Lochner et al. Journal of Shipping and Trade (2018) 3:2
DOI 10.1186/s41072-018-0028-z

Journal of Shipping and Trade

ORIGINAL ARTICLE Open Access

Analysis of maritime team workload and communication dynamics in standard and emergency scenarios

Martin Lochner[1*], Andreas Duenser[1], Margareta Lutzhoft[2], Ben Brooks[2] and David Rozado[3]

* Correspondence: martin.lochner@utas.edu.au
[1]Commonwealth Scientific and Industrial Research Organisation, 15 College Road, Sandy Bay 7005, Australia
Full list of author information is available at the end of the article

Abstract

The introduction of next-generation technologies to the maritime shipping industry, including Portable Pilotage Units, Remote Pilotage, advanced situation awareness aids, and Autonomous Shipping, creates an urgent need to understand operator workload during Bridge Team operations, and co-operations with shore based personnel. In this paper we analyse mental workload of maritime Captains, Pilots and Tug Masters during standard and emergency scenarios, using traditional measures (SWAT, ISA), communications analysis, and the collection of simultaneous electro-dermal activity (EDA) of team members. Results indicate that the EDA measure overcomes some of the problems with paper-based techniques, and has excellent temporal resolution for emergency events. Implications for testing of novel technologies are discussed.

Introduction

The need to understand operator workload is a key requirement across numerous sectors, including maritime shipping (Lützhöft et al., 2011), nuclear power operations (Sheridan, 1981), air traffic control (Loft et al., 2007), driving (Trick et al., 2009), and many other contexts that impose a high demand on the human attentional system. While the physical elements of workload are generally well understood (e.g. De Zwart et al., 1996), the concept of cognitive workload, or mental workload, is less self-evident. One classic definition of mental workload by Hart & Staveland (Hart & Staveland, 1988) is "the perceived relationship between the amount of mental processing capability or resources and the amount required by the task". Human attention is by nature a limited resource, and decades of research have been conducted into its strengths and its limitations. We have a remarkable ability to divide attention across multiple foci, both in physical space, and conceptually. Nevertheless, under certain conditions our comprehension of a situation can break down, with the result being that accidents happen, causing damage to property, human life, and to the environment. In the context of shipping and trade, developing a clear methodology for measuring maritime operator workload has the potential benefit of improving efficiency and safety, by better understanding the human error component that is common in many maritime accidents. In addition to understanding workload of specific

individuals during emergency events, the research reported here also investigates how members of a maritime operations team, including tug operators, Vessel Traffic Service, pilots, and the bridge team react together to deal with emergency events. Our results have implications when considering how to test human performance with novel technological systems, both on the ship's bridge and at remote locations such as a tug, or VTS facility.

Efficiency and safety have long been key drivers of change in the maritime industry. Because of the large volumes and profits involved, and the critical nature of maritime accidents, technological solutions to age old problems of navigation have been employed to various degrees during the past half-century. Technological systems including RADAR, SONAR, GPS, VTS, ECDIS, AIS, and others, have benefitted maritime operations in many ways, but have not necessarily resulted in lower operator workload. Rather, in many cases the result has been just the opposite, where operators in high workload situations have a tendency to ignore maritime decision support systems (Grabowski & Sanborn, 2001). Furthermore, there are examples of technology contributing to failures, for example RADAR (Andrea Doria-Stockholm incident, 1956) and ECDIS (Ovit incident, 2013). To give an idea of the hybrid complexity that can exist on a ship's bridge, one of the present authors (Lützhöft & Nyce, 2014) reports that a container vessel that was manufactured in the 1960s, and which had been converted to a passenger liner in 1990 prior to being inspected by the author in 2001, had an assortment of 15 different manufacturer's brands on the bridge equipment and an offshore supply ship built in 2005 had close to 30 brands. The integration work required to safely operate such a system is a clear strain on the operator's physical and mental capacity (Lützhöft, 2004). It is no surprise then, that a large proportion of modern maritime accidents is attributed to human error, which in turn has been directly linked to mental workload (Hetherington et al., 2006). Note, that this in turn does not necessarily mean an overt mistake was performed by a human.

While there is some general agreement that mental workload is the culprit in many maritime accidents, and thus should be the subject of investigation, there is no such concord on the best way to operationalize the concept of mental workload. A number of methodologies have been employed to this end, each of which has advantages and disadvantages (Tsang & Vidulich, 2006). For the current research we first provide a review of the main methods available in the literature, discuss the most commonly used techniques, and provide a rationale for our use of a simple and effective electrophysiological technique known as Galvanic Skin Response (GSR), or alternately as Electro-Dermal Activity (EDA).

We present three case studies in this paper. In each case, the participants were experienced maritime professionals consisting of Ships Master / Captain (responsible for safe conduct of ship), Pilot (a local addition to the bridge team, who in practice takes over the manoeuvring and leads the communication), Tug Master (tugs are small powerful vessels that assist in manoeuvring large ships in restricted waters, either connected by rope/wire or pushing), Helmsman (performs the steering, on orders from master/pilot but no other tasks), and Vessel Traffic Service (VTS) Operator (VTS is a shore-based information service, much like air traffic

control but with no mandate to give orders). First of all, we present an analysis of operator workload using the ISA (Instantaneous Self Assessment) and the SWAT (Subjective Workload Assessment Technique) that are commonly employed in the literature (Cain, 2007). These measures were chosen for their prevalence in the workload literature, and because they are straightforward to administer and analyse. Second, an analysis of communications patterns during emergency manoeuvres is presented as an additional means of understanding operator workload within the maritime environment. These studies illustrate some of the drawbacks to using the standard ISA/SWAT methodology, and provide some insight regarding communication patterns during an emergency event. Finally, we conducted a series of maritime operations while collecting GSR/EDA measures for the key team players: the Captain/Master, the Pilot, and the Tug Master. The use of GSR/EDA measurement allowed us to collect workload measures from a distributed team of maritime personnel as they performed routine and emergency manoeuvres in a large maritime ship simulator. It has the clear advantage of detecting the onset and relative level of operator stress (a robust correlate of mental workload), and further, of capturing this information for multiple individuals within a distributed team operating environment.

Mental or 'cognitive' workload

The ability to assess and understand human performance, particularly in critical tasks where the actions carried out have major significance for safety and productivity, is a long standing goal of human factors research. Stemming from capacity-based models of human cognition, e.g., (Wickens, 2008; Baddeley, 1992), the concept of cognitive workload is based on the notion that as task demands increase, the individual is required to exert an increasing amount of his or her limited cognitive resources to maintain a steady level of performance. Workload has been assumed to follow the 'Yerkes-Dodson law' (i.e. 'the inverted U') where performance improves for low to medium levels of workload, but drops with higher workload levels (Staal, 2004). Increasing evidence, however, informs us that the true pattern may vary depending on the type of activity, and environmental characteristics. In terms of human cognition, there is evidence that performance tends to decrease in a linear relation to workload (Marshall, 2002).

Individual workload has been assessed in many ways, and a number of detailed reviews are available (e.g., Tsang & Vidulich, 2006). We will briefly mention the main categories, and discuss their applicability to studying team workload in a safety-critical environment – namely the ship's bridge. First, *primary* measures of performance on the given task can be used to infer workload. This means a direct measure of performance on the task of interest, with the notion that decreased performance indicates high workload. It should be evident, however, that task performance can be affected by other factors besides workload (e.g., competence, distraction, equipment failure, etc.). Further, a performance failure in such environments could be catastrophic, and it is therefore desirable to have an alternate measure to pick up load before a failure.

