THE UNIVERSITY OF
WESTERN AUSTRALIA

*Achieve International Excellence*

# Situation Awareness Measurement Techniques for Submarine Track Management

## Final Report Prepared for the Asian Office of Aerospace Research Development (funding agreement #AOARD-12-4028)

Report Prepared by Vanessa Bowden and Shayne Loft

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*

| | | |
|---|---|---|
| 1. REPORT DATE<br>**22 FEB 2013** | 2. REPORT TYPE<br>**Final** | 3. DATES COVERED<br>**05-03-2012 to 04-12-2012** |

4. TITLE AND SUBTITLE
**Situation awareness measurement techniques for submarine track management**

5a. CONTRACT NUMBER
**FA23861214028**

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
**Shayne Loft**

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
**University of Western Australia,35 Stirling Highway,Crawley 6009,Australia,AU,6009**

8. PERFORMING ORGANIZATION REPORT NUMBER
**N/A**

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
**AOARD, UNIT 45002, APO, AP, 96338-5002**

10. SPONSOR/MONITOR’S ACRONYM(S)
**AOARD**

11. SPONSOR/MONITOR’S REPORT NUMBER(S)
**AOARD-124028**

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT
**The first objective was to complete a literature review of the benefits and risks that task automation poses for SA and performance for tasks where individuals monitor and control complex dynamic displays (report delivered to AOARD in April 2012). The second objective was to further establish the predictive validity of SPAM in a higher fidelity simulated submarine track management task by using SPAM to predict task performance. We compared the predictive validity of SPAM to the predictive validity of the Situation Awareness Global Assessment Technique (SAGAT). The third objective was to demonstrate the sensitivity of SPAM and SAGAT to manipulations of task automation, and to examine the impact of task automation on performance. The fourth objective was to examine whether SPAM or SAGAT administration were disruptive by assessing their impact on subjective workload and performance.**

15. SUBJECT TERMS
**Psychology, Human Attention & Performance**

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT<br>**unclassified** | b. ABSTRACT<br>**unclassified** | c. THIS PAGE<br>**unclassified** | **Same as Report (SAR)** | **31** | |

## Executive Summary

***The UWA Project Team, in alphabetical order, consists of:***

- Dr Vanessa Bowden (UWA – School of Psychology)
- Janelle Braithwaite (UWA – Oceans Institute)
- Stephanie Chen (UWA – School of Psychology)
- Associate Professor Shayne Loft (UWA – School of Psychology)
- Daniel Morrell (UWA – School of Psychology)
- Daphne Tan (UWA – School of Psychology)

***Background.*** The Situation Present Assessment Method (SPAM) has been identified as a potential tool to measure Situation Awareness (SA) in the submarine environment as part of its research into human and system performance. The current research builds on previous work (Loft & Morrell, 2011) that found SPAM could predict performance in simulated submarine track management and was sensitive to changes in task demands, information uncertainty, and task disruptions. The current research further assesses the utility of SPAM for measuring SA and examines the impact of automation in a higher fidelity simulated submarine track management task.

***Objectives.*** The first objective was to complete a literature review of the benefits and risks that task automation poses for SA and performance for tasks where individuals monitor and control complex dynamic displays (report delivered to AOARD in April 2012). The second objective was to further establish the predictive validity of SPAM in a higher fidelity simulated submarine track management task by using SPAM to predict task performance. We compared the predictive validity of SPAM to the predictive validity of the Situation Awareness Global Assessment Technique (SAGAT). The third objective was to demonstrate the sensitivity of SPAM and SAGAT to manipulations of task automation, and to examine the impact of task automation on performance. The fourth objective was to examine whether SPAM or SAGAT administration were disruptive by assessing their impact on subjective workload and performance.

***Method.*** One-hundred and seventy seven participants completed a simulated submarine track management task. Participants monitored a conventional radar (picture) display and an adjacent waterfall display. Following training on Day 1, half the participants were allocated to a task automation condition and the other half completed Day 2 without task automation. On Day 2, participants completed one test scenario where their SA was assessed using SPAM, one scenario where SA was assessed using SAGAT, and one control scenario where neither SPAM nor SAGAT were administered. After each of the three scenarios, participants also completed the Situation Awareness Rating Technique (SART), a subjective measure of SA. Performance was measured on a closest point of approach (CPA) task, a contact (vessel) classification task, and an emergency surface task. Participants classified the contacts represented on the picture display as 'Friendly', 'Enemy', or 'Merchant' depending on the contacts behaviour. Participants were also required to mark the CPA of other contacts to Ownship. The emergency surface task required information to be integrated from both the classification task and the CPA task. Participants in the task automation condition had the closest point of approach task automated for some contacts. Subjective workload was measured at various points during each trial using the Air Traffic Workload Input Technique (ATWIT), and post-trial using the NASA Task Load Index (NASA-TLX).

***Results.*** *Predictive Validity of SPAM and SAGAT:* SPAM significantly predicted unique variance in all three performance tasks (closest point of approach, contact classification, and emergency surface task) over and above the variance predicted by the subjective SA and subjective workload measures. SPAM accuracy was a more reliable predictor than SPAM response time. SPAM and SAGAT measures were significantly correlated, although the amount of shared variance was minimal. *Sensitivity of SPAM and SAGAT to Automation*: The SPAM measure was not sensitive to automation (neither was SAGAT). However, the automation manipulation also failed to impact on subjective workload or on most performance measures, which suggests that the automation may not have been effective. *Effect of Automation on Performance:* The automation of the closest point of approach task for some contacts (vessels) actually caused poorer closest point of approach decisions for the other vessels *not* automated, compared to the condition not provided automation at all. *Automation Usage*: Although participants were told that the automation was 100% reliable, only 40% of participants in the automation condition accepted every automated closet point of approach decision. *Disruptive Effects of SA Measurement:* The administration of SPAM did not significantly increase subjective workload (as measured by the ATWITT and NASA-TLX) or task performance compared to the control condition with no SA measurement. However, SAGAT administration significantly increased both within-scenario and post-scenario rated subjective workload compared to the control condition.

***Conclusions.*** Both SPAM and SAGAT had significant predictive validity in a simulated submarine track management task. It was important that SPAM and SAGAT predicted task performance variance over and above the subjective SA measure (SART) because there is no doubt that SART is the easiest measure to prepare and administer. However, the administration of SAGAT significantly increased subjective workload. Thus, predictive validity provided by SAGAT compared to SPAM came at the cost of increased participant subjective workload rating. Since SAGAT disrupts the operator from the tasks at hand, this may limit its use in the submarine control room. In contrast, SPAM administration does not require the task operation to be paused, and also better simulates the command team interaction typical of operational settings. The finding that task automation impaired other non-automated performance task elements is consistent with the findings of our literature review that automation can disengage the operator from the task and impact decision making. The fact that not all participants used the automation, despite it being 100% reliable, is indicative that some participants may not have trusted the automated decisions. In naval operations, it may be more important for SA that track managers work to resolve certain elements of uncertainty rather than passively reacting to automation system output. Care must be taken generalising the findings to real naval operations given the laboratory task and the student participant sample. This notwithstanding, we have taken a crucial further step towards demonstrating that SPAM can potentially be used to assess operator SA in naval submarine track management.

# 1. Background and Objectives

The sensor and information systems employed on submarines are designed to assist human decision making. Understanding the effectiveness with which such technology supports submarine decision making is difficult. The challenge partly stems from the difficulty in definition and measurement of the 'situation awareness' (SA) that is required for command to make decisions. SA is often assumed to be provided by display technology but it really only exists in the heads of decision makers at any level of the submarine command team. SA has been shown to provide the foundation for safe and efficient performance in work systems as diverse as combat aviation (Vidulich, McCoy, & Crabtree, 1995), anaesthesiology (Gaba, Howard, & Small, 1995), and air traffic control (Durso, Hackworth, Truitt, Crutchfield, & Nikolic, 1999). While a handful of papers have reported broad task analysis of submarine personnel (Ehret, Gray, & Kirschenbaum, 1997, 2000; Kirschenbaum, 2011), there has only been one study to date that has defined and measured SA in simulated submarine track management (see DSTO-UWA Research Agreement number 352520; Loft & Morrell, 2011). The next phase of the research (funded by both the Australian Defence Science and Technology Organisation - DSTO, and AOARD) that is reported in the current document continues to examine SA measurement and the impact of task automation in simulated submarine track management.

There are a variety of subjective methods used to measure SA, including asking operators to rate their own SA, or having subject matter experts evaluate the degree to which other operators exhibit SA. However, objective query-based techniques provide more accurate SA measurement. The Situational Awareness Global Assessment Technique (SAGAT) is the most widely used and validated query-based SA measure (Salmon, Stanton, Walker, & Green, 2006). SAGAT focuses on the product of SA using recall techniques, uncovering information of which the operator is consciously aware. SAGAT employs timed freezes in a task, during which elements in the display are blocked, and operators recall what was happening at the time of the freeze. However, the fact that SAGAT disrupts the operator from the tasks at hand may limit its use in the submarine control room. This is because much of the cognitive work of the control room team involves the management of uncertainty stemming from the fact the submarine's passive sensors do not yield veridical range information. Temporarily removing access to the display would lead to the mental picture of the situation being quickly lost. Hence, assessing SA in real-time appears the most advantageous in a practical sense for submarine track management.

