to predict MWL for engine department tasks. With the ANN classification success of eye response data and subjective ratings together with decreased performance results, the authors stated that eye response measurement can be used to predict MWL. Fan et al. (2021) evaluated the functional connectivity for watchkeeping and decision making during routine watchkeeping performance of the officers via fNIRS montage and found that the right lateral area of the prefrontal cortex has been sensitive to watchkeeping and decision making. This has opportunity to predict safety-critical performance.

Furthermore, as being one of the components of triangulated measurement strategy, performance measures of navigational duties have to be clearly defined and modelled in order to determine the redlines of performance and task load. Watch keeping officer experiences different cases those are not be able to evaluated with certain rules or limitations in regard to safety of navigation. So, the situations and the performance of officer on duty should be evaluated according to present conditions of traffic density, geography, visibility or navigational conditions. In literature, Gould et al. (2009) used the TARGETs method to assess performances of officers by expert evaluations. Task-generated (observable safety-critical navigation tasks) and event-generated (responses to external objects such as safe passing criteria; these are evaluated as "just acceptable or not" by experts) criteria were used in evaluation by experts. Besides, course deviation (XTE) and ship control (turn rate, rudder angle, speed) measures were scored in their study. In another study, course changes, rule following, target acquisitions, closest point of approach (CPA) and time to closest point of approach (TCPA), test manoeuvre, bearings taken, headings entered and track keeping were evaluated as task performance parameters (Robert et al., 2003). For example, keeping the CPA value more than 1 nm (nautical miles) is good performance while less than 0.8 nm is near miss and less than 0.5 nm is collision. While mean speed, mean frequency of engine rudder and course orders, mean frequency of fixes, CPA and XTE were used as performance measures for the landfall approach (Cook et al., 1981), fewer manoeuvring order command, fewer communication and more CPA were evaluated as better performance results (Grabowski and Sanborn, 2003). The operators were evaluated in three main criteria in a study; degree of deviation, decision making time and collision avoidance ability. Look out of other vessels, control of ship speed and course, position fixing, radio communication, collision avoidance (Embrey et al., 2006), detection range of targets, COLREG compliance, CPA, position report, communication and attention (Kircher and Lutzhoft, 2011) have been also used in performance measures conducted in the studies. On the other hand, Schuffel et al. (1989) used only XTE for performance measurement in their study.

Considering the above-mentioned elements, this study aims to measure mental workload of officers during the increased navigational task demands, adopting the self-reported and eye responses. With the comparison of MWL and the developed dynamic navigation performance measure, it is aimed to define upper redline of task demands in terms of safety of navigation in this study. The following hypothesis are studied:

- Different level of navigation tasks should draw out different levels of MWL and performance results.
- The developed performance measure for navigation tasks with eye response measurement is reliable for safety of navigation and can indicate the red lines of task demand.

2. METHODS

2.1 PARTICIPANTS

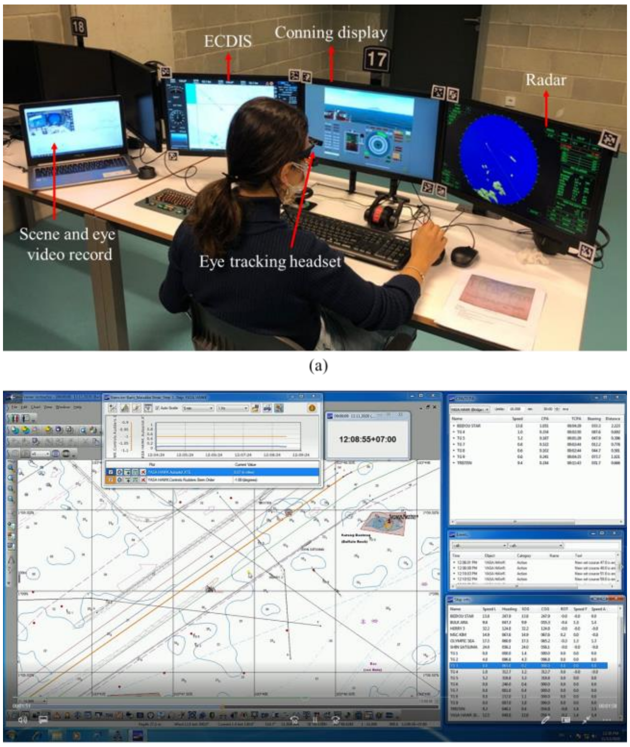
12 participants (5 female) were recruited to perform navigation scenario in bridge simulator in this study. At least, participants must have had an Oceangoing Watchkeeping Officer certificate and one contract sea experience as officer in merchant ships. The mean age was 28.4 (*SD* = 4.8) and the mean period of service of participants was 12.4 months (*SD* = 7.9). All participants gave informed consent form to be participant before performing the tasks in simulator. This study was approved by Medical and Engineering Sciences Human Research Ethics Committee of Istanbul Technical University.

2.2 EXPERIMENTAL TASK

The study was conducted in bridge simulator of Piri Reis University with navigation tasks based on Malacca Straight passage.

2.2 (a) Bridge Simulator

Participants performed the navigation tasks in bridge simulator (Figure 2a) located in Piri Reis University Seaside Campus Simulator Centre. The ship which was used for trials is a chemical tanker which has 183.0m length over all, 32.2m breadth with 60976.0t displacement and 13.0m maximum draft. The simulator has three screens which are ECDIS, RADAR and Conning Display that contains visual settings and auto pilot panel adding to one engine telegraph, one steering wheel. Navigational data was sampled at 1 Hz (TRANSAS, 2014). Additionally, the whole performance of participant as tracks on charts and other variables were recorded as video format from the computer located in control room (Figure 2b).



(a)

(b)

Figure 2. Bridge simulator (a), recording the participant performance (b)

2.2 (b) Tasks

Navigation scenarios have been varied being used different level of difficulties in mostly visibility, traffic density and geography parameters. Gould et al. (2009) used the variables as geography, visibility and traffic density for navigation scenario with 4 different levels of difficulty. Collision threat, target behaviour and traffic were used as variables for navigation scenario, which was conducted as 6 minutes and 18 times, in another study (Robert et al., 2003). Similar to the study of Gould et al. (2009), visibility, traffic density, geography, equipment condition and speed restriction were determined as difficulty variables in the study of Grabowski and Sanborn (2003). In this study, the difficulty level of navigation scenario was gradually adjusted according to traffic density, visibility and geography by combining in 4 steps as:

- Step 1; high visibility, low traffic density, easy geography
- Step 2; high visibility, moderate traffic density, easy geography
- Step 3; moderate visibility, high traffic density, moderate geography
- Step 4; low visibility, high traffic density, hard geography

Participants performed the navigation scenario according to the procedures stated in Table 1, in Malacca Strait, Singapore (Figure 3) because of that area has heavy traffic and there are lots of fishing boats and vessels making short cuts, make the passage more difficult.