Secondary task methods of measuring cognitive workload involve the addition of a so-called secondary task, performance on which varies depending upon the

hypothesized 'spare capacity' remaining for the user. Wickens' (Wickens, 2008) influential Multiple Resource Theory (MRT) takes advantage of this framework, and evolves the concept to include separate resources for different processing modalities such as for visual versus manual information. MRT has empirically shown that processing resources are parsed along the lines of modality, where a visual task and a manual task, for example, may be performed without immediate processing conflicts. Despite such successes, however, the inclusion of secondary task is not generally ecologically valid - for example requiring a participant to detect the onset of a peripheral light while executing the main task concurrently - and can be assumed to impact performance on the primary task. As such, the addition of a secondary task may be considered impractical for assessing performance in a safety-critical environment.

A very popular method of evaluating user workload is to employ subjective questionnaires which are administered to the participant either during[1] or after[2] the activity of interest. Such methodology is wide-spread in the literature (Funke et al., 2012) provide a review including 18 such studies, spanning from 1987 to 2010; and this is certainly only a cross-section). While easy to administer, and relatively informative and easy to interpret, such methodology has some major problems. In the case of the mid-trial tasks, the questionnaire breaks up the flow of the task – again impacting ecological validity, and limiting its usefulness in real-world situations. The post test measures are also troublesome, as the results are based on the recall of mental workload, rather than an immediate index at the time of interest. In both cases, the tests are not generally suitable for mission-critical situations where the task flow cannot be interrupted.

A solution to these difficulties exists in physiological (Tsang & Vidulich, 2006) or physio-behavioural (Funke et al., 2012) measurements. A number of techniques available today can directly tap into the physiological signals of individuals, thus gaining insight regarding physiological states, which have been previously shown to be associated with the concept of cognitive workload (Funke et al., 2012). These include but are not limited to: brain imaging techniques (functional MRI, Positron Emission Tomography), electrophysiological techniques (electroencephalography, electromyography, galvanic skin response / electro-dermal activity), respiration, heart rate, heart rate variability, and blood flow. Although almost any measure of physiology can be a potential indicator for stress, and accordingly, workload, the number becomes more restricted when we consider that the measure is desirably carried out within the expected operational environment, with no (or minimal) impact on the ecological validity of the task (i.e. minimally affecting the standard or typical operation). In short, we want to test workload without interfering with the task. To this end, in the current set of studies we first show how typical measures of workload perform in our operational context. Second, we develop our understanding of workload in the maritime operations context by investigating communications patterns within the maritime operations team. Finally, we investigate EDA as a simple but effective technique to measure workload.

Team workload

While individual cognitive workload (Sweller et al., 2011) has been considered extensively over the past 20 years, research into workload within a team setting is still gaining

momentum in the literature (Funke et al., 2012). Despite the possibility that factors influencing team workload may be characteristically different (or more complex) from those impacting individual workload, it is often the case that team workload is nevertheless measured using an amalgamation of the individual workload measurements. For example, a researcher may collect workload ratings from each member in the team, and create a team average from these ratings. The applicability of individual workload measures to a team setting is open for discussion; however, well-accepted measures of team workload are currently unavailable. Although there are many ways to assess individual workload, the suitability of these for assessing team workload in a realistic setting must be considered. In this research, we employ simultaneous measurement of workload for team members, and use a team average for comparison across workload levels.

Simulation #1: Using ISA / SWAT to measure team workload

Much of the prior research into cognitive workload has employed subjective paper-based tests of workload, as described above. In order to determine the suitability of such questionnaires for measuring team workload in a maritime operations setting, we tested the maritime operations team (Captain/Master, Pilot, Tug Master and Helmsman) within standard training runs in the maritime full-mission simulator, using both the ISA and the SWAT metrics.

Method

Participants

Participants were all experts recruited for the study. Participants consisted of a distributed maritime operations team, including an experienced Captain (male, 31 years old, 5 years experience in current role), Pilot (male, 52 years old, 10 years experience in current role), Helmsman (male, 65 years old, unknown years experience), and Tug Master (male, 50 years old, 2 years experience in current role).

Apparatus

The simulators used were the Australian Maritime College (AMC) full mission ship's bridge and tug simulators.

Kongsberg Full Mission Ship Simulator: The current simulation was carried out in a Kongsberg custom maritime simulator at the Australian Maritime College, in Launceston, Tasmania. This simulator suite includes a maximum of (Marshall, 2002) separate operational rooms, seven of which (excluding the main bridge and main control room) can be setup in various arrangements, such as tugs, secondary ships etc. (Fig. 1).

Design and procedure

Two consecutive days of testing were undertaken in the AMC maritime simulator. For each day, participants went through two runs, one of high workload, and one of low workload. In this case, workload was specifically controlled by manipulating simulator parameters such as wind speed, current (and therefore drift), as well as concurrent maritime traffic. Both ISA and SWAT scores were collected. The ISA is a single numeric workload rating (from 1 to 5) requested from the participant at intervals during the run, while the SWAT is more complex, involving ratings in three

Fig. 1 AMC custom 'full mission' maritime simulator main bridge, in Launceston, Australia

categories, 'Time Load', 'Mental Effort', and 'Psychological Stress', with ratings taken both during and after the run. For day one, ISA measures were taken at 5 min intervals, and the SWAT measure was taken once during the run at approximately 30 min, and once immediately after the run. For day two, ISA measures were taken at 3 min intervals. Due to poor feedback on the first day, the SWAT measure was not collected on day 2.

Results

The scores for Day 1, Low vs. High workload are graphed below. In each case, the ISA score is on the y-axis, and the time in minutes is on the x-axis. Low-workload runs are displayed on the left, and high workload on the right. (Fig. 2).

ISA analysis

As a measure of team workload we compared the day 1 mean team ISA scores for the low ($M = 2.76$, $SD = 0.53$) and high workload ($M = 2.62$, $SD = 0.59$) runs using a t-test. The scores did not differ significantly ($t = 0.37$, $df = 6$, $p = .72$). Similarly, team ISA scores at day 2 did not differ significantly ($t = -1.39$, $df = 12$, $p = .19$) between the low ($M = 2.18$, $SD = 0.31$) and high ($M = 2.47$, $SD = 0.46$) workload runs. Because the Day 2 results were non-significant, the graphs were excluded from this text, and are available in Appendix.

SWAT analysis

We analysed the SWAT scores using a 2 × 2 mixed model ANOVA with the time at which participants filled out the questionnaire (during run / after completion) as repeated measures factor and low / high workload as between subjects factor and mean team SWAT scores as dependent variable. While both main effects were not significant, the interaction between both factors was significant ($F_{(1, 5)} = 12.71$, $p = .02$). This interaction is visualised in Figs. 3 and 4. (Table 1).

Discussion

ISA measures

Measuring of workload ratings using this methodology was a challenge in the simulator, and it is likely that this difficulty would increase in real life operational settings. One