The Situation Present Assessment Method (SPAM; Durso & Dattel, 2004) is potentially suited to the submarine environment because it measures SA in real-time without pausing task operations. SPAM distinguishes workload from SA by warning the operator that a question is in the queue, and waiting until the operator accepts the question. SPAM probe acceptance time is measured as the time between when the experimenter asks the participant whether they were 'ready' to the time that the participant accepts the question. Following this, the SA question is asked and SPAM response time is measured as the time between when the experimenter completes asking the question to the time the participant answers (Durso & Dattel, 2004). The logic underlying SPAM is that operators who have better SA will know where to find appropriate information and thus be able to respond faster or more accurately. It is arguable that SPAM simulates the command team interactions typical of submarine operations (i.e., live unsolicited requests for information from the Officer on the Watch to the track manager). SPAM has been effectively used to measure SA and to predict performance in work domains of air traffic control and driving (Durso, Bleckley, & Dattel, 2006; Durso & Dattel, 2004; Durso et al., 1999).

Loft and Morrell (2011; DSTO-UWA Research Agreement number 352520) conducted the first set of experiments using SPAM in simulated submarine track management. The first objective was to determine if SPAM was a sensitive measure of SA and if it predicted performance; under conditions of varying uncertainty, workload, and task disruption. Loft and Morrell (2011) developed a simulation of submarine track management in consultation with the DSTO called SIM(sub). In three laboratory experiments, participants performed SIM(sub) tasks such as deciding whether they could fire on enemy contacts using rules of engagement and detecting contact heading changes. Participants also responded to SPAM probes designed to measure their SA of the current and future state of the display. In a fourth experiment SPAM was applied within a small team setting under more realistic field conditions (using the T106 combat system) and with trained Royal Australian Navy (RAN) submariners.

In Experiments 1-3 participants were less accurate and slower to respond to SPAM questions that assessed SA of the future display situation compared to SA of the current display situation. This finding was replicated across manipulations of uncertainty, workload, and task disruption, and was also replicated in Experiment 4 using the T106 combat system with trained RAN operators. Thus, it was difficult for participants to project future states of the simulated tack management display, even though the dynamic elements on the track management display were evolving quite slowly.

Loft and Morrell (2011) also found that SPAM differentially predicted performance. In Experiments 1, 2, and 3 the time that participants took to decide if they could fire on enemy contacts was dependent on their ability to access and interpret information related to the current display; that is, this task was most uniquely predicted by current SPAM questions. In contrast, performance tasks that required participants to prioritise when and where to allocate attention were predicted by future SPAM questions, which reflected participants' SA of how the relationships between contacts on the display would evolve. However, in Experiment 2 the relationship between current SA and performance was moderated by the contact uncertainty manipulation. In this case, SA for the current display was more likely to predict performance under conditions of low uncertainty, and SA for the future display was more likely to predict performance under conditions of high uncertainty. In Experiment 3, task disruptions reduced participants' SA and performance. The number of vessels on the display (the workload manipulation) negatively impacted on performance but had no influence on SA.

The experiments by Loft and Morrell (2011) suggest that SPAM can potentially be a sensitive measure of SA and could be used to predict performance in submarine track management. It was also established that the SPAM technique was sensitive to task demands such as workload impacts, information uncertainty, and task disruptions of the type that would be expected at sea. However, there are many research questions pertaining to the potential use of SPAM in submarine track management left unanswered by Loft and Morrell (2011). First, it is crucial to demonstrate the predictive validity of the SPAM measurement method in a higher fidelity simulation of submarine track management and in which participants receive more task training. Second, it is important to examine the extent to which SPAM can predict task performance when compared to the predictive validity of the more well-established SAGAT measure. Third, it is possible that SPAM or SAGAT probes may be distracting, providing additional loading that would take away from primary task performance. To examine this, we need to include additional control conditions where SPAM or SAGAT questions are not administered during the simulated submarine track management task. In addition, it is important to examine the impact of task automation on SA and task performance.

The first objective of the current project was to complete a literature review of the benefits and risks that task automation poses for SA and performance for tasks where individuals monitor and control complex dynamic displays (this report was delivered in April 2012). The second objective was to further establish the predictive validity of SPAM in a higher fidelity simulated submarine track management task. We compared the predictive validity of SPAM to the predictive validity of SAGAT. The third objective was to demonstrate the sensitivity of SPAM and SAGAT to manipulations of task automation, and to examine the impact of task automation on performance. The fourth objective was to examine whether SPAM or SAGAT administration are disruptive by assessing their impact on subjective workload and performance.

# 2. Method

## 2.1. Participants

179 participants from the undergraduate population at UWA completed two days of testing. There were 57 males and 122 females with a mean age of 21.3 years.

## 2.2. The SIM(sub)3.0 Program

SIM(sub)3.0 was developed and runs in the MATLAB programming environment. Participants observed and interacted with two display screens (two 22 inch monitors arranged side-by-side). The left monitor presented a simplified but conventional maritime radar (tactical picture) display and the right monitor presented a submarine sonar bearing history or ‘waterfall’ display (see Figure 1). Participants were required to attend to both displays in order to perform several different tasks; a contact classification task, a closest point of approach task, and an emergency surface task. Each of these tasks is described in greater detail in the sections below.

### 2.2.1. Contact Classification task

Participants classified the contacts (vessels) presented on picture display as ‘Friendly’, ‘Enemy’ or ‘Merchant’. Accurate classification of contacts required the participants to make rule based judgements using the information presented on both their picture and waterfall displays. To complete a classification, participants used the mouse to select a contact and then entered a classification option using labelled keys (‘F’ for friendly, ‘E’ for enemy, and ‘M’ for merchant). Once a contact was classified, the icon on the radar display changed from a yellow circle to the icon designating the new classification (green triangle for friendly, red square for enemy, and inverted white triangle for merchant) and the associated track on the waterfall display changed from a yellow line to the classified icon colour (see Figure 2).

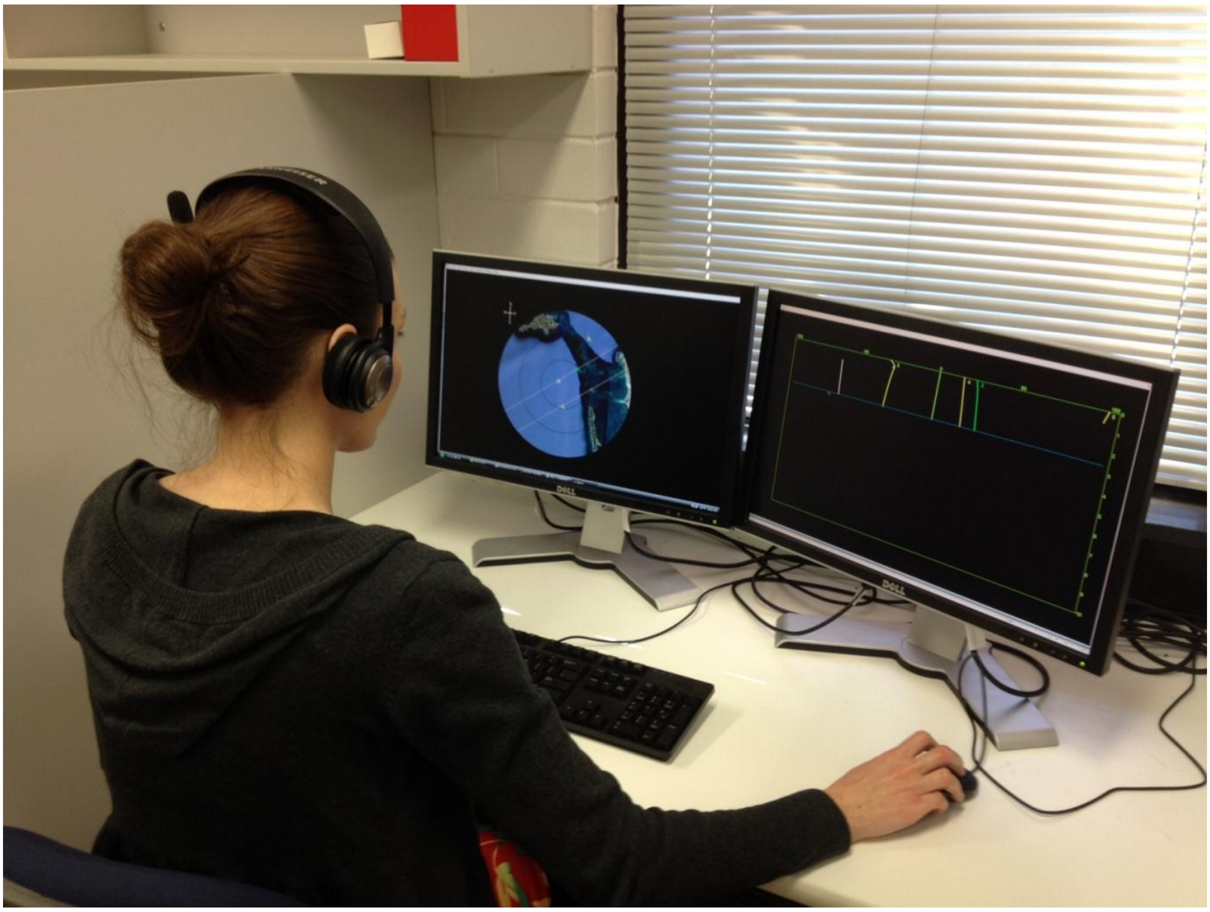

Figure 1. Photograph showing the experimental setup. The left monitor shows the picture display, and the right monitor shows the waterfall display. Participants wore headphones throughout the experiment and responded with a combination of mouse and keyboard entries.

Participants were able to determine which waterfall track belonged to which contact by checking the contact number associated with the track (small number to the right of the start of each track on the waterfall display). Participants were able to re-classify contacts as many times as they wanted while that contact remained present on the picture display.

Contacts were classified according to the following rules:

1. **Friendly** – To be classified as a friendly, the contact must have remained within 5 km of Ownship for at least four minutes.
2. **Enemy** – To be classified as an enemy, the contact must have zigged three times within a six minute period (where a zig is an angular change in direction of at least 45 degrees, performed over less than 30 seconds).
3. **Merchant** – To be classified as a merchant, the contact must have remained within a shipping lane for at least four minutes.