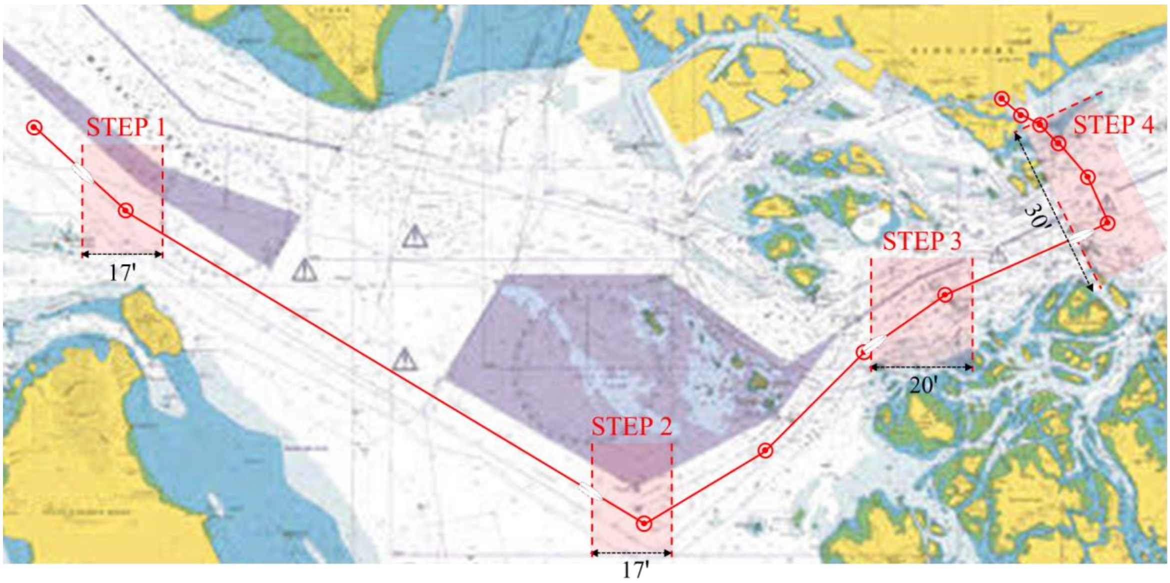


Figure 3. Navigation area used in simulator with route legs and performance measurement areas as stated in steps. Image obtained from Admiralty Chart BA 3833.

Table 1. Navigation in Malacca Strait procedure

Step	Task (sec.)	Task description
1	T1 (0-300)	Proceed to next waypoint with minimum XTE and detect the target on starboard bow side
	T2 (300-420)	React for collision avoidance
	T3 (420-570)	Make visible course change to starboard
	T4 (570-780)	Proceed with safe CPA
	T5 (780-1020)	Return to planned course
2	T1 (0-120)	Proceed to next waypoint with minimum XTE and detect the targets on head
	T2 (120-240)	Alter the course for safe CPA and for avoiding the fishing nets
	T3 (240-360)	Proceed with safe CPA and detect the target on starboard bow side
	T4 (360-480)	Proceed with safe CPA
	T5 (480-800)	Proceed with safe CPA and not be out of the traffic separation
	T6 (800-1020)	Alter the course to port for next waypoint and detect the fishing boat targets
3	T1 (0-240)	Proceed to next waypoint with minimum XTE by considering the fishing nets, detect the targets on port bow side
	T2 (240-300)	Alter the course for safe CPA and for avoiding the fishing nets
	T3 (300-420)	Proceed with safe CPA and not be out of the traffic separation
	T4 (420-540)	Proceed with safe CPA in decreased visibility and not be out of the traffic separation
	T5 (540-840)	Proceed with safe CPA and detect the target on starboard bow side
	T6 (840-1200)	Detect the target on starboard bow side and react for collision avoidance
4	T1 (0-360)	Proceed to next waypoint with minimum XTE, detect the targets on port bow side
	T2 (360-540)	Alter the course to port for next waypoint and proceed with safe CPA
	T3 (540-800)	Alter the course to starboard for safe CPA
	T4 (800-1100)	Return to planned course considering the current and detect the targets on port bow side
	T5 (1100-1250)	Proceed with safe CPA to fishing targets in more decreased visibility, detect the target on starboard side
	T6 (1250-1350)	Detect the fishing targets and proceed with safe CPA
	T7 (1350-1600)	Detect the fishing targets and proceed with safe CPA
	T8 (1600-1800)	Proceed to Loading Port with minimum XTE

The speed of own vessel is 10 to 13 knots and the XTE is 0.05 nm during the whole steps. Participants performed the navigation with auto pilot, but they can use hand steering for big course alterations and in emergency cases.

2.3 MEASUREMENT PROCEDURE

Before the measurements, the participants signed a consent form. The instructions of the scenarios were told to participants that they would be completing a navigation from middle of the Malacca Strait to entrance of Singapore Port as 4 steps in total 84 minutes. They were to navigate as they would in real-life, they communicated with simulator control room as if they communicate with other vessel and vessel traffic services. The performances of the participants were recorded via the computer located in simulator control room for the performance parameters. At the same time, eye measures were recorded. Additionally, the participants evaluated the their subjective MWL levels with NASA-TLX after the end of each step. To validate the performance scores of the participants, one ocean going Master expert evaluated the performances of the participants to assess the actions “just acceptable or not”.

2.3 (a) Performance Measurement

Navigation performances were evaluated using the targeted acceptable responses to generated events or tasks (TARGETS) method (Fowlkes et al., 1994). Differently, targets corresponding to the events were weighted according to the degree of importance in related event / task. Moreover, the performance results of the participants were scored as 0, 0.5 and 1 against the evaluation “just acceptable or not”. By the way, it was aimed to make performance measurement quantify in this study. In literature, Kim et al. (2010) tried to make performance measurement quantify, but they used constant limits for performances and that evaluation was not sufficient for variable navigational conditions. In a similar way stated in the study of Gould et al. (2009), tasks were evaluated separately as safety critical and track keeping in this study. Those were stated as task generated activities which are “observable safety-critical navigation tasks” and event-generated activities which are “responses to external objects”. Differently, performance scores were equal to the weighted sum of the scores of all criteria of both activities in this study.

Performance criteria were determined according to issues stated in literature and the opportunities of simulator environment (Table 2). 3 experts weighted the importance of each criteria for each step and for each task with fuzzy numbers because of that the level of importance of navigation criteria can vary to the navigational conditions.

Eventually, the performance score of the participant was calculated with the weighted sum of the score values:

$$P_i = \sum_{\alpha=1}^{p} w_{\alpha} \cdot \gamma_{\alpha} + \sum_{\nu=1}^{q} w_{\nu} \cdot \eta_{\nu} \qquad (1)$$

where $\gamma\alpha$ and $\eta\nu$ are the score values (0, 0.5 and 1) for safety critical navigation tasks and trackkeeping tasks respectively, where $w\alpha$ and $w\nu$ are the weights of safety critical navigation tasks and trackkeeping tasks respectively.