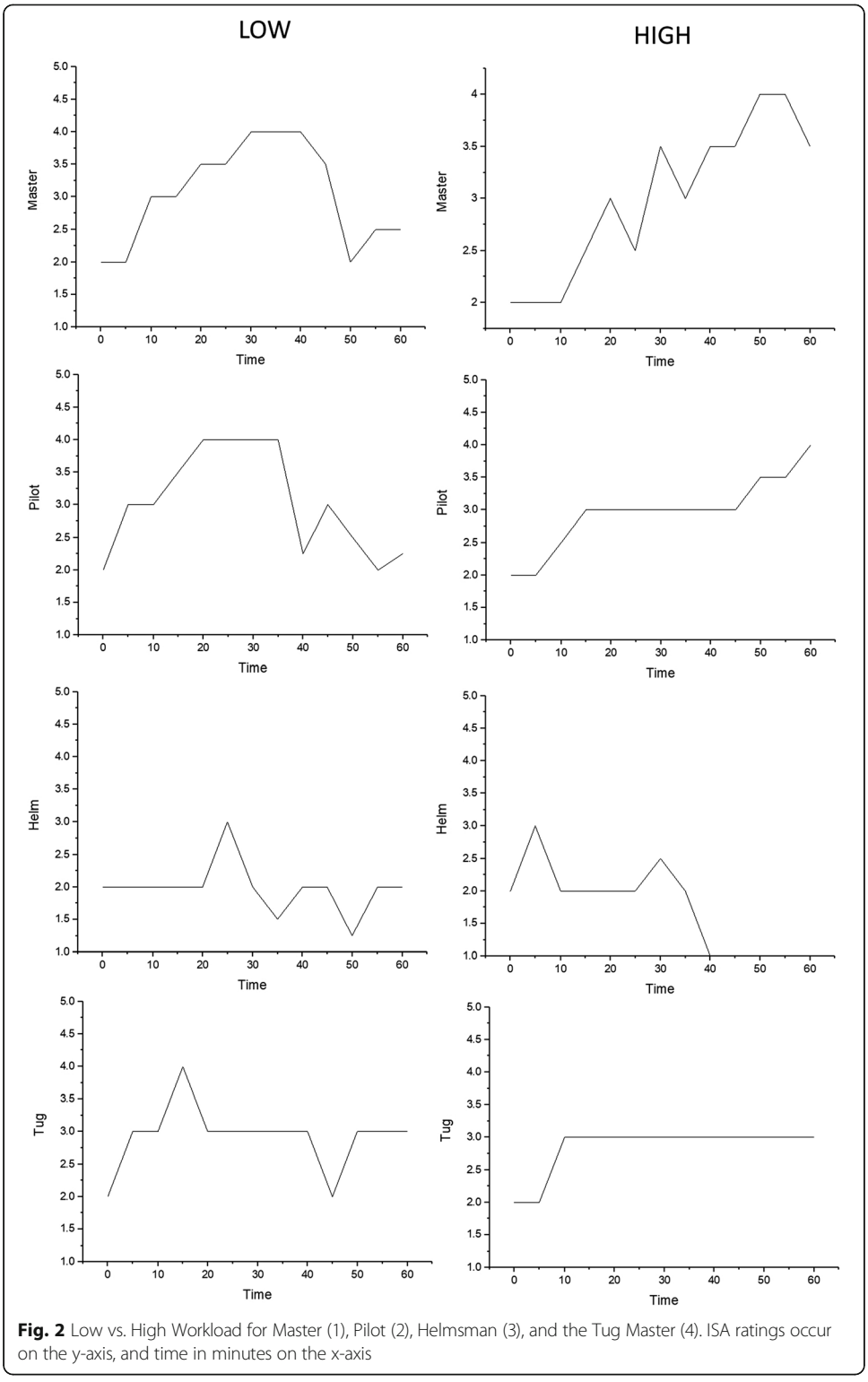


Fig. 2 Low vs. High Workload for Master (1), Pilot (2), Helmsman (3), and the Tug Master (4). ISA ratings occur on the y-axis, and time in minutes on the x-axis

issue is the time-stamping, or coordination, of data across experimenters at each station. Although the administration of the ISA demands rigid time intervals (e.g. 5 min) between measurements, these fluctuate slightly depending upon the availability of the participant to reply to the verbal prompt, and the ability of the experimenter to

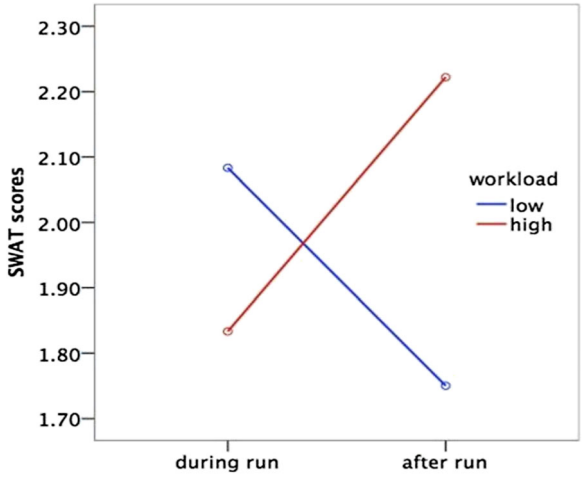


Fig. 3 SWAT mean team scores interaction between time (SWAT during or after simulation run) and low vs. high workload

administer timely cues. One possible remediation for this difficulty would be an automated logging system for workload ratings – though the user would still need to respond to the system at the appropriate time. One can likewise infer that the simple act of verbally prompting a participant for his / her workload may impact the operation currently underway.

Despite this issue, the ISA appears to have worked well, in that it shows coherent gradients across users during the simulation run, and that the pattern of performance differs in the 'high workload' and 'low workload' runs. Although the high and low workload conditions are statistically equivalent when comparing the average score over a run, a look at the graphs above, particularly the Captain and Pilot, indicate a very different pattern between the High and Low workload conditions. Specifically, in the Low load conditions, performance for the pilots and captains appears to peak during the middle of the run, whereas in the High load conditions the ratings continue to rise until the end of the run. This pattern is not evident for the helm and tug operators, indicating that perhaps they were not as affected by the workload manipulation.

There were indications that the 5 min intervals may be too long. Specifically, a brief, but nonetheless serious, incident can start and finish within a 5 min time period, and be missed completely on the rating scales. This was seen to happen at the 11:28 mark during the High workload condition on Day 1. In this case, the vessel crashed into the breakwater, and the Pilot voluntarily reported that he was momentarily at a workload of 5. This was missed by the standard ISA recording interval of 5 min. Reducing the interval, however, does not seem to be a viable solution, as verbal feedback from the participants on day 2 indicated that the 3-min ISA scoring interval was "annoying", because it was given too often.

The ISA scoring 1–5 was considered by one participant "straightforward and subconscious". The scale steps 1–5 were judged easy to remember but not fine-grained enough. Some ratings were between scale steps, which may indicate a preference for a more finely grained scale. Finally, another issue to consider is that the verbalization (speaking out loud) of individuals workload scores may have had an influence of other team members interpretation of the workload occurring at any given time.