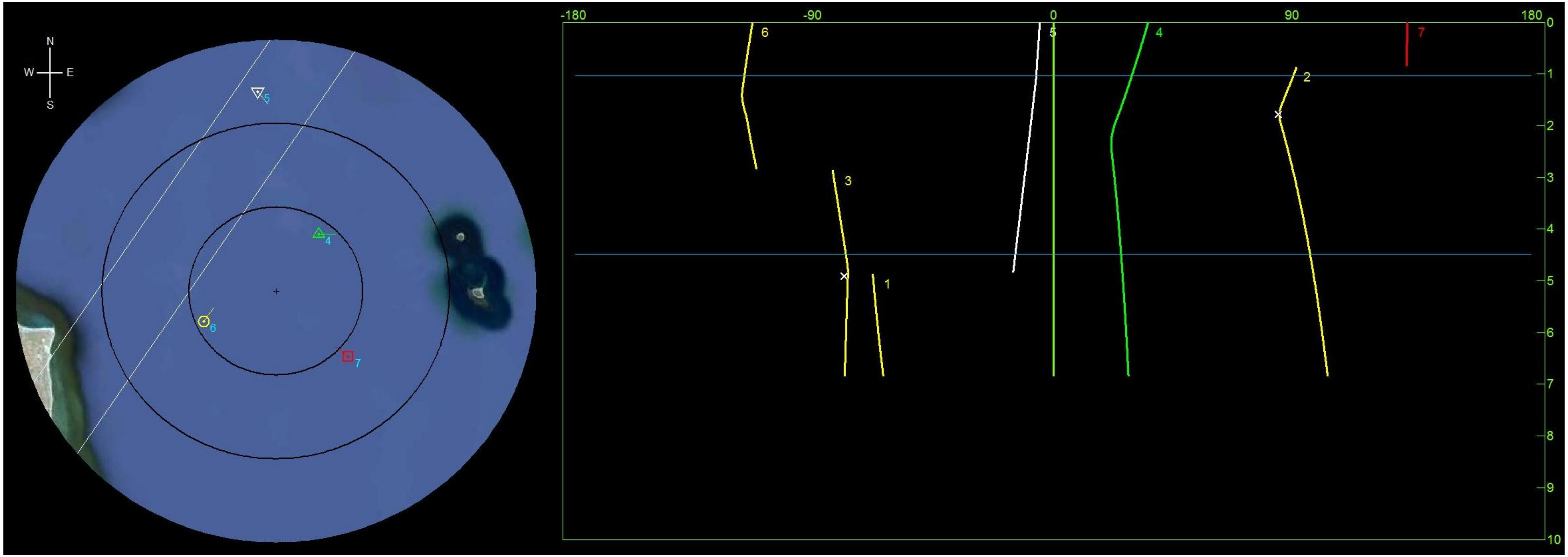

Figure 2. The left image is the picture display, and the right image shows the waterfall display in SIM(sub)3.0. On the picture display, the concentric rings indicate the distance from Ownship in the centre of the screen. The rings extend in 5 km increments. The parallel white lines indicate a shipping lane. Four contacts are currently displayed in this figure: the red square indicates a contact that has been classified as an enemy, the green triangle is a friendly, the white inverted triangle is a merchant, and the yellow circle has yet to be classified. The lines projecting from the centre of each icon indicate the current heading of each contact and the numbers indicate which contact each is. The waterfall display has the angle relative to Ownship on the x-axis, and the time elapsed in minutes on the y-axis. Each contact track is numbered and the colour matches the classification given to the contact. Participants can mark the closest point of approach with a small white x, and can mark horizontal blue lines to keep track of timing. When a contact abrupts off the screen (e.g. vessel 3), the track terminates.

Successful application of these rules required a participant to judge how long each contact had been behaving a certain way (for example, how long contact 4 had been within 5 km of Ownship). As such, participants needed to track time. To assist with this, the y-axis of the waterfall display indicated time elapsed in minutes (using a one minute scale). By right-clicking on the waterfall display participants could leave a blue horizontal line to mark the beginning/end of certain target contact behaviours. In addition to this, each participant was provided with a pencil and paper to write notes throughout the experiment. Participants could, for example, record when a contact changed course or moved into a different target area.

Contact classification performance was measured using:

1. **Accuracy** – The total number of correct classification actions performed.
2. **Response Time** – The speed at which a correct classification action was made. This response time was relative to when the contact had spent enough time performing the classification rule target behaviour (e.g., had spent four minutes in a shipping lane).

During each scenario there were a total of nine contacts that remained on the display long enough to be classified. These contacts were on screen for a total of eight minutes each. There were always four contacts on the display. Every two minutes a contact would abrupt out (disappear from the picture display and cease to be updated on the waterfall display), and a new contact would abrupt in. One purpose of this was to remove classified contacts from the display and to allow new unclassified contacts to appear without building up participant workload by increasing the number of contacts to monitor. The second purpose was to simulate operational settings where contacts abrupt in and out of the display. Contact movement on the display was constant at 30 km/hr and the contact position and heading was updated every 0.25 seconds. Ownship remained stationary.

### 2.2.2. Closest Point of Approach (CPA) task

In addition to classifying contacts, participants were required to mark the closest point of approach (CPA) of contacts to Ownship. To determine CPA, participants needed to monitor the change in track gradient on the waterfall display as well as the changes in contact heading on the picture display. Participants were asked to mark any CPAs that occurred on the waterfall. A CPA could be marked by clicking the closest point of approach on the contacts waterfall display track with the mouse. When the participant left-clicked on this point, a CPA designation icon was then attached to the contact track (analogous to the labelling conducted by track managers). In this case, a small white cross appeared wherever the participant left-clicked on the waterfall display.

The SIM(sub) program determined a contact CPA to Ownship as the point where the radial distance between a contact and Ownship changed from decreasing to increasing. In other words, a CPA occurred when a contact went from heading towards Ownship, to heading away from Ownship. The number of CPAs in each scenario ranged between 9 and 12. Not all contacts had a CPA – for example, a contact might have headed consistently away from Ownship. To be scored as correct, participants needed to mark CPAs on the waterfall display within ± 5 degrees on the x-axis and ± 0.4 minutes on the y-axis from the actual position that the CPA occurred.

CPA task performance was measured by:

1. **Accuracy** – The number of CPAs marked out of the total number of CPAs presented.

### 2.2.3. Emergency Surface Task

In addition to the Contact classification task and the CPA task, participants also completed an emergency surface task. The purpose of this emergency surface task was to introduce a performance measure that required information to be summarised from both the classification task (predominantly using radar display) and the CPA task (predominantly using waterfall display).

This task involved sending a recommendation to the Officer on Watch that Ownship was safe to make an emergency surface. Participants did this by pressing the 's' key. Each of the scenarios lasted for 24 minutes, and within each scenario there was a single window of between 2 and 3 minutes duration where the following emergency surface conditions were satisfied:

1. No unidentified or enemy vessels were within 5 km of Ownship (participants used the picture display to judge this).
2. At least one CPA had occurred in the last 5 minutes (participants used the waterfall display to judge this).

Emergency surface performance was measured by:

1. **Accuracy** – Whether the surface was performed during the correct time window.

2. **Response Time** –The speed of the emergency surface after the required conditions were satisfied.

When participants performed an emergency surface the message "*Emergency surface successful*" appeared in red in the centre of the picture display.

## 2.3. Situation Awareness Measurement

On Day 1 of the experiment participants were trained on how to perform the simulated submarine track management tasks. On Day 2 they completed three test scenarios. In one test scenario they were required to respond to online SPAM questions (hereafter referred to as the SPAM condition). In another scenario they answered offline SAGAT questions (SAGAT condition). In their third scenarios, participants were not presented with either SPAM or SAGAT questions (control condition). In all three scenarios, including the control conditions, the SART subjective measure of SA was administered at the end of each scenario in the form of a written questionnaire.

### 2.3.1. SPAM Administration

SPAM questions were presented at regular intervals. There were 160, 220, or 280 seconds delay between SPAM questions. The questions were delivered over the headphones, and this auditory presentation provided face validity (it more closely replicates the conditions in a submarine control room compared to questions presented visually on the screen as in Loft and Morrell, 2011). Each question began with a ready prompt ("*Are you ready for a question?*") and was followed by a SPAM question (e.g. "*Is vessel four currently heading towards the shipping lane?*"). The SA probes were designed to assess Level 1, 2, and 3 SA. Within each SPAM scenario there were: two Level 1 questions, three Level 2 questions, and one Level 3 question (there were a total of six questions delivered in each SPAM scenario). The question template for SPAM scenarios was as follows:

**Level 1 SA:**

1. Which vessel is currently in *X*?
2. Is vessel Y currently heading towards *X*?

**Level 2 SA:**

1. How many times has vessel *Y* zigged?
2. Has there been a CPA in the past five minutes?
3. Has vessel *Y* been in area *X* for at least *Z* minutes?

**Level 3 SA:**

1. Could vessel *Y* be in area *X* within the next four minutes?

*Note that the values of *X*, *Y,* and *Z* varied for each of the three scenarios.