Table 2. Performance criteria for navigation scenario

Type of task	Main task	Detailed task	Symbol
Safety critical navigation tasks	Collision avoidance	Keeping a safe CPA	γ_{11}
		Rule following (COLREG)	γ_{12}
		Detection range of targets	γ_{13}
		Time to response	γ_{14}
		Communication & true reaction	γ_{15}
	Identify and communicate navigation landmarks		γ_2
	Identify hazards (report & action)		γ_3
Trackkeeping tasks	Crosstrack variability (XTE)		η_1
	Time to return to course		η_2
	Ship control	Rudder angle	η_{31}
		Turn radius	η_{32}
	Radar performance		η_4

To validate the performance scores of the participants, one ocean going Master expert evaluated the performances of the participants to assess the actions “just acceptable or not”. These evaluations were matched with the performance scores. The rates of true positive and false positive were analysed in ROC curves with the help of the thresholds set to performance score value. It was expected to assess the performances of officers with the help of the statistically significant threshold value of performance score. A receiver operating characteristic (ROC) is a technique for evaluating classifiers based on their performance (Fawcett, 2006). Graphical plot of sensitivity (true positive rate which is the ratio of positives correctly classified to total positives) is used to analyse the tendency of true positive and false positive rates (the ratio of negatives incorrectly classified to total negatives). The area under the ROC curve (AUC) has been used as statistical metric to show the accuracy of the classification (Fawcett, 2006). AUC value represents the classification performance - excellent (AUC > 0.9), good (0.8 < AUC < 0.9), fair (0.6 < AUC < 0.8) and failed (below 0.6) test (Singh et al., 2013).

2.3 (b) Subjective Rating

NASA-TLX was used in this study to evaluate the MWL levels of participants. This scale has been widely used in many human-machine interaction studies such as maritime (Gould et al., 2009; Orlandi and Brooks, 2018, Wu et al., 2017, Yan et al., 2019), aviation (Lehrer et al., 2010, Sirevaag et al., 1993), automobile (Sauer et al., 2013), traffic control centre (Fallahi et al., 2016), nuclear power plant (Gao et al., 2013) and mental task experiment (Miyake et al., 2009).

NASA-TLX is a multidimensional task load assessment tool, developed by Hart and Staveland (1988). NASA-TLX has 6 sub-scales which are mental, physical and temporal loads (task related), performance and effort (behavioural and skill related) and frustration (individual related). Participants weight the sub-scales to determine the intensity of each factor to total workload. Each sub-scale is evaluated independently from 0 to 20 and the sub-scales are weighted from 0 to 5. Finally, the weighted sum of the task load assessment is found as a score between 0 and 100 (Hart, 1986).

2.3 (c) Eye Response Measurement

Eye measures have been widely used in MWL studies. Pupil dilation occurs when task demand increases (Causse et al., 2010), but gives insufficient data to state the magnitude of arousal (Embrey et al., 2006). Pupil dilation is an autonomic sympathetic nervous system response that covers attention, interest or emotion (Bergstrom et al., 2014). Pupil diameter change is also correlated highly with error rate (Gao et al., 2013). Eye blink rate decreases when continued monitoring is required (Brookings et al., 1996; Ryu and Myung, 2005, Sirevaag et al., 1993; Veltman and Gaillard, 1996; Wilson, 2002) while closure duration and eye blink latency decrease with increased task demand (De Waard, 1996; Embrey et al., 2006). In high MWL, eye blink interval is longest and blink duration is shortest (Borghini et al., 2014; Hwang et al., 2008; Lean and Shan, 2012; Veltman and Gaillard, 1996).

Eye responses of the participants were recorded by Pupil Core eye tracking headset (Pupil Labs, Germany) including 1 eye camera with a rate of 200Hz at 192x192 px, 1 world camera with a rate of 60 Hz at 720p. Pupil Player Software version 2.1.0 was used to import data. Calibration of the device was carried out for each participant at the beginning of the performances. With the help of the headset, standard deviation of pupil diameter (Eq. 2), percentage of large pupil dilation (PerLPD) (Eq. 3), blink rate as frequency (Eq. 4), average eye closure duration (Eq. 5) and percentage of eye closure (PERCLOS) (Eq. 6) were analysed as the features of eye response in this study.

$$\text{pd\_std} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(D_i - \bar{D})^2} \tag{2}$$

$$\text{pd\_lpd} = \frac{(D_i - \bar{D})}{\bar{D}} \tag{3}$$

where D_i is the diameter of pupil and $\bar{D}$ is the mean of pupil diameter.

$$\text{br\_freq} = \frac{b}{t} \tag{4}$$

$$\text{br\_aecd} = \bar{d} \tag{5}$$

$$\text{br\_perclos} = \frac{\sum_{i=1}^{n} d_i}{t} \tag{6}$$

where b is the total number of blinks, t is the total duration, $\bar{d}$ is the mean of the closure durations of blinks and d_i is the closure duration of blink.

2.4 STATISTICAL ANALYSIS AND ARTIFICIAL NEURAL NETWORK MODEL

To correlate the relation between task load level and performance of officer and between task load level and eye responses data, correlation analysis was used. *t*-Test was used to test the difference of performance of officers and eye responses data at two workload levels. Additionally, ANOVA test was implemented to test the difference of NASA-TLX scores among 4 navigation steps. In all cases, α level of 0.05 was used to find out statistically significance. The statistical analysis was conducted using SPSS software, version 24.

ANN has been often used for classification in literature. This classifier has lots of advantages such as feedforward and backpropagation options, high processing speed, generalization ability. The classifier has a structure like a neuron (perceptron) which consists of similar input and output structure (Fausett, 1994; Polikar, 2006). The process of perceptron training includes the modifying the weights (connections between neurons) and finding the best weight. In literature, there are several training algorithms to form the relationship of input and output. Each neural network consists of the nodes, input layer, hidden layers and output layer. The number of hidden layers and nodes vary to the structure of the problem (Fausett, 1994). In this study, while five eye responses data form the input layer, two task load levels (low and high) form the output layer. In ANN structure, 2 hidden layers have been used (Figure 4). The data have been divided to training, validation and test in the ratio of 0.7, 0.1 and 0.2 respectively. ANN has been trained with random training data set. A *tansig* transfer function (Figure 5) have been used for hidden layers. In ANN structure, *trainlm* (Levenberg-Marquardt backpropagation) training function has been used as training method. To determine best classification structure of ANN, MSE values have been noted corresponding to the number of neurons (from 1 to 22) of hidden layers. ANN model has been structured and analysed in MATLAB R2014a.

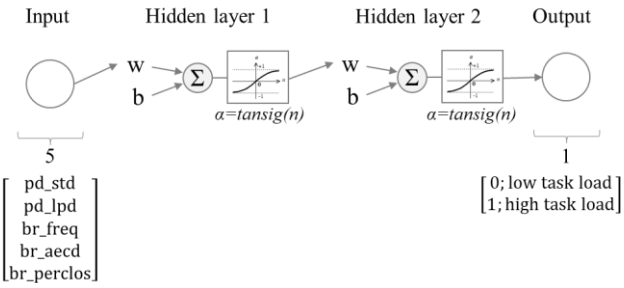


Figure 4. ANN structure used in this study.