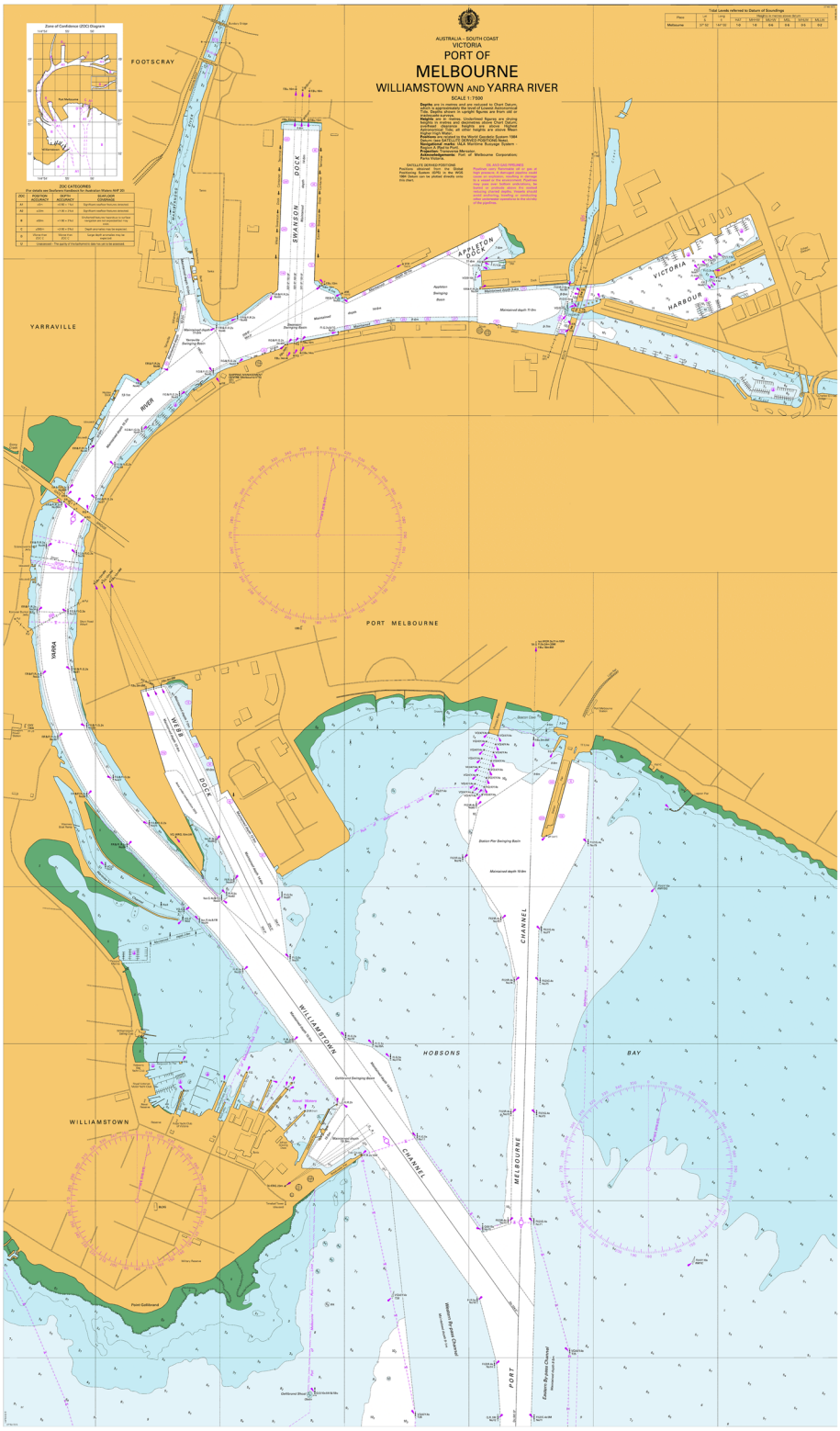


Fig. 4 Nautical Map of Port of Melbourne entry

Table 1 SWAT ratings for all runs. Measures were collected during each run at approximately 30 min in, and immediately after each run. * indicates missing scores

		During Run		After run	
Participant	Factor	Low load	High load	Low load	High load
Master	Time load	2	2	2	3
	Mental effort	2	2	2	2
	Psychological stress	2	2	2	2
Pilot	Time load	3	2	2	3
	Mental effort	2	2	2	3
	Psychological stress	2	2	2	2
Helmsman	Time load	2	1	1	*
	Mental effort	2	2	2	*
	Psychological stress	1	1	1	*
Tug	Time load	3	1.5	2	2
	Mental effort	2	2	2	2
	Psychological stress	2	1	1	1

SWAT measures

The SWAT measure was taken twice for each run: once during the run (at approximately 30 min into the run), and once after the run had completed. No visible relationships were evident, either to the ISA measures for the same run, or to operational environment of the bridge. Scoring on the SWAT was marginally lower in the high workload run, however this effect was not significant. This pattern is the opposite as that found in the ISA scores, which correctly identified the high workload runs. It must be considered, however, that the SWAT was not administered close to any high workload situations.

Overall, participants regarded the SWAT measure as too complex and wordy for quick judgements, i.e. during navigation. Further, it was perceived to interrupt the workflow to a greater extent than the ISA. One contradictory comment by a participant was that the SWAT measure works well – in particular that the 3-level scale was good, but that the description had too many words.

The statistical analysis showed an interesting interaction, which, with the amount of data collected in this study has to be interpreted rather cautiously. The results indicate that while during the simulation SWAT scores were slightly lower in the high workload condition and slightly higher in low workload condition, this was reversed after completion of the simulation where higher SWAT scores were given in the high workload condition and vice versa. From this it can be concluded that, after the run had completed, the SWAT was able to differentiate between high and low workload conditions, but that it was less successful when administered during the run.

Summary

In summary, though these measures were able to capture the workload of participants in the maritime simulator, it is evident that there are problems on a number of levels. Specifically, the measures themselves may impact workload; the measures were either too far apart to be meaningful, or so close in timing as to seriously disrupt the tasks being measured. The SWAT was unable to capture workload differences between the runs; and

finally, it was considered to be too wordy and complex for any rapid assessment of workload in an operational environment. Given this assessment, we undertook to develop more sophisticated methods of workload analysis for operational team environments.

Simulation #2: Using communications patterns to measure team workload

In order to further develop our model for team interactions and workload on a sea-going vessel's Bridge, between the Maritime Pilot and ship's Captain, as well as the Vessel Traffic Service (VTS) and auxiliaries such as local tug boats, we deployed a team of 5 researchers during a port operations training simulation being undertaken by a local Port Authority. Researchers logged the verbal, VHF, and local intercom communications from all parties during an 80 min simulation.

For this training operation, a fairway transit into the port of Melbourne, in the state of Victoria, Australia, was chosen. In addition to a team of 5 researchers from our group, there was a multi-disciplinary group from Melbourne Ports, including a Captain and an Officer of the Watch, one Helmsman; two VTS operators (one experienced VTS operator and one trainee); and a control-room based operator for the two tugboats. The purpose of the simulation was to model a standard entry into the port for a Panamax container ship, the Offen 4100. The vessel transit was aided by two tugs. An additional vessel, the Hual Trooper, was in transit directly behind the Offen 4100.

In order to assess how the team reacted to an unexpected serious event, a key emergency situation was timed to occur during a critical stage of the transit. In this case, the Offen 4100 experienced a main engine failure (bow thrusters remained operational) at approximately 40 min into the run. Incidentally, the timing of this failure was particularly serious, as the vessel was undergoing a turn in the fairway, and an oil refinery was visible on the shore at this point.

Method

Participants

A group of 5 personnel from the Melbourne Port participated in this simulation. These included a Captain, a Helmsman, an Officer of the Watch, and two VTS operators. Additionally, a staff member from UTAS/AMC operated the tugs from the control room. All participants were male, with an average age of 38 years, and an average experience of 17 years in current role.

Apparatus

This simulation also took place in the AMC custom maritime simulator in Launceston, Tasmania. For this simulation the main Bridge and Control Centre were employed, as well as one of the satellite operational rooms, which was equipped as a Vessel Traffic Service centre. The Bridge and VTS centre are equipped to feel and function like their real-world equivalents, while the Control Centre used to operate the tugs is designed to control all operational components of the simulation. See Fig. 1 above.

Design

For this simulation run, the following vessels were involved: Offen 4100, (Panamax container vessel, 281 m, top speed 20 knots); Hual Trooper (Car Carrier, 200 m, top speed 21

knots); Keera (34 m tug) and the Marysville (25 m tug). The scenario involved standard communications between the three main stations (Bridge, VTS, and Tug Master) via VHF radio, and additionally with the Hual Trooper on VHF radio. The VTS was located in a separate room, and manned by two VTS operators (one experienced, one trainee). The separate operations centres communicated by way of a real VHF radio system.

The goal of the simulation was to achieve safe docking of the Offen 4100 at the Port of Melbourne. This involved standard issues such as dealing with another vessel in transit (the Hual Trooper), tugs, and the VTS. Additionally, the team was tested during an emergency event, in which the main engines of the Offen 4100 failed at a critical juncture during the fairway transit, at approximately 40 min into the run. The reaction of the Bridge team and the VTS to such an event is essential for safe operations in a real-world environment, and thus the goal of the simulation was to provide experience in such a situation.

Results

Communications were logged at all stations during the simulation. For each instance of a communication was classified according to the mode, the sender, and the destination. For each participant, observers logged all significant verbal communications occurring throughout the duration of the run. Note, only communications related to the task at hand were logged; conversational or training related discussion was not logged. These data are tabulated in the following graphs.