Participants were instructed to respond to the ready prompt by pressing the 'Y' key when they were ready to answer the SPAM question. They had a 30 second window following the ready prompt in which they could accept the question. If the participant failed to press 'Y' during this interval, no SPAM question was asked. Once a participant had indicated readiness, the SPAM question was presented over headphones. To answer a question, the participants pressed single digit numeric keys (0 – 9), or 'Y' for yes, or 'N' for no. They were instructed to answer all questions based on the current state of their displays at the time the question was asked (as opposed to based on when they responded). The reason for this was that the situation on the displays was constantly evolving and we wanted to ensure that all participants were responding to a similar current display. There was a 30 second window following the SPAM question in which participants could enter a response. If the participant failed to enter any response, or answered after the 30 second window elapsed, then the SPAM question was considered missed and scored as incorrect. SPAM questions were also scored incorrect if participants provided the wrong answer. To help participants answer questions about the future state of the display, participants could right-click with the mouse on any contact on the radar display to bring up a four-minute trajectory circle placed around the contact. This light green circle indicated how far that particular contact would be able to move in any direction within the next four minutes.

SPAM was measured by:

1. **SPAM Accuracy** – The proportion of SPAM questions correctly answered.

2. **SPAM Probe time** – The speed at which participants indicated readiness by accepting the question (for correct answers only). This response time likely indicates how hard the participant was working at the time the probe was administered.

3. **SPAM Response time** – The time taken to answer SPAM questions after the probe was accepted and the content of the question was delivered (for correct answers only).

### 2.3.2. SAGAT Administration

In the SAGAT condition, the simulation was paused (and the screen blanked) four times each scenario. During this task pause the participants' SA was queried by the presentation of a number of SAGAT questions. The SAGAT procedure we employed, and the template we used for SAGAT questions, was based on recommendations by Jones and Endsley (2004). The questions were presented visually (i.e., in written form) on the display. When a SAGAT freeze occurred within a scenario, a blank map was presented to participants on the picture display (the scenario map and all contacts were removed) and the waterfall display was blanked. The participants were first instructed to click on the blank map to locate one of the four contacts which was on the screen at the time of the freeze. The vessel they needed to locate was specified, e.g. "*Place vessel 4 on the screen*". This SAGAT question assessed Level 1 SA and was always the first SAGAT question presented.

Participants were then presented five additional SAGAT questions. Half of these questions were identical to those presented during SPAM administration. Participants were asked one additional Level 1 SA question (two Level 1 in total – including the vessel location question), three Level 2 questions, and one Level 3 question. There were a total of six questions delivered during each SAGAT freeze and the question template for SAGAT scenarios was as follows:

**Level 1 SA (two asked per freeze):**

1. Place vessel *Y* on the screen.
2. Which vessel is currently in *X*?
3. Is vessel *Y* currently heading towards *X*?

**Level 2 SA (three asked per freeze):**

1. How many times has vessel *Y* zigged?
2. Has there been a CPA in the past five minutes?
3. Has vessel *Y* been in area *X* for at least *Z* minutes?
4. What was the last vessel to appear/disappear from the display?
5. Has vessel *Y* had a CPA?
6. What was the last vessel to enter area X?

**Level 3 SA (one asked per freeze):**

1. Could vessel *Y* be in area *X* within the next four minutes?
2. Could vessel *Y* be north/south/east/west of Ownship in four minutes time?

*Note that the values of *X*, *Y*, and *Z* varied for each of the three scenarios.

SAGAT performance was measured by:

1. **SAGAT Accuracy** – The proportion of SAGAT questions correctly answered.

2. **SAGAT Response time** – The time taken to answer SAGAT questions (for correct answers only).

### 2.3.3. SART Administration

Following completion of each of the three test scenarios, participants completed the 10-dimension paper-based SART questionnaire (Selcon & Taylor, 1990). Participants provided ratings on a number of scales regarding the degree to which they perceived: 1) the demands on their resources, 2) their supply of resources, and 3) their understanding of the situation. A SART score was generated by combining scores on the three subscales. This combined score provided a measure of subjective SA [SA = Understanding – (Demand-Supply)].

## 2.4. Measuring Workload

### 2.4.1. NASA-TLX

The NASA-TLX (Hart & Staveland, 1988) was administrated at the completion of each scenario. The NASA-TLX is a two-part paper-based procedure for measuring subjective workload based on 'magnitude of workload' ratings and the 'sources of workload' weightings. Participants made magnitude of workload ratings for the scenario on six 20-point scales: mental demand, physical demand, temporal demand, performance, effort, and frustration. After finishing each of the three test scenarios, the participants completed 15 pair-wise scale comparisons (e.g., mental demand vs. frustration) for each task type (CP task, contact classification task, and emergency surface task). That is, participants rated which source of workload (e.g., mental demand vs. frustration, temporal demand vs. effort, etc) contributed most to their workload for each task. These pair-wise comparisons provided the unique relative weighting for each of the scales (mental demand, physical demand, temporal demand, performance, effort, and frustration). Participants' overall NASA-TLX score was determined by multiplying the magnitude of workload ratings for each scale by the corresponding workload weighting for the corresponding scale, adding the values for all the scales together, and then dividing the total by 15.

### 2.4.2. ATWIT

The Air Traffic Workload Input Technique (ATWIT - Stein, 1985) measured mental workload in real-time by presenting workload probes on screen that required participants to indicate what level of workload they were experiencing at that particular moment. The advantages to using an online workload measure is that recorded changes in workload *within* a scenario (unlike the NASA-TLX which only provides a single estimate of subjective workload at the end of a scenario and is thus more reliant on participant memory).

ATWIT probes were delivered four times during each scenario. The workload probes in the scenarios where SAGAT was administered were delivered either 15 seconds before a SAGAT freeze, or 15 seconds after a SAGAT freeze (equal numbers were delivered before and after a SAGAT freeze). In the SPAM scenarios, the ATWIT probes were also delivered 15 seconds before or after the SPAM question. It is crucial to note that the presentation of ATWIT was not a predictive cue that a SPAM/SAGAT question would follow. This is because ATWIT probes were only presented before two of the six SPAM questions, and before two of the four SAGAT question sets. ATWIT probes appeared on the screen with the message "*Rate your workload*" (where 1 was very low workload, and 10 was very high workload). Participants clicked on the number between 1 and 10 that most accurately represented how hard they were currently working. Each workload probe only remained on the screen for 10 seconds, after which it disappeared.

## 2.5. Procedure

### 2.5.1. Day 1

The experiment was conducted over two days (with one day separation). Due to the complexity of the task, a two hour training session was designed to introduce participants to simulated submarine track management. The training day began with a 30 minute computer-driven presentation which covered all of the requirements of the experiment. This presentation also included a short narrated video of the task. Participants were encouraged to ask questions at any point on the training day. Following completion of the training presentation, the participants completed two practice scenarios, one with embedded SAGAT questions and one with embedded SPAM questions. The practice sessions also included all other measurements (NASA-TLX, ATWIT and SART were all delivered). At the end of each training scenario participants were presented with written feedback on the screen and the experimenter discussed any areas of poor performance with the participant – reiterating the relevant task details or rules as required. This feedback included the participants' CPA and contact classification accuracy, emergency surface success, and the proportion of SPAM or SAGAT questions answered correctly. Both training scenarios were completed without automation. At the beginning of Day 2 the participants received a short 10 minute refresher presentation where the important information from Day 1 was condensed and repeated. Participants in the automation condition had extra slides in their refresher presentation that introduced how automation would be provided.

### 2.5.2. Day 2

On Day 2 participants completed three scenarios. The experiment consisted of a mixed (within subjects /between subjects) design. The within subjects factor was SA measurement (SAGAT, SPAM, or control). In one scenario participants were required to respond to online SPAM questions. In another scenario they answered offline SAGAT questions. In their third scenario, participants were not presented with either SPAM or SAGAT questions (control condition). In all three scenarios, including the control conditions, the SART subjective measure of SA was administered at the end of each scenario. In each scenario ATWIT and the NASA-TLX were administered.

### 2.5.3. Automation Manipulation

Participants were randomly allocated to either the automation or no-automation condition. Participants in the no-automation condition were required to mark all CPAs that occurred on the waterfall display, as per training. Participants in the automation condition had some of their CPAs automatically marked on the waterfall display. This automation marked a cross on the waterfall display whenever a contact reached its closest point of approach to Ownship. The automation however only applied to CPAs that were inside the 5 km radius from Ownship (within the first 5 km ring marked from Ownship on the picture display; see Figure 2). Thus, participants in the automation condition were explicitly told that they were still required to manually mark the CPAs occurring in the area more than 5 km from Ownship. When a CPA was automatically marked, a message appeared on the screen that said "*Accept CPA?*" and that message remained until the participant clicked either '*accept*' or '*reject*'. An auditory tone coincided with the CPA notification to capture participants' attention.

If participants clicked '*accept*' then the automation alert message disappeared and the automated CPA marking remained on the waterfall display. If participants clicked '*reject'* then the alert message disappeared and the automated CPA marking was removed from the waterfall display. Automation messages time out after 90 seconds if participants click neither accept nor reject. At this point the automation message box disappeared and the automated CPA was removed from the waterfall display. The automation applied to approximately half of the scripted CPAs (the other half occurred outside the 5km to Ownship so were required to be manually detected).

# 3. Results

## 3.1. Data Management and Descriptive Statistics

Response time outliers were removed at the within-subject level. That is, any response time more than 2.5 standard deviations from a participant mean for the contact classification task, SPAM question set, or the SAGAT question set were excluded. This resulted in two participants' data being trimmed for the contact classification task, and one for the SAGAT question set.

To ensure that there were no significant differences in skill acquisition over training between the participants assigned to the automation condition and participants assigned to the no-automation condition, the performance data from the training scenarios on Day 1 were analysed. All participants completed training without the assistance of automation, regardless of whether they were subsequently assigned to the automation or no-automation condition on Day 2. The Day 1 data revealed no significant difference between participants in the automation and no automation condition on accuracy or response time on the emergency surface task (*t*s < 1), CPA task (*t*s < 1), or contact classification task (*t*s < 1). On Day 1 there were also no significant differences between participants in the automation and no automation condition on the SPAM measures (*t*s < 1) or SAGAT measures ($t(163) = 1.63$, $p = 0.11$). There was also no difference between the two groups on ATWIT-before ($t < 1$), ATWIT-after ($t(326) = 1.56$, $p = 0.12$), NASA-TLX ($t(329) = 1.48$, $p = 0.14$), or the SART measures ($t < 1$). Thus, as would be expected from the fact we randomly assigned participants to the Day 2 automation and no automation conditions, there were no differences in responses from these two groups on Day 1.