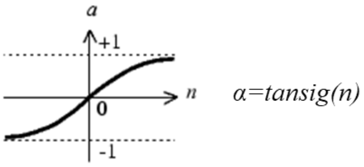


Figure 5. Tangent sigmoid (*tansig*) transfer function.

3. RESULTS

3.1 THE VALIDATION OF PERFORMANCE SCORES OF THE PARTICIPANTS

ROC curve analysis has been performed for validation of developed officer performance model. Recorded performances of the participants were evaluated as "just acceptable or not" for each task by one ocean going Master expert. According to the analysis, the value of AUC is 0.984 (p < 0.0001) (Sensitivity; 97.67, Specificity; 93.12) and the cut-off value is 52 for these performance scores (Figure 6).

3.2 PERFORMANCE DATA

Performance data show that there is a negative significant correlation between performance score and task load (p < 0.01). Correlation analysis are shown in Table 3. The results of performance measurement showed that the performance scores are significantly different (t = 3.967; p < 0.01) in low and high task loads. Table 4 presents the t-Test of performance data between low and high task load.

3.3 NASA-TLX SCORES

The NASA-TLX scores of each step evaluated by the participants have been statistically analysed and summarized in Table 5. ANOVA results show that there are significant differences in the NASA-TLX scores of 5 different dimensions and in total, among 4 steps which have different task load levels, i.e., MD (p < 0.01), P (p < 0.05), TD (p < 0.01), E (p < 0.01), F (p < 0.01) and total (p < 0.01).

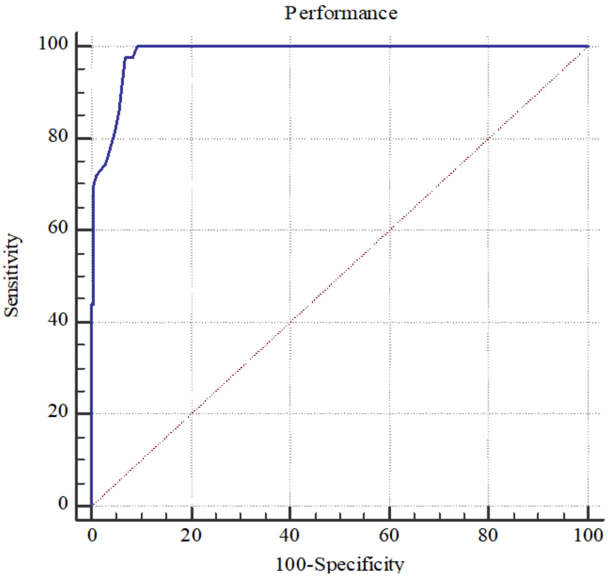


Figure 6. ROC curve analysis for developed officer performance model.

3.4 EYE RESPONSES DATA

Table 6 presents the correlation analysis between eye responses data and task load level. According to the analysis, large pupil dilation (pd_lpd) and lower average eye closure duration (br_aecd) are significantly correlated with higher task load. Table 7 shows the t-Test of eye responses data between low and high task load. According to the t-Test, there is a significant difference in only large pupil dilation (t = -2.618; p = 0.009). However, there are no significant differences in standard deviation of pupil diameter (pd_std), blink frequency (br_freq), average eye closure duration (br_aecd) and percentage of eye closure (br_perclos).

3.5 ANN MODEL ON EYE RESPONSES DATA TO CLASSIFY MWL

Data set was trained with various network structures. After the training the data set with all network structures (from 1 to 22 neurons for 2 hidden layers), 5-20-20-1 network structure was found to have minimum training, test and validation errors. The *MSE* values of all network structures are presented in Figure 7. Therefore, 5-20-20-1 ANN structure was selected for this model.

According to results of ANN model, the classification success was found as 79.2% (all data). ROC curve analysis and error matrix of training, test, validation and all data are given in Figure 8. Although there is no high classification success, ANN model used in this study has sufficient classification accuracy between high and low task load. This indicates that the prediction of MWL of officers can be realized based on eye responses data.

Table 3. Correlation between performance score and task load level

		Performance score	Task load level
Performace score	Spearman’s rho Correlation	1.000	
	Sig. (1-tailed)		
Task load level	Spearman’s rho Correlation	-0.484**	1.000
	Sig. (1-tailed)	<0.001	

**. Correlation is significant at the 0.01 level (1-tailed).

Table 4. *t*-Test of performance data between low and high task load.

	Low task load (*M* ± *SD*)	High task load (*M* ± *SD*)	*p*
Performace score	80.33 ± 19.099	68.48 ± 25.204	<0.001**

**. $p \leq 0.01$.

Table 5. ANOVA of NASA-TLX scores among 4 navigation steps.

	Step 1 (*M* ± *SD*)	Step 2 (*M* ± *SD*)	Step 3 (*M* ± *SD*)	Step 4 (*M* ± *SD*)	*p*
Mental demands	3.33 ± 2.15	10.22 ± 4.04	14.28 ± 5.71	20.03 ± 6.34	<0.001**
Performance	5.61 ± 5.16	5.17 ± 2.94	6.89 ± 4.21	10.00 ± 5.04	0.045*
Temporal demands	0.83 ± 1.19	6.36 ± 7.11	9.72 ± 7.21	14.53 ± 10.53	<0.001**
Efforts	3.33 ± 2.56	6.97 ± 4.89	9.39 ± 4.89	14.72 ± 5.68	<0.001**
Frustration	1.50 ± 1.27	6.31 ± 5.89	7.08 ± 5.23	13.75 ± 10.41	0.001**
NASA-TLX score	14.61 ± 8.97	35.03 ± 16.16	47.36 ± 14.24	73.03 ± 10.20	<0.001**

*. $p \leq 0.05$, **. $p \leq 0.01$.

Table 6. Correlation between eye responses data and task load level.

		Task load	pd_std	pd_lpd	br_freq	br_aecd	br_perclos
Task load	Spearman’s rho Correlation	1.000					
	Sig. (2-tailed)						
pd_std	Spearman’s rho Correlation	0.009	1.000				
	Sig. (2-tailed)	0.900					
pd_lpd	Spearman’s rho Correlation	0.189**	-0.212**	1.000			
	Sig. (2-tailed)	0.005	0.002				
br_freq	Spearman’s rho Correlation	-0.009	-0.024	0.119	1.000		
	Sig. (2-tailed)	0.893	0.723	0.078			
br_aecd	Spearman’s rho Correlation	-0.133*	0.279**	-0.092	0.221**	1.000	
	Sig. (2-tailed)	0.05	<0.001	0.173	0.001		
br_perclos	Spearman’s rho Correlation	-0.075	0.138*	-0.011	0.834**	0.633**	1.000
	Sig. (2-tailed)	0.269	0.040	0.868	<0.001	<0.001	

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Table 7. *t*-Test of eye responses data between low and high task load.