Communications mode (Fig. 5)

Team communications

Discussion

It is evident from these data that in the current context, the Pilot is the main issuer of communications to all parties. This fits well with the typical embodiment of port operations, where the Pilot is given the task of navigating the ship to a safe berth. We can infer from this that the pilot is experiencing a relatively high workload. Interestingly, however, (and perhaps not common knowledge to those

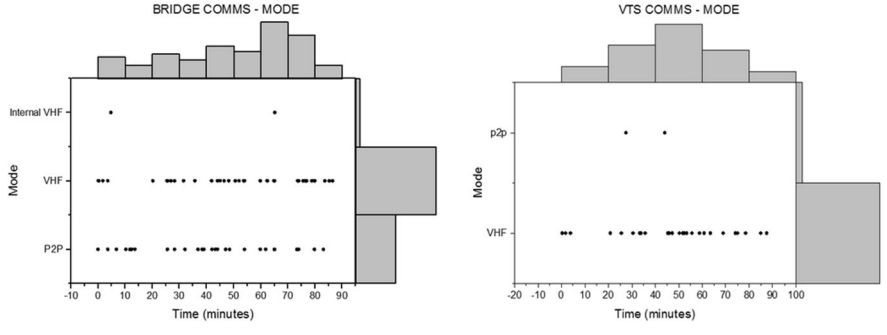


Fig. 5 Mode of communications at Bridge and VTS station. VHF stands for Very High Frequency, and P2P stands for Person to Person communication

outside of the maritime context), the Captain of the ship maintains legal responsibility for the ship through this process, despite his or her limited operational activity during port manoeuvres, as illustrated by the data in this study. In many cases, particularly in Oceania, the Captains will be of non-English origin, and communications between the Captain, tugs, and VTS can prove difficult without the Pilot as an intermediary due to language barriers. This relationship could be vastly enhanced by an explicit shared model of all operations surrounding the berthing of the ship in the local port of call.

In regards to the critical event, there is a definite increase in communications in the period following the engine failure. The progression of events can be followed by examining the "Comms - Destination" window of Fig. 6 above. When the engine fails at 40 min into the run, orders to the Helmsman cease. VTS traffic increases, and the Captain begins to receive increased communications around this time. Communications with the tugs increases, and can be seen to peak at their arrival on site, at approximately 60 min into the run. Further, it is evident that more communications were carried out between the Pilot and the tug Keera, as compared to the tug Marysvale. It is clear from this analysis that paying attention to communications patterns during maritime operations is one way to detect events of interest. In this case, the engine failed at a critical juncture (approaching a bend in the transit, with an oil refinery in the path of the vessel), and the communications pattern in the period immediately following the engine failure reflected this event.

An interesting discussion point is the issue of shared context between the bridge team and the tug boat(s) involved in port manoeuvres. While traditionally these dynamics are organized verbally over VHF, and maintained in working memory, there is evidence that as ship sizes increase (thus reducing visibility of the tugs), and as there is greater complexity in modern ports, this mental model of port operations could benefit from being actualized in some form of shared visual representation. This will be investigated in the third simulation, below.

In terms of workload analysis, the communications analysis is helpful, in that we can see who is doing most of talking, and therefore who is the most active. As described in the introduction, this is an example of a primary measure, where workload is inferred directly from the performance of the individual. However, this method is not entirely satisfactory, as we are only acquiring an analog of workload, rather than a direct understanding of the operators' workloads. For

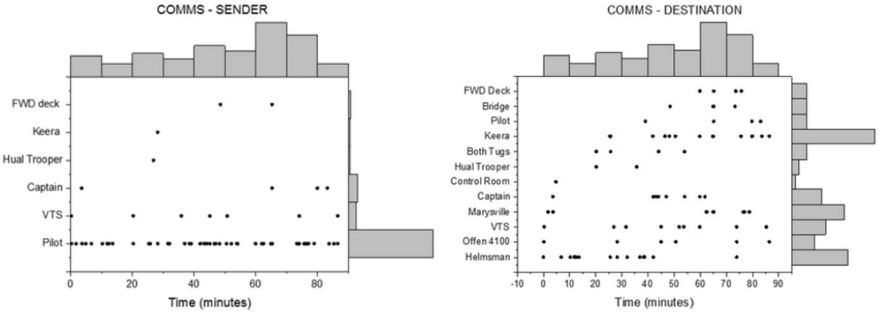


Fig. 6 Full list of communications within the maritime operations team, listed by Sender (left) and by Destination (right). FWD Deck (forward deck) controls the ship's anchor. Communications that were not specifically for the Captain or Pilot were labeled 'Bridge'. If the pilot had an order for both tugs, this was labelled 'Both Tugs', as opposed to individual tug orders (Keera, Marysvale)

example, different pilots could issue the same number of commands, despite the fact that one is considerably stressed by the combination of events taking place, while a different individual is relatively calm.

Simulation #3: Using GSR/EDA to measure team workload

In light of the preceding analysis, and faced with the question of operator performance and workload when considering the inclusion of new technologies in the maritime environment, our team decided to investigate synchronous electrophysiology (Electro-Dermal Activity) among members of the bridge-pilot-tug team. This methodology has the advantage that each participant's EDA signal can be monitored wirelessly via a remote transmitter / receiver apparatus, and the electrodes can be designed as to not interfere overly with standard operating procedures. As such, we were able to obtain synchronous electro-dermal responses from multiple team members, without negatively impacting the realism of the simulated ship-board procedures. In this manner, we collected simultaneous EDA signals from the Pilot, the Captain, and the Tug master, during runs in which a critical event occurred.

Prototype enhanced PPU device to facilitate team performance

This research was carried out in the context of testing a prototype maritime informatics device (study ongoing), which provided shared navigation information to the members of the distributed pilotage team, including Captains, Pilots, and Tug masters. The underlying notion behind this technology was that by increasing the shared conceptual space between team members, performance during standard and emergency procedures could be improved.

Method

Participants

Participants recruited for the study were all experts. Participants consisted of a distributed maritime operations team, including an experienced Captain (male, 31 years old, 5 years experience in current role), Pilot (male, 52 years old, 10 years experience in current role), Helmsman (male, 65 years old, unknown years experience), and Tug Master (male, 50 years old, 2 years experience in current role).

Apparatus

The same general testing apparatus was used for this simulation, with the addition of the GSR/EDA logger-sensor devices and the prototype PPU (shared context) device.

Kongsberg Full Mission Ship Simulator: Simulation runs were carried out AMC's 'full mission' ship simulator, which includes a full-scale ships bridge appropriate to a large ocean going vessel, as well as a number of secondary ship simulators, appropriate for tugs or smaller ships.

Logger-sensors: Using modular GSR/EDA logger-sensors (Neulog, 2014), equipped with RF transmitter modules and custom electrodes (two electrodes attached to a person's wrist with a velcro band to avoid issues with the traditional finger-mounted electrodes when operating ships controls),

Prototype PPU device: This research was carried out in the context of testing a prototype pilotage informatics device intended to improve the safety and efficiency of ship movements in ports.

Design and procedure

A single factor design (prototype device present vs. prototype device absent) was used for this study, with an emergency event occurring midway through each run. Participants either conducted the individual scenarios using standard maritime protocol (i.e. all comms carried out over UHF radio), or used the prototype PPU device to facilitate an understanding of relative position between tugs and vessel. Participants were provided information and consent forms and conducted a training and familiarization run, after which actual experiment commenced. Simulations were carried out over two days. The Pilot and Captain worked together on the ship's bridge (with a live Helmsman, who had no EDA monitor), while the Tug Master operated a separate simulated vessel from a remote control centre. Apart from the baseline simulation run in which no specific events occurred, the testing runs were designed to focus upon a particular hazard common to maritime environments.

Emergency events

In this analysis we focus on two emergency events: an engine failure (Series 1), and a loss of VHF communications (Series 2). In Series 1, the engine failures were timed to occur at the 8 min mark in each run. In Series 2, the communications failures were timed to occur at approximately 6 min and 8 min, in Run 1 and Run 2, respectively.