Table 1 summarises the descriptive statistics for each condition (SPAM, SAGAT, and Control) on Day 2 collapsed across the automation and no-automation between-subject conditions. Participants performed well on both the emergency surface and the contact classification task, but found the CPA task more difficult. It should be noted that response times for SPAM questions presented in Tables 1 and 2 are for correct SPAM responses only. Thus, one potentially important correlation that is not shown in Tables 1 and 2 is the relationship between SPAM accuracy and overall SPAM response time (for both correct and incorrect SPAM responses). There was a trending negative correlation between SPAM accuracy and SPAM response time ($r = -0.126$, $p = .09$). This indicates that SPAM accuracy decreased as SPAM response time increased. This is important because it shows that participants were more likely to answer SPAM incorrectly if they took longer to make SA decisions. Although this correlation does not reach significance, it clearly indicates that it was not the case that faster SA responders were also less accurate (in that case we would have expected a positive correlation, i.e., speed-accuracy trade-off).

Correlations between variables as a function of the three within subject conditions (SPAM, SAGAT, and Control) are presented in Tables 2, 3, and 4, respectively. The before-ATWIT, after-ATWIT, and the NASA-TLX workload measures were highly correlated, indicating concurrent validity. Interestingly, SPAM probe acceptance time (a proxy for workload according to the designers of SPAM – Durso & Dattel, 2004) did not correlate with the workload measures. The subjective SA measure (SART) did not correlate significantly with any performance or SA measure. However, SART correlated with the subjective workload measures in all three experimental conditions; participants rated that they had higher SA when they also rated that they were under a higher level of workload.

Table 1. Descriptive statistics (standard deviations are in parentheses) for the SPAM, SAGAT, and control conditions. The data below are collapsed across the automation and no-automation conditions. Response times (RT) are in seconds.

**SA Measures**

| | *SPAM probe RT* | *SPAM accuracy* | *SPAM RT (correct only)* | *SAGAT accuracy* | *SAGAT RT* | *SART* |
|---|---|---|---|---|---|---|
| SPAM | 1.73 (1.31) | 0.86 (0.13) | 1.97 (1.01) | - | - | 5.25 (1.41) |
| SAGAT | - | - | - | 0.67 (0.19) | 6.01 (1.51) | 5.29 (1.39) |
| Control | - | - | - | - | | 5.18 (1.46) |

**Task Performance**

| | *Emergency surface hit rate* | *Emergency surface RT* | *CPA hit rate (outer)* | *Classification hit rate* | *Classification RT (sec)* | |
|---|---|---|---|---|---|---|
| SPAM | 0.76 (0.43) | 31.99 (42.34) | 0.54 (0.25) | 0.74 (0.23) | 19.30 (13.43) | |
| SAGAT | 0.74 (0.44) | 40.99 (46.53) | 0.56 (0.25) | 0.74 (0.22) | 19.51 (13.64) | |
| Control | 0.75 (0.43) | 30.57 (36.76) | 0.56 (0.27) | 0.76 (0.22) | 18.80 (12.97) | |

**Subjective Workload**

| | *ATWIT-before* | *ATWIT-after* | *NASA-TLX* | | | |
|---|---|---|---|---|---|---|
| SPAM | 5.16 (1.78) | 5.32 (1.68) | 10.50 (3.45) | | | |
| SAGAT | 5.13 (1.67) | 5.56 (1.65) | 10.82 (3.18) | | | |
| Control | 5.02 (1.72) | 5.20 (1.69) | 10.22 (3.27) | | | |

Table 2. Correlations between performance, SA, and workload measures for the **SPAM condition**. Significant correlations are indicated by asterisks (* is significant at the 0.05 level, and ** is significant at the 0.01 level).

| | 1. Emergency surface RT | 2. CPA hit rate (outer) | 3.Classification hit rate | 4.Classification RT | 5. SPAM probe RT | 6. SPAM RT | 7. SPAM accuracy | 8. ATWIT (before) | 9.ATWIT (after) | 10. NASA-TLX | 11.SART |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. | - | -0.252** | -0.170* | 0.293** | 0.065 | 0.105 | -0.085 | 0.102 | 0.102 | 0.165 | -0.006 |
| 2. | | - | 0.179* | -0.163* | -0.152* | -0.204** | 0.229** | -0.024 | -0.013 | -0.068 | 0.064 |
| 3. | | | - | -0.146 | 0.053 | -0.176* | 0.301** | -0.187* | -0.147* | -0.269** | -0.045 |
| 4. | | | | - | 0.034 | 0.109 | -0.180* | 0.150* | 0.103 | 0.216** | 0.026 |
| 5. | | | | | - | 0.202** | 0.027 | 0.134 | 0.097 | 0.147 | 0.001 |
| 6. | | | | | | - | -0.297** | 0.149* | 0.162* | 0.133 | -0.035 |
| 7. | | | | | | | - | -0.086 | -0.112 | -0.095 | -0.087 |
| 8. | | | | | | | | - | 0.884** | 0.696** | 0.293** |
| 9. | | | | | | | | | - | 0.725** | 0.342** |
| 10. | | | | | | | | | | - | 0.384** |
| 11. | | | | | | | | | | | - |

Table 3. Correlation between performance, SA, and workload measures for the **SAGAT condition**. Significant correlations are indicated by asterisks (* is significant at the 0.05 level, and ** is significant at the 0.01 level).

| | 1. Emergency surface RT | 2. CPA hit rate (outer) | 3. Classification hit rate | 4. Classification RT | 5. SAGAT accuracy | 6. SAGAT RT | 7. ATWIT (before) | 8. ATWIT (after) | 9. NASA - TLX | 10. SART |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | - | -0.023 | -0.345** | 0.211** | -0.159 | -0.021 | -0.027 | -0.102 | 0.040 | -0.077 |
| 2. | | - | 0.192** | -0.232** | 0.277** | -0.107 | -0.007 | -0.020 | -0.067 | -0.033 |
| 3. | | | - | -0.237** | 0.444** | -0.055 | -0.039 | 0.005 | -0.193* | 0.075 |
| 4. | | | | - | -0.258** | 0.121 | 0.050 | 0.065 | 0.091 | -0.102 |
| 5. | | | | | - | 0.018 | -0.028 | -0.060 | -0.176* | 0.053 |
| 6 | | | | | | - | 0.012 | 0.065 | -0.02 | -0.061 |
| 7. | | | | | | | - | 0.889** | 0.662** | 0.233** |
| 8. | | | | | | | | - | 0.679** | 0.337** |
| 9. | | | | | | | | | - | 0.270** |
| 10. | | | | | | | | | | - |

Table 4. Correlation between performance, SA, and workload measures for the **Control condition**. Significant correlations are indicated by asterisks (* is significant at the 0.05 level, and ** is significant at the 0.01 level).

| | 1. Emergency Surface RT | 2. CPA hit rate (outer) | 3. Classification hit rate | 4. Classification RT | 5. ATWIT (before) | 6. ATWIT (after) | 7. NASA- TLX | 8. SART |
|---|---|---|---|---|---|---|---|---|
| 1. | - | -0.038 | -0.203* | 0.312** | 0.057 | 0.056 | 0.078 | 0.027 |
| 2. | | - | 0.233** | -0.132 | 0.014 | -0.082 | -0.222** | -0.004 |
| 3. | | | - | -0.117 | -0.100 | -0.128 | -0.238** | -0.004 |
| 4. | | | | - | 0.074 | 0.064 | 0.124 | -0.101 |
| 5. | | | | | - | 0.866** | 0.695** | 0.354** |
| 6. | | | | | | - | 0.695** | 0.309** |
| 7. | | | | | | | - | 0.326** |
| 8. | | | | | | | | - |

## 3.2. Predictive Validity of SPAM and SAGAT SA Measurements

SAGAT accuracy was found to be significantly correlated with SPAM accuracy, $r = 0.20$, $p < 0.05$. SAGAT response time was found to be significantly correlated with SPAM response time, $r = 0.30$, $p < 0.01$. Concurrent validity was therefore demonstrated, but there was only a modest amount of shared variance between the SAGAT and SPAM measures.

Hierarchical multiple regressions were then conducted to examine the extent to which SPAM and SAGAT predicted unique variance in each performance measure over and above the variance in performance explained by the automation manipulation and by the subjective SA (SART) and workload (ATWIT & NASA-TLX) measures. In the first block of the hierarchical multiple regressions we entered automation (categorical variable), SART, NASA-TLX, ATWIT-Before and ATWIT-After. In the second block of the hierarchical multiple regressions we entered either SPAM (probe response time, and accuracy and response time to question) or SAGAT (accuracy and response time). Please note that separate hierarchical multiple regression analyses were conducted to assess the predictive validity of SPAM and SAGAT, because these two SA measures were administered in different scenarios (i.e., they represented a within-subjects variable).