	Low task load (*M* ± *SD*)	High task load (*M* ± *SD*)	*p*
pd_std	2.412 ± 0.752	2.445 ± 0.716	0.741
pd_lpd	0.006 ± 0.065	0.031 ± 0.076	0.009**
br_freq	0.227 ± 0.094	0.229 ± 0.107	0.900
br_aecd	0.268 ± 0.144	0.248 ± 0.114	0.239
br_perclos	0.062 ± 0.041	0.058 ± 0.040	0.533

**. $p \leq 0.01$.

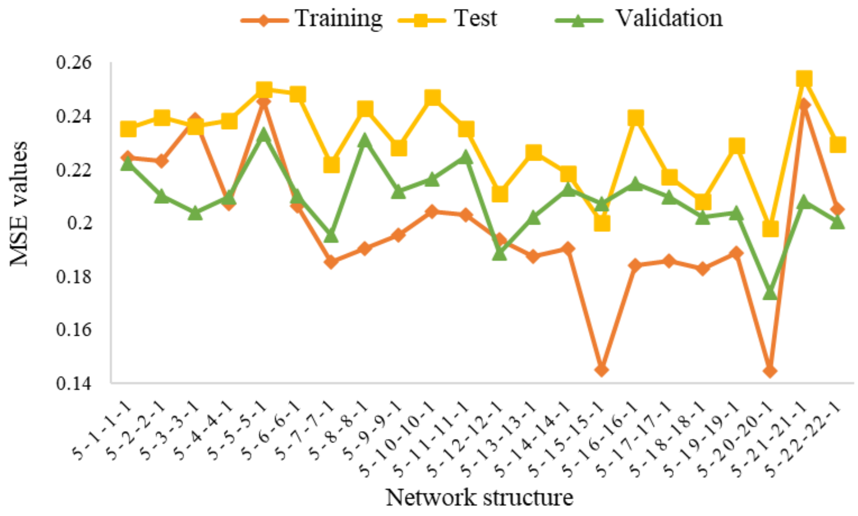


Figure 7. The *MSE* values corresponding to network structure.

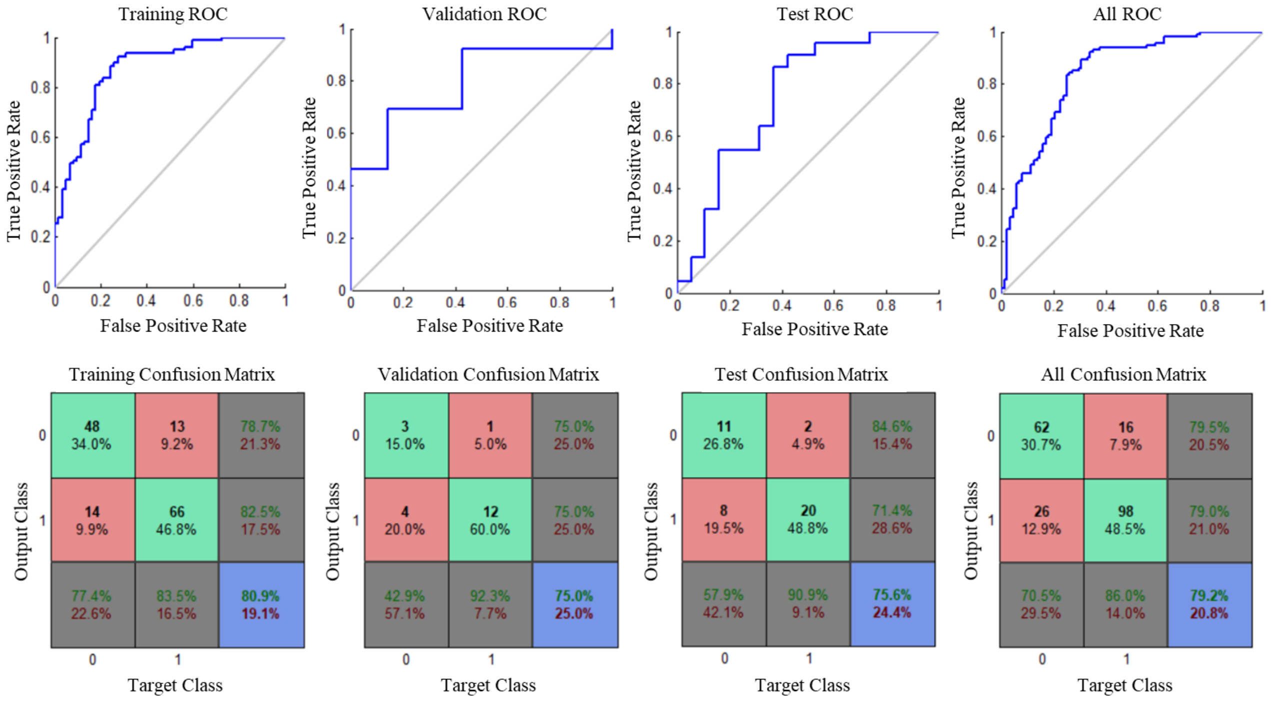


Figure 8. ROC curve analysis and error matrix of ANN classifier.

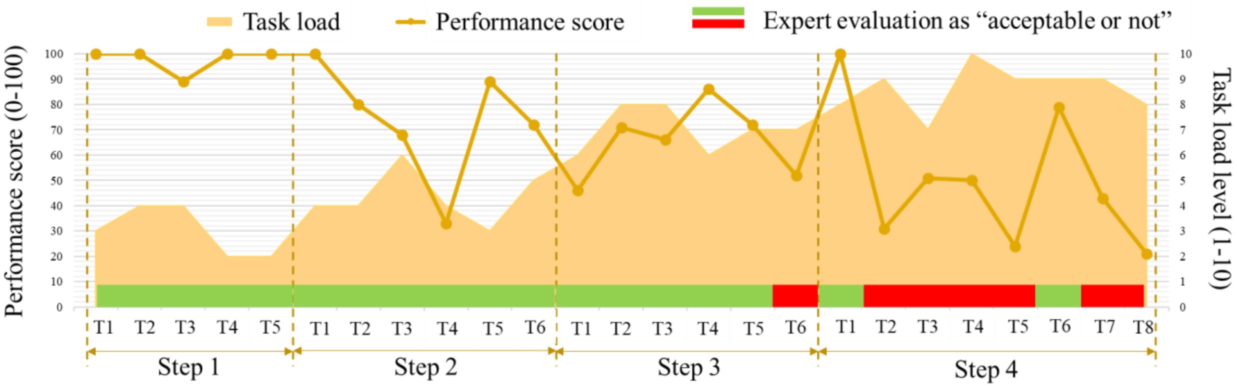


Figure 9. The performance – task load graphic of a participant.

4. DISCUSSION

This study aims to measure MWL of officers during the increased navigational task demands, adopting the self-reported and eye responses. The navigation scenario was created with different difficulty levels being gradually adjusted according to traffic density, visibility and geography by combining in 4 steps. Task load level (1-10) were determined according to the frequency of cognitive task transaction and the human information processing resources used in navigation tasks. This approach was adapted from OFM-COG analysis developed by Lee and Sanquist (2000). The results of developed officer performance model shows that there is a negative significant correlation between performance and task load. Figure 9 shows the task load-performance graphic of a participant. For all participants, when the task load is 7 and more, the performances of the participants have more tended to be evaluated as “not acceptable” by experts during the measurements. For this reason, the tasks which task load level is greater than or equal to 7, have been evaluated as “high task” and the tasks which task load level is less than 7, have been evaluated as “low task”. Moreover, the performance scores are significantly different in low and high task loads according to the statistical analysis.