Results

Data collection

Using modular GSR/EDA logger-sensors, we collected simultaneous EDA signals from the Pilot, the Tug master, and the Ship's Master / Captain), during both standard training runs, and during runs in which a critical event occurred. Participants were given a period of 5 min for the EDA signal to settle prior to the onset of the trial.

Analysis

Raw values in microsiemens for each participant were converted to standardised (Z) scores to facilitate comparison between individuals (see Equation 1). Values were logged at a rate of 5 Hz for the duration of the simulation runs, which lasted between 30 and 60 min. For the comparative analysis, a modified version of the Z score was calculated (see Equation 2), replacing the group mean and group standard deviation with baseline terms, which were computed by taking the average and standard deviation of the 20 data cycles which occurred immediately prior to the onset of the trial. In this manner, it was possible to determine whether, on average, the GSR signal was higher or lower in each experimental condition.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

$$b = \frac{x - \mu(b)}{\sigma(b)} \tag{2}$$

Equations 1 and 2: Mean and standard deviation terms in a standardized score (z-score, top) are replaced with the mean of the baseline, μ(b), and standard

deviation of the baseline, σ(b). b scores therefore indicate EDA values standardized to the baseline, which consists of average EDA values during the 2 s immediately preceding the run.

Data filtering

Raw EDA signals, as captured from the device, are generally very 'noisy', in the sense that small movements of the participant can produce many small, and some large, transient errors in the data. Following the methodology of Bakker et al., (Bakker et al., 2011), we filtered the data using a median filter with a window size of 110. This means that the momentary EDA value is replaced by the median value that occurs within a 110 data frames (so, about 10 s on either side of the momentary signal). Because the function uses a 'rolling median' as time progresses, very little of the true underlying signal is lost: see Fig. 7. This has the result of smoothing the signal and removing all but the largest transient errors in the signal, while maintaining the actual EDA level.

Data Discretization

For further interpretability, we used a discretization filter on the data, which in essence rounded the standardized EDA value to the nearest whole number. In this manner, we were able to obtain a numeric rating for each participants EDA level at each point in the simulation. We calculated the percentages of these EDA levels for the teams for each run. Figure 8 shows a comparison of these percentages for the two runs (standard communication / new prototype) on the two day.

This comparison illustrates that on both days, the team showed a higher level of EDA activation (i.e. showed a higher percentage of high EDA values) in the standard communication runs compared to the runs that employed the new PPU prototype. (Figs. 9 and 10).

Comparison graphs: PPU device absent vs. PPU device present

Discussion

The current data set is encouraging, however, is indicative of some of the issues that can interfere when collecting physiological data in a realistic environment. Because the research participants were real bridge-team members carrying out realistic simulations,

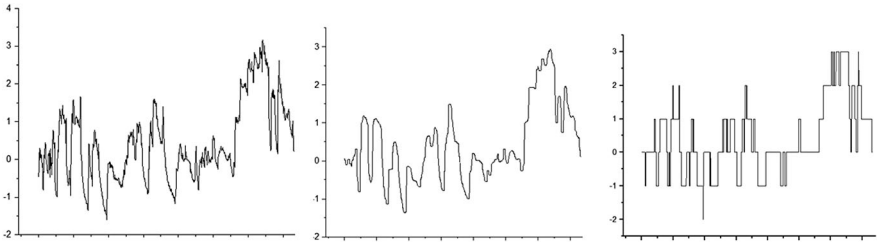


Fig. 7 Raw (left), Median Filtered (centre), and Discretized (right) EDA signals for one of the participants

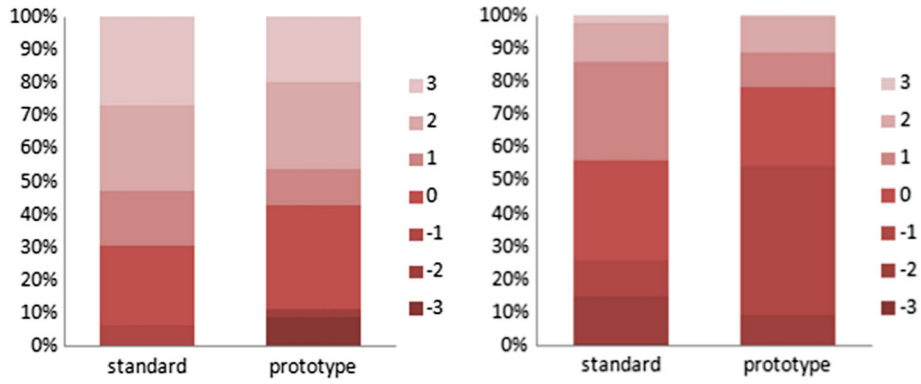


Fig. 8 Percentages of discretised EDA levels for the team; left: day 1; right day 2

it was often difficult to re-adjust or to check the EDA electrodes once the simulation was running. Likewise, there were situations when the electrical current for a particular user's electrodes were either under-saturated, showing a floor effect (no movement of the graph due to a bottomed-out signal), or over-saturated, showing a ceiling effect (no movement of the graph due to a maxed-out signal). The former could have been due to poor electrode contact, while the latter could be due to excessive skin conductivity as a result of body temperature (the Tug simulator room, in particular, was very warm). Outlier analysis and other filtering techniques enable us to substantially improve the interpretability of the data.

We can report that the EDA measurement technique is very promising for the measurement of team workload. In particular, it solves some of the key problems with subjective paper based measures such as SWAT and ISA, including the need to interrupt the task at hand in order to collect workload ratings. Such an act undoubtedly changes the task flow, thus at least partially invalidating the measure, and is unnecessary with the EDA apparatus.

Further, examination of the EDA signals in relation to the emergency events, in particular those in Session 2, indicate that this methodology can illustrate how team

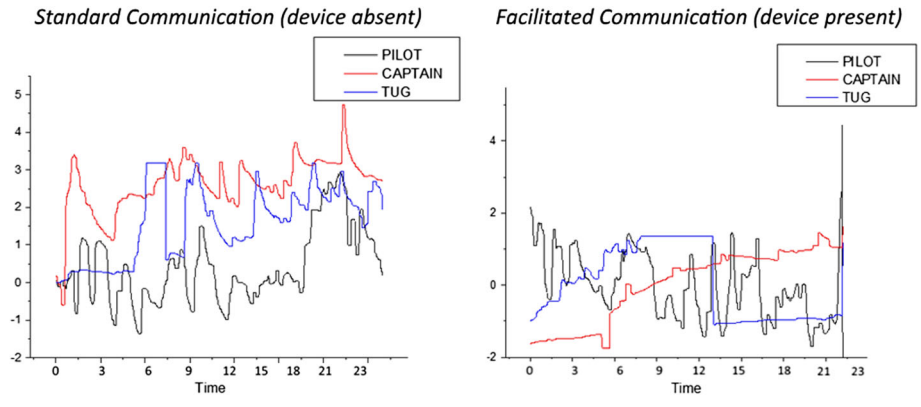


Fig. 9 Session 1: Run1 (Inbound to turning basin - engine fail) vs. Run 2 (Inbound to turning basin - engine fail - with prototype). * note, pilot EDA maxed out due to hot simulation room