### 3.2.1. CPA Accuracy

The automation provided to some of the participants on Day 2 was designed to assist participants by automating the detection of CPAs that occurred within 5 km of the Ownship. If participants accepted the automation then CPA performance should be 100% accurate for these contacts and there is no variance in CPA for the SA measures to predict. However, participants in the automation condition were still required to detect the CPA's for contacts located outside the 5 km zone from Ownship, and thus we can assess whether SPAM and SAGAT can predict variance in this performance. But first, to confirm that there was no difference in the inherent difficulty of detecting contact CPAs that occurred within 5 km compared to those that occurred outside the 5 km zone, we compared the hit rates for inner and outer radius CPAs for the participants not provided automation. For these participants in the no-automation condition, there was no significant difference in CPA performance for contacts located in the inner compared to outer 5km radius ($t > 1$). Thus, to allow for comparison between the automation and no-automation condition (i.e., to allow automation to be entered as a categorical predictor into the regression equation), we only included CPA performance for the vessels located outside the 5 km zone.

In the first regression, CPA hit rate was entered as the criterion variable and SPAM probe response time, SPAM accuracy, and SPAM (question) response times were entered as the predictors in the second block. The first model was found to be significant, $F(5, 167) = 3.53$, $p < 0.01$, with automation a significant predictor ($\beta = -0.29$, $p < 0.01$). The provision of automation resulted in a poorer CPA hit rate (for contacts outside the 5 km radius that needed to be detected manually) than the provision of no-automation. The first model explained 10% of the variance in the CPA hit rate. The second model was also significant, $F(8, 164) = 4.46$, $p < 0.01$, and the significant predictors were automation ($\beta = -0.27$, $p < 0.01$) and SPAM accuracy ($\beta = 0.19$, $p < 0.05$). The $R^2$ change for the second model was significant ($p < 0.01$) and the second model explained an additional 8% of variance in CPA.

A second regression was conducted with CPA hit rate entered as the criterion, and SAGAT accuracy and response time entered as the predictors in the second block. The first model was not significant (*F* < 1) and only accounted for 1% of the variance in CPA hit rate. The second model was significant, *F* (7, 166) = 2.45, *p* < 0.05, and it accounted for an additional 8% of variance in CPA. There was only one significant predictor in the second model and that was SAGAT accuracy (β = 0.28, *p* < 0.01).

In summary, neither the workload measures nor SART (the subjective measure of SA), were significant predictors of CPA performance. However, both SPAM and SAGAT accuracy predicted CPA performance above the factors entered in the first model (ATWIT, NASA, SART, and automation). SPAM response time was also a significant predictor, though SAGAT response time was not. The data indicate that as participants' SA increased, as measured by SPAM and SAGAT, they detected more CPAs that occurred for vessels outside the 5 km radius. Automation was a significant negative predictor of performance, but in the SPAM condition only. This indicates that automating CPAs for vessels located within the 5km radius of Ownship has a negative impact on the ability of participants to manually detect CPAs for vessels that occurred outside the 5 km radius, a point we return to later in this report.

### 3.2.2. Contact Classification Accuracy

Classification accuracy was entered as the criterion and SPAM probe response time, SPAM accuracy, and SPAM (question) response time were entered in the second model. Both models were significant (*F* (5, 167) = 3.54, *p* < 0.01 and *F* (7, 165) = 5.34, *p* < 0.01, for the first and second models respectively). The first model accounted for 10% of the variance in contact classification accuracy, with NASA-TLX a significant predictor (β = -0.29, *p* < 0.01). The second model accounted for an additional 10% of variance in contact classification accuracy, with NASA-TLX (β = -0.29, *p* < 0.01) and SPAM accuracy (β = 0.27, *p* < 0.01) both significant predictors.

In the SAGAT condition, classification accuracy was entered as the criterion, and SAGAT response time and accuracy were entered as the predictors in the second model. Both models were significant (*F* (5, 168) = 3.29, *p* < 0.01 and *F* (7, 166) = 7.68, *p* < 0.01, for the first and second models respectively). The first model accounted for 9% of the variance in contact classification accuracy, with NASA-TLX a significant predictor (β = -0.36, *p* < 0.01). The second model accounted for an additional 16% of variance, and NASA-TLX (β = -0.24, *p* < 0.01), ATWIT-after (β = 0.35, *p* < 0.05), and SAGAT accuracy (β = 0.41, *p* < 0.01) were significant predictors.

In summary, both SPAM and SAGAT accuracy predicted contact classification accuracy over and above the factors entered in the first model (ATWIT, NASA, SART, and automation). As participants' SA increased, as measured by SPAM and SAGAT, the accuracy of contact classification increased. Subjective workload was also a significant predictor of contact classification accuracy; as participants' subjective workload ratings increased so did their accuracy on the contact classification task.

### 3.2.3. Contact Classification Response Time

In addition to classification accuracy, we repeated the regression analyses outlined above for response times (correct only) to the contact classification task. For the SPAM condition, the first model was almost significant, *F* (5, 164) = 2.09, *p* = 0.07, with NASA-TLX a significant predictor (β = 0.29, *p* <

0.05). The first model accounted for 6% of the variance in contact classification response time. The second model was significant, $F$ (7, 162) = 2.35, $p$ < 0.05, with NASA-TLX (β = 0.29, $p$ < 0.05) and SPAM accuracy (β = -0.17, $p$ < 0.05) both significant predictors. The second model added an additional 3% of variance in contact classification response time.

For the SAGAT condition, the first model was not-significant, $F$ (5, 167) = 1.08, $p$ = 0.37, and accounted for 3% of the variance in classification response time. The second model was significant, $F$ (7, 165) = 2.42, $p$ < 0.05, and accounted for an additional 6% of the variance in classification response time. The only significant predictor in the second model was SAGAT accuracy (β = -0.25, $p$ < 0.01).

In summary, the hierarchal regression analyses conducted on the contact classification response time data were consistent with the analyses for the contact classification accuracy data. SPAM accuracy and SAGAT accuracy predicted unique variance in classification task response time over and above the control variables. Participants with higher SA as measured by SPAM and SAGAT responded more quickly to the contact classification task (correct responses only).

### 3.2.4. Emergency Surface Task Accuracy

To examine the extent to which SPAM and SAGAT predicted accuracy on the emergency surface task (binary data; success versus failure) we used logistic regression. In the logistic regression, emergency surface accuracy was the criterion, and SPAM probe response time, SPAM accuracy, and SPAM (question) response time were entered as predictors in the second model (predictors in the first model remained the same as in the hierarchical regressions). The first model was significant, $\chi^2$ (5, N = 173) = 11.16, $p$ = 0.048. The Wald criterion indicated that only NASA-TLX was close to making a significant contribution to the model ($p$ = 0.09) and the first model accounted for 9% of the variance (as determined by the Nagelkerke R-square). The second model was also significant, $\chi^2$ (2, N = 173) = 15.70, $p$ < 0.01, and accounted for an additional 22% variance in emergency surface accuracy. SPAM accuracy ($p$ = 0.02) and SPAM response time ($p$ = 0.02) were significant predictors.

For the SAGAT condition, emergency surface accuracy was the criterion, and SAGAT accuracy and response time were entered as predictors in the second model. The first model was not significant ($p$ = 0.30). The Wald criterion indicated that only NASA-TLX was close to making a significant contribution to the first model ($p$ = 0.06). The first model accounted for 5% of the variance (as determined by the Nagelkerke R square). The second model was significant, $\chi^2$ (2, N = 174) = 15.91, $p$ < 0.01 and accounted for an additional 17% variance in emergency surface accuracy. SAGAT accuracy ($p$ = 0.001) and SAGAT response time ($p$ = 0.03) were significant predictors.

In summary, SPAM and SAGAT accuracy and response time predicted unique variance in the accuracy of the emergency surface task, over and above the SART and workload control variables. Participants with higher SA as measured by SPAM and SAGAT responded more accurately to the contact classification task.

### 3.2.5. Emergency Surface Task Response Time

We repeated the regression analyses outlined above for response time (correct only) to the emergency surface task. Emergency surface response time was entered as the criterion variable, and

SPAM variables were entered as the predictors in the second model. The first model was not significant, *F* < 1, and accounted for 3% of the variance in emergency surface response time. The second model was also not significant, *F* < 1 and only added an additional 1% to the total variance accounted for. For the SAGAT condition, the first model was non-significant, *F* (5, 123) = 1.35, *p* = 0.25, and accounted for 5% of the variance in emergency surface response time. The second model was also not significant, *F* (2, 121) = 2.33, *p* = 0.10, and accounted for an additional 4% of the variance.

## 3.3. Automation and the Disruptive Effects of SPAM and SAGAT

### 3.3.1 Use of Automation

Given tha participants were told that the CPA automation tool was 100% reliable it was expected that the vast majority of people would choose to accept the decisions by made the automation on every occasion. However, when the use of automation was examined, only 40% of participants in the automation condition accepted the automated CPA decision on every occasion. Overall, 26% of participants rejected one automated CPA decision, which could have been due to the participant still learning how the automation functioned or due to participants missing an automation alert message due to data limitations in allocation of attention. However, that still left 34% of participants who rejected automation on multiple occasions. The average number of automation rejections from this last group of participants was five. We re-analysed all the data presented in this report with these 34% of participants from the automation condition excluded (including the analyses reported in the sections below). The rationale here was that the effects of automation on the various measures may be more detectable by only including participants that consistently used the CPA decisions provided by the automation. However, we essentially found the same patterns of data as that included in this report when we re-analysed the data in this fashion.

### 3.3.2. Sensitivity of SPAM and SAGAT to Automation

In order to examine the sensitivity of SPAM and SAGAT to automation, three independent-samples t-tests were conducted with automation as a between subjects variable. There was no significant effect of automation on SPAM response time (*t* < 1), SPAM accuracy (*t* < 1), SAGAT response time (*t* (177) = 1.07, *p* = 0.29), or SAGAT accuracy (*t* < 1). This suggests that automation had no effect on participants' objective SA as measured by either SPAM or SAGAT.