Due to fact that the task performance cannot alone indicate any change in MWL (Young et al., 2015), the other measures (of triangulated measurement strategy) should be analysed. The results of NASA-TLX scores show that when task load increases, MWL perceived by the participants significantly increases. The score of physical demand that is one of the six dimensions of the scale has been evaluated as “0” because of that the participants haven’t perceived any physical demand during the steps of scenario. According to the results, mental demand is more dominant than others. The effect of temporal demand and effort follow mental demand. The reason for the low frustration effect compared to the others is thought to be due to the measurements being carried out in the simulator environment. This distribution of weights of the dimensions contributed NASA-TLX results can predict MWL for the navigation scenario.

Eye responses measurement have been widely used in MWL studies. Although the selectivity of eye blinks and pupil diameter to MWL is low (De Waard, 1996) and pupil dilation gives insufficient data to the magnitude of arousal (De Waard, 1996), it is stated in literature that pupil diameter and endogenous eye blinks are related to workload and ocular activity is more sensitive to visual demands. Hwang et al. (2008) stated in their study, which is conducted by simulated nuclear power plant tasks, eye blink interval is longest and blink duration is shortest when MWL is high. Similarly, in another study eye blink rate decreases when MWL increases during air traffic control tasks (Wilson and Russell, 2003).

In this study, the decrease of average eye closure duration was significantly correlated with the increase of MWL. This result contributed to literature. Additionally, large pupil dilation occurred when MWL increased. In maritime related studies, similar results have been stated. Pupil dilation occurred and blink duration decreased when task difficulty increased in engine room tasks (Yan et al., 2019). Similarly, pupil diameter of the participants increased when the complexity of berthing operations increased in a study which is conducted by marine pilots (Orlandi and Brooks, 2018).

According to classification efforts of eye responses on high task load and low task load levels and performance scores of the subjects, the red lines of task demand became apparent in this study. Continuing from the aim of Orlandi and Brooks (2018) and the contributions to MWL prediction in marine engine operations of Yan et al. (2019), the red lines of task demand in ship navigation was determined in this study. Classification of eye responses and the distinction of the task loads according to the performances of the subjects have ensured the task load to be separated as high task load and low task load.

As a future perspective, Seafarer-Centric Safety System focuses mainly the safety of the ship by taking the considerations of operational parameters and physiological parameters of the responsible operator. Therefore, the system needs the operational data from related equipment and the physiological data of the operator. Figure 10 presents the sample design for future Seafarer-Centric Safety System. The inputs of the Cognitive Seafarer-Ship Interface (CSSI) were formed with the outputs of high task load details for navigation and the eye responses given as features (classified in this study). CSSI processes the task loading together with physiological data of the officer and gives an output as “Risky” for safety of navigation in “The future Seafarer-Centric Safety System design” to be used on ships or at the Shore Control Centre for autonomous ships in future. It should be noted that this design can be used in all autonomous levels except the last level, the fully autonomous ship. Monitoring and evaluation of the officer in charge will be important for ship safety, as the workload will be present during navigation (which will vary according to the system design) at the autonomous levels where the human is in the system by operating or monitoring.

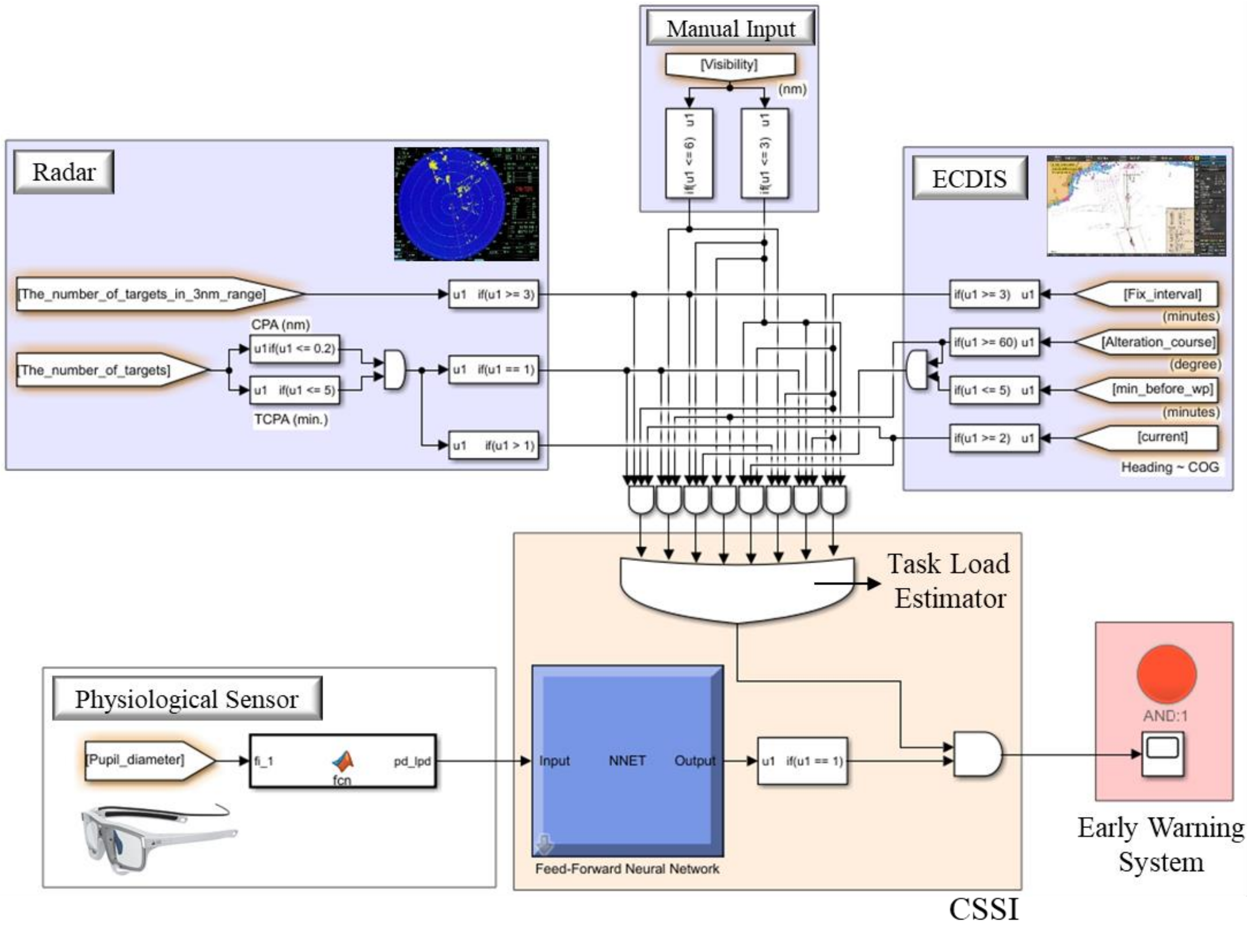


Figure 10. The detailed future Seafarer-Centric Safety System design (created in Matlab 2020a Simulink).