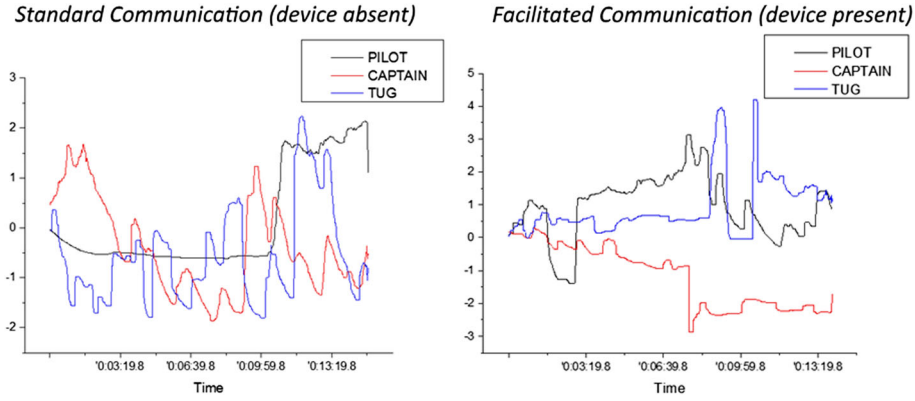


Fig. 10 Session 2: Run 1 (loss of communications - without prototype) vs. Run 2 (loss of communications – with prototype). Note EDA spikes for Tug at 6 and 9 min (right) indicate collisions with main vessel. Likewise the synchronous elevation of Tug and Pilot signals (left) indicate onset of communications failure

member's workload fluctuates together. Specifically, during the communications failure in Session 2 Run 1, the Tug collided with the main vessel at two points, which are clearly visible in the Tug's EDA measure at 6 and 9 min. Further, in Session 2 Run 2, at the onset of the communications failure around the 10 min mark, it can be clearly seen how the Tug and Pilot EDA levels increase sharply, while the Captains EDA signal remains at a lower level. This potentially indicates that the Pilot and Tug Master found the communications failure more taxing, while the Captain remained relatively calm.

The comparison of runs with standard communication and the new PPU prototype revealed that the standard runs show higher EDA response in team members. This new prototype in development is designed to improve collaboration between team members and thus may help in reducing workload. Our results indicate that the EDA measurement in fact was sensitive enough to measure reduced activation, as a proxy of workload, in simulation runs that employed the new prototype.

Generally we have shown that electrophysiological measure responds sharply in the context of stressful events in the simulation. These events are readily visible in the EDA signal, and this can aid in a detailed analysis of the events during the run, and their effect on the team members. Finally, it is evident from these brief analyses that the EDA signal, when collected for a team of individuals, has great potential for showing the *synchrony* with which individual team members respond to a stressful event. In this regard, these measures may be very well suited to understanding and developing team mental models of workload during stressful events in the maritime operating environment.

Strengths and weaknesses of GSR/EDA measures

Our analysis of these data have enabled us to compile a list of some of the positive and negative aspects, related to the use of EDA to measure stress levels in a simulated maritime team operating environment. (Table 2).

Note that the weaknesses are mostly implementation difficulties, which can be overcome with better equipment, some and signal analytics, and/or more experience with taking such measurements. While the advantages inherent in this methodology are enough to encourage further research along these lines, it is furthermore evident that proper design of the equipment and simulation facility (including proper cooling of the simulator rooms), coupled with an improved understanding of analysis techniques and data processing capability, could allow this methodology to become a versatile and easy-to-administer means of tracking team workload levels.

General discussion and conclusion

In this paper we investigated a variety of methods to study cognitive workload in a simulated maritime operations environment. The simulated ship, Tug and VTS operating environments allowed us to have a high level of operational control, while conducting the study in a safe and repeatable manner. At the same time the simulation technology is maximally akin to real world settings and procedures, and therefore may provide some valuable insights that would not be obtainable in a classical laboratory setting. While high mental workload is generally implicit in critical task environments, particularly during emergency manoeuvres, it is worth nothing here that some steps can be taken to alleviate workload in routine but critical environments. Particularly, automation of routine tasks can leave the human operator with spare capacity for essential decision making; however there is a further risk here of automation-related errors compounding such emergency situations (Parasuraman & Riley, 1997).

We have studied standard workload questionnaires, analysis of communication patterns, and electro-dermal activity as measures of operator workload. Our findings indicate that each of these methods have unique advantages and disadvantages and, depending on circumstances, may be more or less suitable to measure workload. The ISA and SWAT questionnaires were generally able to measure workload levels. These scales are commonly used, which is beneficial in that it facilitates easy comparison with previous work in the domain. There also an established understanding in the research community regarding how they may or may not be used, and further, information regarding the tool's reliability and validity metrics is well known. There are, however, several shortcomings of questionnaire use, including problems with recall, rationalisation, and, as discussed earlier, interruption of workflow.

Communications are shown here to be a valuable tool that can be used to analyse team dynamics, and develop an understanding of how information flows between operator stations. In this research, we have demonstrated a clear pattern of communications surrounding an emergency event (engine failure), and have identified the Pilot and

Table 2 Strengths and weaknesses of using EDA to measure workload in maritime operations

Strengths of EDA measures	Weaknesses of EDA measures
immediate response to stress events immediate ability to monitor performance no interruption of task (ecological validity) good ability to demonstrate correlation of activation between team members 'behavioural synchronicity', or synchrony	inconsistency of GSR signal susceptibility of GSR signal to heat, other disruptions

Tugs as the parties most active during the emergency. It should be noted here, however, that the collection of communications data were relatively effort-intensive, and that standard methods of analysis have yet to be developed. The methodology used in the present Simulation 2 required five researchers for data collection, and considerable time transcribing verbal dialog; the recording phase was especially tricky given the speed at which communications were carried out during the emergency manoeuvres. An analysis of the recorded audio could facilitate this process, but would have its own drawbacks, accordingly.

The measurement of Electro-Dermal Activity has shown considerable promise in studying workload, as a correlate of physiological stress, in real time. A major advantage when using EDA is that the operator's workflow does not have to be interrupted, and further, the operator's level of EDA response can be measured continuously.

This last point bears particularly on the usefulness of this measure to interpret the workload of a team as a whole. Because the relative increases and decreases in workload can be seen in real-time, a monitoring device, or personnel, can have instant feedback regarding the operating capacity within a given team. In this research, we saw clearly that team workload levels tended to fluctuate simultaneously for key team members when the emergency event occurred.

Although standard metrics such as the SWAT and ISA are able to distinguish high and low workload conditions, they may be unsuitable to real world environments because they break up the task flow, and could even introduce some risk into the operating environment. New methodology such as communications analysis and EDA measurement may bridge the gap, allowing for real time measurement of workload without disrupting the task flow. Ideally, a combination of the measures employed in this series of studies would be the best way to measure operator workload, and specifically team workload in a maritime operations setting. Given this context, we find considerable reason to further investigate the use of remote electrophysiology to assess operator workload in complex, real world team environments.

Regarding implications of this research for shipping and trade, there are a few interesting points to consider. Primarily, we have shown that there are multiple ways of assessing the impact of introducing new technologies to the ship's bridge. Given the proliferation of new technologies on the horizon, including automation and supplementary information system, there may be some advantage in applying these methodologies to test the impact of a new device, prior to its commercial implementation. Further, we have demonstrated the ability to simultaneously monitor the workload of both bridge personnel, and remote personnel such as tug or VTS operators, without impacting operator performance, which is necessary if we wish to monitor workload in a critical task environment. This could provide an opportunity for testing the impact of such experimental technologies as remote pilotage (Hadley, 1999), or shared mental model systems (e.g. Owen et al., 2013) for improving efficiency and safety in port entry.