### 3.3.3. Effects of SA Administration (and Automation) on Subjective Workload – ATWIT

A 3 condition (SPAM, SAGAT, control) x 2 automation (no automation, automation) x 2 ATWIT probe placement (before, after) mixed-factor ANOVA was conducted to determine whether SA measurement or automation increased subjective workload as measured by the ATWIT probes. Condition and ATWIT probe placement were within-subjects factors, and automation was a between-subjects factor. The dependent variable was the ATWIT subjective workload score (between one and ten). The data are presented in Figure 3.

There was a significant main effect of condition, *F(*2, 352) = 4.82, *p* < 0.01, partial-$\eta^2$ = 0.05. Post-hoc paired-samples t-tests revealed that the SAGAT condition (*M* = 5.35) had significantly higher ATWIT subjective workload scores than the control condition (*M* = 5.11; *t*(178) = 3.18, *p* < 0.01). There was no significant difference between SPAM ($M$ = 5.24) and either the SAGAT condition $t$(178) = -1.35, $p$ = 0.18, or the control condition, $t$(178) = 1.67, $p$ = 0.10. ATWIT probe placement had a main effect on ATWIT workload, $F$(1, 176) = 47.72, $p$ < 0.01, partial-$\eta^2$ = 0.21; workload was higher for probes administered after SA questions ($M$ = 5.36) than for those administrated before SA questions ($M$ = 5.11). There was no main effect of automation, however there was a trend towards automation reducing subjective workload (automation $M$ = 5.04, no automation $M$ = 5.43), $F$(1, 176) = 2.81, $p$ = 0.10, partial-$\eta^2$ = 0.02.

The aforementioned main effects of condition and ATWIT probe placement were qualified by a significant interaction between condition and probe placement, $F$(2, 352) = 5.23, $p$ < 0.01, partial-$\eta^2$ = 0.03. As illustrated in Figure 3, the increase in ATWIT subjective workload scores in the SAGAT condition was most pronounced for the probes delivered 15 seconds *after* the SAGAT freeze occurred, which is exactly what would be expected if SAGAT administration was responsible for the subjective workload increase. This suggests that SAGAT has the potential to be disruptive by increasing operators' subjective workload. There was no conclusive evidence this was the case for SPAM, although the p value for the SPAM vs. control post-hoc comparison was trending toward significance ($p$ = 0.10).

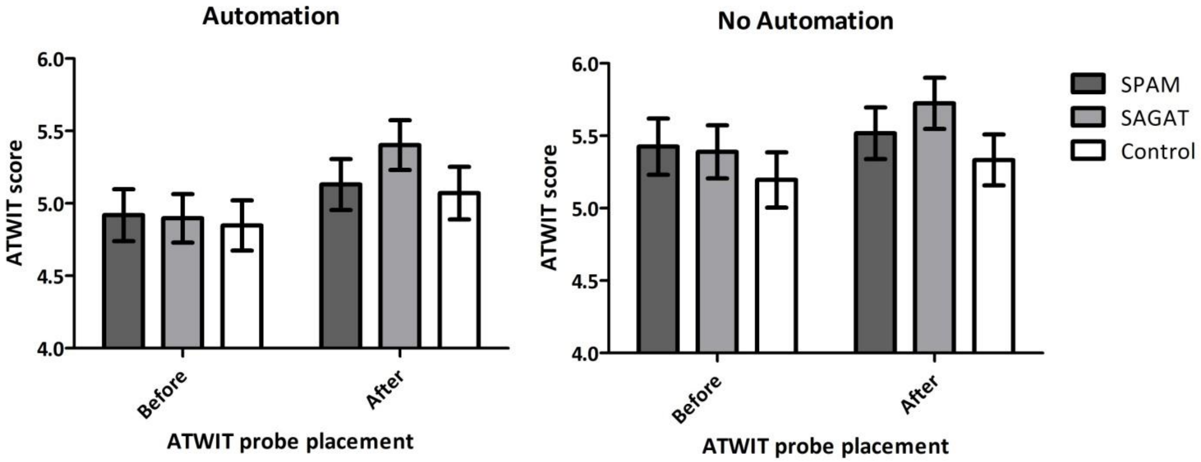

Figure 3. ATWIT subjective workload score for the three within-subjects conditions (SPAM, SAGAT, and Control). The left graph shows ATWIT scores for the automation condition, and the right graph shows ATWIT scores for the no automation condition. ATWIT scores are separated into ATWIT probes that occurred before an SA question, and AWTIT probes that occurred after. Error bars represent the standard error of the mean.

### 3.5.2. Effects of SA Administration (and Automation) on Subjective Workload – NASA-TLX

A 3 condition (SPAM, SAGAT, control) x 2 automation (no automation, automation) mixed factor ANOVA was conducted to determine the extent to which automation and the SAGAT/SPAM administration increased subjective workload as measured by NASA-TLX. Condition was the within-subjects factor, and automation was the between-subjects factor. The dependent variable was the NASA-TLX, where higher NASA-TLX scores indicated higher subjective workload.

As illustrated in Figure 4, there was a significant main effect of condition on the NASA-TLX *(F(*2, 348) = 6.47, $p$ < 0.01, partial-$\eta^2$ = 0.04). Post-hoc paired-samples t-tests revealed that the SAGAT condition ($M$ = 10.83) had significantly higher subjective workload than the control condition ($M$ = 10.22; $t$(176) = 3.46, $p$ < 0.01), and the SPAM condition ($M$ = 10.50; $t$(176) = -2.14, $p$ < 0.05). There was no significant difference between the control and SPAM conditions, $t$(175) = 1.38, $p$ = 0.17. There was no main effect of automation on NASA-TLX workload, *F(*1, 174) = 1.49, $p$ = 0.22, partial-$\eta^2$ = 0.01.

Thus, entirely consistent with the ATWIT data, the NASA-TLX subjective workload scores suggest that SAGAT has the potential to be disruptive by increasing operator subjective workload. There was no conclusive evidence that this was the case for SPAM.

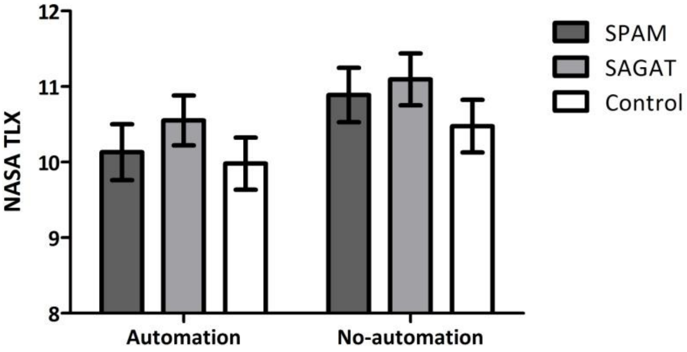

Figure 4. NASA-TLX for the three within-subjects conditions (SPAM, SAGAT, and Control), separated into automation and no-automation conditions. Error bars represent the standard error of the mean.

### 3.5.3. Effects of SA Administration (and Automation) on CPA hit rate (outer 5km)

A 3 condition (SPAM, SAGAT, control) x 2 automation (no automation, automation) mixed factor ANOVA was conducted on the extent to which automation and SAGAT/SPAM administration impacted task performance as measured by CPA hit rate (for manually detected CPAs occurring outside the 5 km radius). Condition was the within-subjects factor, and automation was the between-subjects factor. The dependent variable was the CPA hit rate.

There was no significant effect of condition on CPA hit rate *(F* < 1) or automation, however there was a trending main effect of automation *(F(*1, 178) = 3.34, $p$ = 0.07, partial-$\eta^2$ = 0.02), where automation led to a poorer CPA hit rate (automation $M$ = 0.53, no-automation $M$ = 0.58). This was qualified by an interaction between condition and automation *(F(*2, 356) = 3.57, $p$ < 0.05, partial-$\eta^2$ = 0.02). Figure 5 shows that this interaction was due to the CPA hit rate being much lower in the automation condition, compared to the no automation condition, when SPAM was being administered during the scenario. In contrast, the effect of automation on CPA performance was smaller for the SAGAT and control conditions.  This result is consistent with outcomes from the hierarchical multiple regression analyses where automaton was a significant negative predictor of CPA performance for contacts located outside the Ownship's 5 km radius under conditions were SPAM was being administered. Specifically, under SPAM conditions, , the automation of CPAs for contacts located inside Ownship's 5 km radius led to 5% poorer manual CPA performance for contacts located outside the ownship's 5 km radius when compared to the no automation condition in which participants were charged with manually detecting all CPAs, regardless of their location.

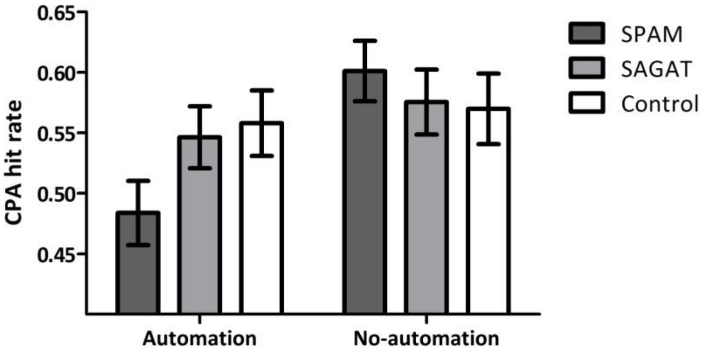

Figure 5.  CPA hit rate (between 0 and 1) for the three within-subjects conditions (SPAM, SAGAT, and control), separated into automation and no-automation conditions.  Error bars represent the standard error of the mean.