According to the design, task load estimator processes the data which are the possible combinations of the outputs from ECDIS, Radar and manual input. These combinations stated in this design, are the high task load indicators which have been tested in this study. Therefore, the combinations that can be evaluated as high task load should be increased in future studies. At the same time, neural network stated in CSSI, process the inputs which are physiological features extracted from physiological sensors and gives an output according to the structure of ANN. When the output of neural network is 1 (indicated as “High task load” in this study) and one of the possible combinations exists in task load estimator, CSSI gives an output for early warning system to be activated. It was stated before that similar study for aircraft was conducted by Liu et al. (2016). One cognitive pilot-aircraft interface was designed with environmental variables of flight and physiological variables of the pilot. The cognitive pilot-aircraft interface can give an output to adjust the level of auto pilot considering the mental strain of pilot and the task load of environmental variables of flight.

However, the method stated in this study has some limitations and assumptions to be underlined. Simulator environment was chosen for measurements due to fact that measurement on real environment on board is dangerous and is difficult to obtain repeatable results of operator errors. The sample group for this research consists of junior deck officers who have minimum one contract sea service. Although it is known that most of maritime accidents result from the deficiencies in cooperation of Master-Pilot-Officer during pilotage or manoeuvres, in one-third of all accidents one officer keeps watch at the bridge (Yıldırım et al., 2019). On the other hand, experience is a major contributor for coping with stressor factors (Jeżewska and Iversen, 2012; Salyga and Kusleikaite, 2011). Considering all of above-mentioned reasons, junior officers are selected for this research and the measurements were taken from the subjects in simulators as if they keep watch alone at the bridge. One of the limits of the study is that the sample group consists of only junior deck officers. Universal usability of the method stated in this study, for all ranks of seafarers and for all specified seaborn operations has to be researched in future studies.

5. CONCLUSIONS

MWL, the effect of demand on operator, is an interaction between operator and task structure. Under conditions where this interaction is unstable, human errors occur. It is known that these errors have serious consequences, especially in maritime. In this study, it was aimed to measure the MWL of the operators according to the increasing workload during simulated ship navigation and it was aimed to contribute to the clarification of upper redline of task demands. Apart from the performance measurement methods used in the literature, a dynamic officer performance measurement method has been developed and validated. Eye responses and NASA-TLX results taken during the measurements showed that the MWL of the participants increased as the task load increased and their performance decreased. According to the ANN model developed in the study, it was seen that the eye responses values can be divided into two classes (as “safe” and “risky” in terms of safety of navigation).

The results of this study also showed that the task demand, whose upper red line became apparent, drew a set of projections and possibilities in which navigation conditions could not be performed by a single operator. This study will contribute to the literature in terms of defining an upper redline of task demands for an operator and monitoring near real-time MWL indicators based on physiological data of operator in the presence of autonomous ships and in navigational conditions where the automation level of ships gradually increases.

6. ACKNOWLEDGEMENTS

This study is part of PhD thesis of first author and was supported by Scientific Research Projects Department of Istanbul Technical University. Project Number: 41710. The authors would like to thank the experts and ocean-going marine officers who are the participants in this experimental research for their valuable contributions.

7. REFERENCES

1. AKHTAR, M. J. & BOUWER UTNE, I. 2015. Common patterns in aggregated accident analysis charts from human fatigue-related groundings and collisions at sea. *Maritime Policy & Management*, 42, 186-206.
2. BERGSTROM, J. R., DUDA, S., HAWKINS, D. & MCGILL, M. 2014. Physiological Response Measurements. *Eye Tracking in User Experience Design*. Elsevier, Waltham, Massachusetts/USA.
3. BORGHINI, G., ASTOLFI, L., VECCHIATO, G., MATTIA, D. & BABILONI, F. 2014. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44, 58-75.
4. BROOKINGS, J. B., WILSON, G. F. & SWAIN, C. R. 1996. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological psychology*, 42, 361-377.
5. CAUSSE, M., SÉNARD, J.-M., DÉMONET, J. F. & PASTOR, J. 2010. Monitoring cognitive and emotional processes through pupil and cardiac response during dynamic versus logical task. *Applied psychophysiology and biofeedback*, 35, 115-123.
6. COOK, R., MARINO, K. & COOPER, R. 1981. A Simulator Study of Deepwater Port Shiphandling and Navigation Problems in Poor Visibility. ECLECTECH ASSOCIATES INC NORTH STONINGTON CT.
7. DE WAARD, D. 1996. *The measurement of drivers' mental workload,* Groningen University, Traffic Research Center Netherlands.
8. EMBREY, D., BLACKETT, C., MARSDEN, P. & PEACHEY, J. 2006. Development of a human cognitive workload assessment tool. *Dalton (UK): Human Reliability Associates*.
9. FALLAHI, M., MOTAMEDZADE, M., HEIDARIMOGHADAM, R., SOLTANIAN, A. R. & MIYAKE, S. 2016. Effects of mental workload on physiological and subjective responses during traffic density monitoring: A field study. *Applied ergonomics*, 52, 95-103.
10. FAN, S., BLANCO-DAVIS, E., ZHANG, J., BURY, A., WARREN, J., YANG, Z., YAN, X., WANG, J. & FAIRCLOUGH, S. 2021. The role of the prefrontal cortex and functional connectivity during maritime operations: an fNIRS study. *Brain and Behavior*, 11, e01910.
11. FAUSETT, L. V. 1994. *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall Englewood Cliffs, New Jersey/USA.
12. FAWCETT, T. 2006. An introduction to ROC analysis. *Pattern recognition letters*, 27, 861-874.
13. FOWLKES, J. E., LANE, N. E., SALAS, E., FRANZ, T. & OSER, R. 1994. Improving the measurement of team performance: The TARGETs methodology. *Military Psychology*, 6, 47-61.
14. GAO, Q., WANG, Y., SONG, F., LI, Z. & DONG, X. 2013. Mental workload measurement for emergency operating procedures in digital nuclear power plants. *Ergonomics*, 56, 1070-1085.
15. GOULD, K. S., RØED, B. K., SAUS, E.-R., KOEFOED, V. F., BRIDGER, R. S. & MOEN, B. E. 2009. Effects of navigation method on workload and performance in simulated high-speed ship navigation. *Applied ergonomics*, 40, 103-114.
16. GRABOWSKI, M. & SANBORN, S. D. 2003. Human performance and embedded intelligent technology in safety-critical systems. *International journal of human-computer studies*, 58, 637-670.
17. HART, S. G. 1986. *NASA Task load Index (TLX). Volume 1.0; Paper and pencil package*. NASA Ames Research Center, California/USA.