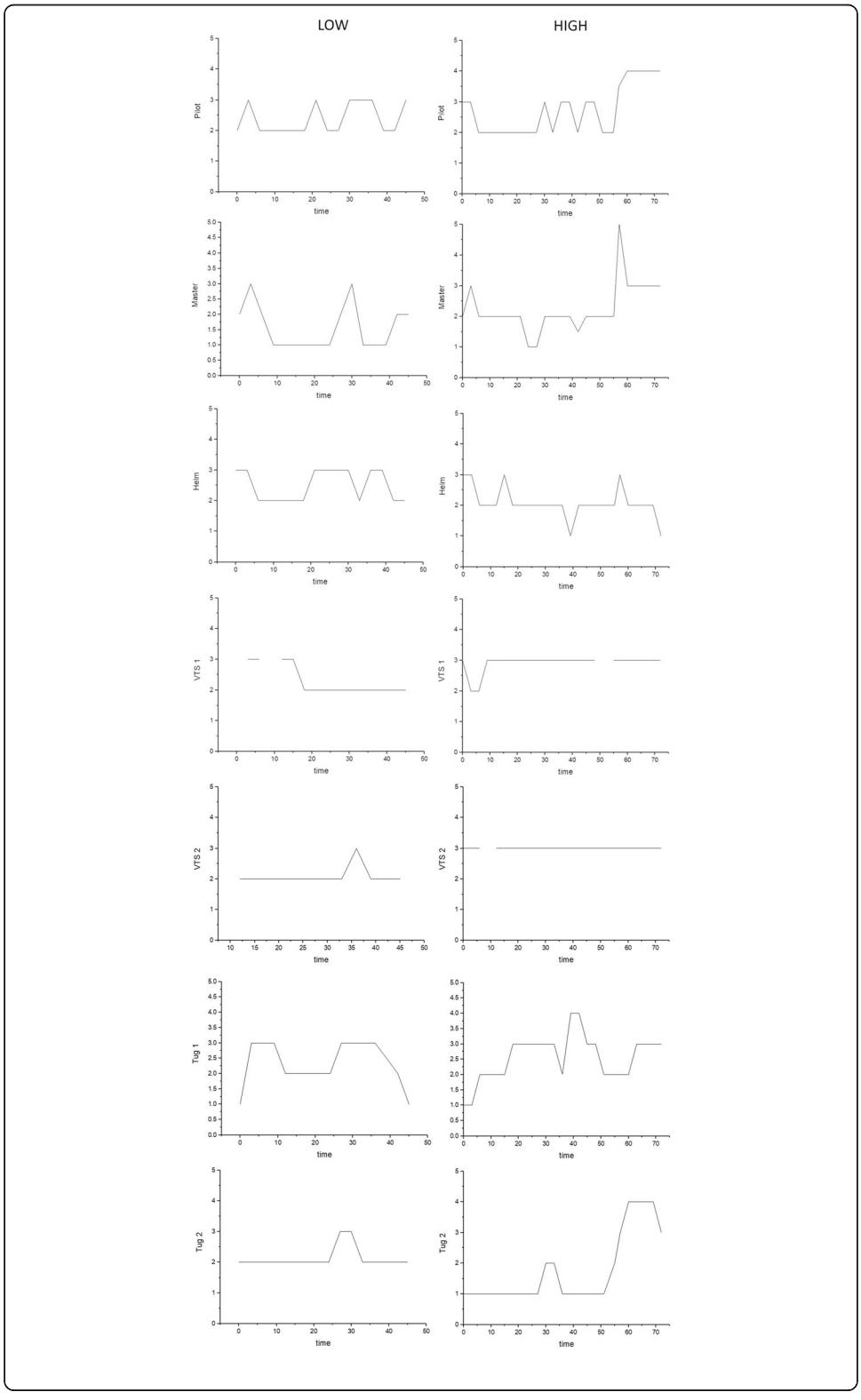
Endnotes

[1]e.g. Instantaneous Situation Awareness or ISA measure (EUROPEAN ORGANIZATION FOR THE SAFETY OF AIR NAVIGATION, 1996)

[2]e.g. NASA TLX (Hart & Staveland, 1988)

Appendix

Simulation 1, Day 2 ISA scores for High vs. Low Workload runs, ISA collected at 3 min intervals.



Acknowledgements
We have no further acknowledgements to make.

Funding
This project was funded by grant #00003477 from Ports Australia and Associate Professor Brooks' involvement was also funded by grant LP120100422 from the Australian Research Council.

Authors' contributions
All authors participated in background research, study design, experimental testing, and writing of this paper.
All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Commonwealth Scientific and Industrial Research Organisation, 15 College Road, Sandy Bay 7005, Australia. [2]Australian Maritime College, Locked bag 1397, Launceston, Tasmania 7250, Australia. [3]Otago Polytechnic, Forth St, Dunedin 9054, New Zealand.

Received: 4 October 2017 Accepted: 25 January 2018
Published online: 23 February 2018

References
Baddeley A (1992) Working memory. Science 255(5044):556–559
Bakker J, Pechenizkiy M, Sidorova N (2011) What's your current stress level? Detection of stress patterns from GSR sensor data. In: Data mining workshops (ICDMW), 2011 IEEE 11th international conference on [internet]. IEEE, pp 573–580 [cited 2015 Jun 30]
Cain B (2007) A review of the mental workload literature. Defence Research And Development Toronto (Canada), Toronto
De Zwart BC, Frings-Dresen MH, Van Dijk FJ (1996) Physical workload and the ageing worker: a review of the literature. Int Arch Occup Environ Health 68(1):1–12
European organization for the safety of air navigation (1996). Ergo (Version 2) for instantaneous self assessment of workload in a real0time ATC simulation environment. [cited 2015 Sep 6]
Funke GJ, Knott BA, Salas E, Pavlas D, Strang AJ (2012) Conceptualization and measurement of team workload: a critical need. Hum Factors J Hum Factors Ergon Soc. 54(1):36–51
Grabowski M, Sanborn SD (2001) Evaluation of embedded intelligent real-time systems*. Decis Sci 32(1):95–124
Hadley M (1999) Issues in remote pilotage. J Navig 52(1):1–10
Hart SG, Staveland LE (1988) Development of the NASA-TLX (task load index): results of empirical and theoretical research. In: Human mental workload. North Holland Press, Amsterdam [cited 2015 Jan 9]
Hetherington C, Flin R, Mearns K (2006) Safety in shipping: the human element. J Saf Res 37(4):401–411
Loft S, Sanderson P, Neal A, Mooij M (2007) Modeling and predicting mental workload in en route air traffic control: critical review and broader implications. Hum Factors 49(3):376–399
Lützhöft M (2004) "The technology is great when it works": maritime technology and human integration on the Ship's bridge. . (ph.D. thesis). Linköping University, Linköping Available at: http://www.diva-portal.org
Lützhöft M, Grech MR, Porathe T (2011) Information environment, fatigue, and culture in the maritime domain. Rev Hum Factors Ergo 7(1):280–322
Lützhöft M, Nyce J (2014) Integration work on the ship's bridge. J Marit Res 5(2):59–74
Marshall SP (2002) The index of cognitive activity: measuring cognitive workload. In: Human factors and power plants, 2002 proceedings of the 2002 IEEE 7th conference on. IEEE, pp 7–5 [cited 2015 Sep 30]
Owen C, Bearman C, Brooks B, Chapman J, Paton D, Hossain L (2013) Developing a research framework for complex multi–team coordination in emergency management. Int J Emerg Manag 9(1):1–7
Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. Hum Factors 39(2):230–253
Sheridan TB (1981) Understanding human error and aiding human diagnostic behaviour in nuclear power plants. In: Human detection and diagnosis of system failures. Springer, US, pp 19–35
Staal MA (2004) Stress, cognition, and human performance: a literature review and conceptual framework. NaSA Tech Memo 212824:9
Sweller J, Ayres P, Kalyuga S (2011) Cognitive load theory. Springer [cited 2013 Nov 15].
Trick LM, Lochner M, Toxopeus R, Wilson D (2009) Manipulating drive characteristics to study the effects of mental load on older and younger drivers. In: Proceedings of the fifth international driving symposium on human factors in driving assessment, training, and vehicle design, big sky, MT, pp 363–369
Tsang P, Vidulich M (2006) Mental workload and situation awareness. In: Handbook of human factors and ergonomics, third edition. Wiley, Hoboken, pp 243–268. doi:https://doi.org/10.1002/0470048204.ch60
Wickens CD (2008) Multiple resources and mental workload. Hum Factors J Hum Factors Ergon Soc 50(3):449–455