18. HART, S. G. & STAVELAND, L. E. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, 52, 139-183.
19. HORBERRY, T., GRECH, M. & KOESTER, T. 2008. *Human factors in the maritime domain*, CRC press, Boca Raton, Florida/USA.
20. HWANG, S.-L., YAU, Y.-J., LIN, Y.-T., CHEN, J.-H., HUANG, T.-H., YENN, T.-C. & HSU, C.-C. 2008. Predicting work performance in nuclear power plants. *Safety science*, 46, 1115-1124.
21. IMO. 2018. *Maritime Safety Committee (MSC), 100th session, 3-7 December 2018* [Online]. Available: http://www.imo.org/en/MediaCentre/MeetingSummaries/MSC/Pages/MSC-100th-session.aspx [Accessed 23.02.2019].
22. JEŻEWSKA, M. & IVERSEN, R. 2012. Stress and fatigue at sea versus quality of life. Gdansk, 11 June 2012. II International Congress on Maritime, Tropical, and Hyperbaric Medicine. *International maritime health*, 63, 106-115.
23. KAHNEMAN, D. 1973. *Attention and effort*. Prentice-Hall Englewood Cliffs, New Jersey/USA.
24. KIM, H., KIM, H. & HONG, S. 2010. Collision Scenario-based Cognitive Performance Assessment for Marine Officers.
25. KIRCHER, A. & LUTZHOFT, M. Performance of seafarers during extended simulation runs. International Conference on Human Factors in Ship Design and Operation, 2011. 53-59.
26. KURT, R. E., KHALID, H., TURAN, O., HOUBEN, M., BOS, J. & HELVACIOGLU, I. H. 2016. Towards human-oriented norms: Considering the effects of noise exposure on board ships. *Ocean Engineering*, 120, 101-107.
27. LEAN, Y. & SHAN, F. 2012. Brief review on physiological and biochemical evaluations of human mental workload. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 22, 177-187.
28. LEE, J. D. & SANQUIST, T. F. 2000. Augmenting the operator function model with cognitive operations: Assessing the cognitive demands of technological innovation in ship navigation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30, 273-285.
29. LEHRER, P., KARAVIDAS, M., LU, S.-E., VASCHILLO, E., VASCHILLO, B. & CHENG, A. 2010. Cardiac data increase association between self-report and both expert ratings of task load and task performance in flight simulator tasks: An exploratory study. *International Journal of Psychophysiology*, 76, 80-87.
30. LIU, J., GARDI, A., RAMASAMY, S., LIM, Y. & SABATINI, R. 2016. Cognitive pilot-aircraft interface for single-pilot operations. *Knowledge-Based Systems*, 112, 37-53.
31. LOUIE, V. W. & DOOLEN, T. L. 2007. A study of factors that contribute to maritime fatigue. *Marine Technology*, 44, 82-92.
32. MIYAKE, S., YAMADA, S., SHOJI, T., TAKAE, Y., KUGE, N. & YAMAMURA, T. 2009. Physiological responses to workload change. A test/retest examination. *Applied ergonomics*, 40, 987-996.
33. ORLANDI, L. & BROOKS, B. 2018. Measuring mental workload and physiological reactions in marine pilots: Building bridges towards redlines of performance. *Applied Ergonomics*, 69, 74-92.
34. ÖZSEVER, B. & TAVACıOĞLU, L. 2018. Analysing the effects of working period on psychophysiological states of seafarers. *International Maritime Health*, 69, 84-93.
35. ÖZSEVER, B. & TAVACıOĞLU, L. 2019. An Extensive Research into Possibility of a Human-Centered Safety System for Fatigue Detection at Sea. *III. Global Conference on Innovation in Marine Technology and the Future of Maritime Transportation*. Izmir, Turkey.
36. POLIKAR, R. 2006. Pattern Recognition. *Wiley Encyclopedia of Biomedical Engineering*. New Jersey/USA.
37. ROBERT, G., HOCKEY, J., HEALEY, A., CRAWSHAW, M., WASTELL, D. G. & SAUER, J. 2003. Cognitive demands of collision avoidance in simulated ship control. *Human factors*, 45, 252-265.
38. RYU, K. & MYUNG, R. 2005. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, 35, 991-1009.
39. SALYGA, J. & KUSLEIKAITE, M. 2011. Factors influencing psychoemotional strain and fatigue, and relationship of these factors with health complaints at sea among Lithuanian seafarers. *Medicina (Kaunas)*, 47, 675-681.
40. SAUER, J., NICKEL, P. & WASTELL, D. 2013. Designing automation for complex work environments under different levels of stress. *Applied ergonomics*, 44, 119-127.
41. SCHUFFEL, H., BOER, J. & VAN BREDA, L. 1989. The ship's wheelhouse of the nineties: the navigation performance and mental workload of the officer of the watch. *The Journal of Navigation*, 42, 60-72.
42. SHERIDAN, T. B. & SIMPSON, R. 1979. Toward the definition and measurement of the mental workload of transport pilots. Cambridge, Mass.: Massachusetts Institute of Technology, Dept. of Aeronautics and Astronautics, Flight Transportation Laboratory, [1979].

43. SINGH, R. R., CONJETI, S. & BANERJEE, R. 2013. A comparative evaluation of neural network classifiers for stress level analysis of automotive drivers using physiological signals. *Biomedical Signal Processing and Control*, 8, 740-754.
44. SIREVAAG, E. J., KRAMER, A. F., REISWEBER, C. D. W. M., STRAYER, D. L. & GRENELL, J. F. 1993. Assessment of pilot performance and mental workload in rotary wing aircraft. *Ergonomics*, 36, 1121-1140.
45. TRANSAS 2014. *Navi-Trainer Professional 5000 Instructor Manual*. Transas MIP Ltd, Portsmouth, United Kingdom.
46. VELTMAN, J. & GAILLARD, A. 1996. Physiological indices of workload in a simulated flight task. *Biological psychology*, 42, 323-342.
47. WIERWILLE, W. W. & EGGEMEIER, F. T. 1993. Recommendations for mental workload measurement in a test and evaluation environment. *Human factors*, 35, 263-281.
48. WILSON, G. F. 2002. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, 12, 3-18.
49. WILSON, G. F. & RUSSELL, C. A. 2003. Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors*, 45, 381-389.
50. WU, Y., MIWA, T. & UCHIDA, M. 2017. Using physiological signals to measure operator’s mental workload in shipping–an engine room simulator study. *Journal of Marine Engineering & Technology*, 16, 61-69.
51. YAN, S., WEI, Y. & TRAN, C. C. 2019. Evaluation and prediction mental workload in user interface of maritime operations using eye response. *International Journal of Industrial Ergonomics*, 71, 117-127.
52. YıLDıRıM, U., BAŞAR, E. & UĞURLU, Ö. 2019. Assessment of collisions and grounding accidents with human factors analysis and classification system (HFACS) and statistical methods. *Safety Science*, 119, 412-425.
53. YOUNG, M., BROOKHUIS, K., WICKENS, C. & HANCOCK, P. 2015. State of science: mental workload in ergonomics. *Ergonomics*, 58, 1-17.
54. YOUNG, M. S. & STANTON, N. A. 2002. Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. *Human factors*, 44, 365-375.