### 3.5.4. Effects of SA Administration (and Automation) on Contact Classification Accuracy and Response Time

A 3 condition (SPAM, SAGAT, control) x 2 automation (no automation, automation) mixed factor ANOVA was conducted to determine the extent to which automation and SAGAT/SPAM administration impacted on contact classification.  Condition was the within-subjects factor, and automation was the between-subjects factor.  This analysis was conducted with classification response time and classification hit rate as the dependent variables. Figure 6 displays the data.  There was no main effect of condition (*F(*2, 356) = 1.20, *p* = 0.30, partial-$\eta^2$ = 0.01) or automation (*F* < 1) (and no interaction).  There was also no effect of condition (*F* < 1) or automation (*F* < 1) on classification response time (and no interaction).

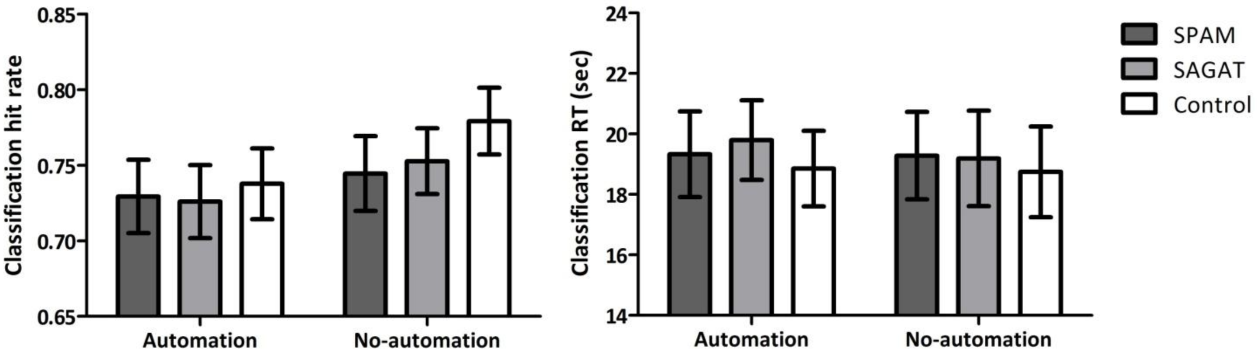

Figure 6. Contact classification performance for three within-subjects conditions (SPAM, SAGAT, and Control), separated into automation and no-automation. The left figure shows contact classification hit rate (0 – 1.0), and the right figure shows response time (seconds). Error bars represent the standard error of the mean.

### 3.5.5. Effects of SA Administration (and Automation) on Emergency Surface Task Accuracy and Response Time.

A 3 condition (SPAM, SAGAT, control) x 2 automation (no automation, automation) mixed factor ANOVA was conducted to determine the extent to which automation and SAGAT/SPAM administration impacted emergency surface task performance. Condition was the within-subjects factor, and automation was the between-subjects factor. This analysis was conducted with emergency surface response time and then emergency surface accuracy as the dependent variables. There was no main effect of condition or of automation on emergency surface accuracy (and no interaction, all *F* < 1). There was also no main effect of either condition (*F(*2, 168) = 1.78, *p* = 0.17, partial-$\eta^2$ = 0.02) or automation (*F* < 1) on emergency surface response time (and no interaction).

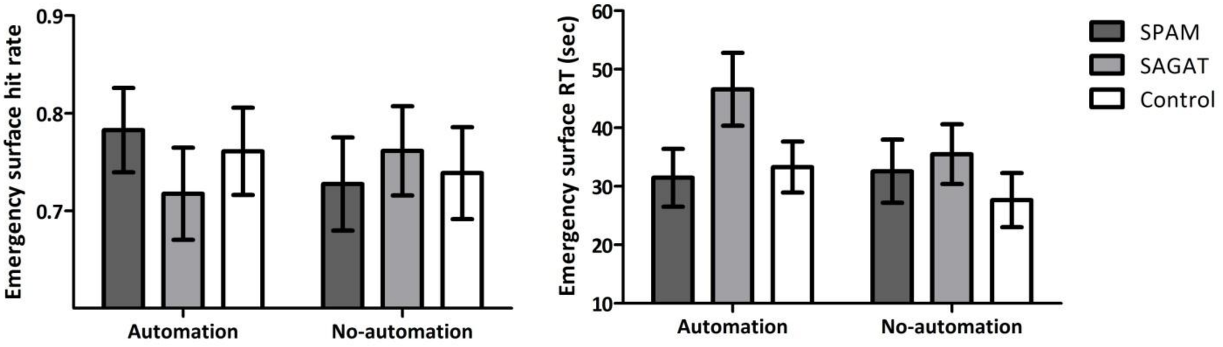

Figure 7. Means for all three within-subjects conditions (SPAM, SAGAT, and Control), separated into automation and no-automation. The left figure shows emergency surface accuracy (0 or 1), and the right figure shows response time (seconds). Error bars represent the standard error of the mean.

## 4. Discussion and Conclusions

The first objective of the current project was to complete a literature review of the benefits and risks that task automation poses for SA and performance for tasks where individuals monitor and control complex dynamic displays. This report was delivered to AOARD in April 2012. The second phase of the project, on which this current report is based, also makes a significant contribution to the DSTO long term work program that aims to provide design advice into future submarine projects. Specifically, the current research builds directly on previous work (Loft & Morrell, 2011) by examining SA measurement and the impact of task automation in higher fidelity simulated submarine track management. The data presented here clearly indicate that both SPAM and SAGAT could predict significant variance in task performance. Crucially, the evidence suggests that the predictive validity of SPAM was comparable to the predictive validity of SAGAT. It was important that SPAM and SAGAT both predicted task performance variance over and above the subjective SA measure (SART) because there is no doubt that SART would be easier to prepare and administer in applied naval settings. Another crucial finding of the current research was that SAGAT administration, but not SPAM administration, increased participant ratings of subjective workload (also see Pierce, 2012). Thus, the predictive validity provided by SAGAT came at the cost of increased participant subjective workload ratings. If SAGAT disrupts the operator from the tasks at hand, this would limit its use in the submarine control room. Assessing SA in real-time using SPAM may prove to be the most advantageous in a practical sense for submarine track management; SPAM administration does not require the task operation to be paused, and also better simulates the kinds of command team interactions typical of operational settings.

The finding that task automation impaired other non-automated elements of the performance tasks is consistent with the findings of our literature review that automation can disengage the operator from the task and subsequently negatively impact decision making. More specifically, the poorer task performance when automated assistance was provided could be due to some form of out-of-the-loop performance deficit (Endsley & Kiris, 1995). Out-of-the-loop performance deficits generally occur when the operator is faced with a failure in the automation, or some other abnormal system occurrence. In the current study, the automation did not fail, but CPA events requiring manual detection were less frequent with the provision of automation. By introducing automation, we further reduced the number of CPAs that participants had to detect manually by half. It is possible that participants in the automation condition allocated less attention to the CPA task. This interpretation is consistent with other research showing that participants returning to manual functioning from high level automation take a longer time to recover compared to individuals operating under lower levels of automation (Endsley & Kiris, 1995; Kaber, Omal, & Endsley, 2000; Manzey, Reichenbach, & Onnasch, 2012) – the greater the degree of automation, the harder it is to return to manual functioning.

Related to this, another explanation for the poor automation outcomes in the current study is that we 'fractionized' the CPA task. Vortac and Manning (1994) demonstrated that automation was less successful when it was only applied to a portion of a complex task. They argued that automation will improve performance and decrease workload only if it reduces the overall number of task modules that a participant has to complete. Given that CPA automation was only applied to a portion of the total

CPAs, it could be argued that this particular type of automation fractionated the task. However, although automating CPAs within a specific region of space did create an additional rule that the participants needed to remember, it did not add an extra task by breaking up a more complex one. This could explain why participants were poorer on the CPA task when they were assigned to the automation condition, but not worse on any other performance or workload measures. While the obvious answer to this issue would be to automate the entire CPA task, this would likely increase the out-of-the-loop performance problems even further (Endsley & Kiris, 1995).

The fact that not all participants used the automation, despite it being 100% reliable, is indicative that some participants may have been protecting themselves from these pitfalls of automation. Many factors affect whether an individual chooses to use the automated assistance provided in a task, and not just the reliability of the automation itself. Individual dispositional factors such as propensity to trust and extraversion have been linked to automation usage (Lee & Moray, 1992, 1994; Paul, Rovira, & Lee, 2011). With the current data there was a correlation between being an automation rejecter (two or more rejections) and gender, where females were more likely than males to reject the automation ($r = 0.133$, $p < 0.05$). It would certainly be of value to conduct future studies looking at individual differences in more detail. Several studies have also shown that trust in automation is linked with perceived reliability across time (Adams, 2007; Cassidy, 2009; Parasuraman, Molloy, & Singh, 1993). The participants in this study had two half hour practice scenarios before commencing the experimental conditions. It could be that more 'proof' of the reliability of CPA automation was required to convince some of the participants, and future research could provide this during training. This is not to say that being hesitant about trusting automation is always a negative trait. When an individual's trust in automation exceeds the reliability and robustness of the automation then it becomes problematic (Parasuraman et al., 1993; Parasuraman, Sheridan, & Wickens, 2008).

One must take care generalising the current findings to real naval operations given the use of the medium fidelity task and the student participant sample. This notwithstanding, we have taken a crucial further step towards demonstrating that SPAM can potentially be used to assess operator SA in naval submarine track management. In future research we will replicate these findings using naval personnel in simulations of Collins Class Submarine combat systems. Gaining an unobtrusive, real-time measure of SA is a priority for Defence researchers attempting to design new information handling systems and evaluate the utility of off-the-shelf technologies. Research in the submarine context is to date limited and the unique requirements of the submarine, together with the specific command team procedures in the submarine control room, need to be better understood so that emerging technology might be exploited effectively. It is expected, for example, that the volume and variety of sensor data may increase dramatically with technological change in coming years. How the data are translated and communicated to the human needs to be carefully managed, filtered, and processed to ensure peak operational performance and safety. The capability of the submarine will depend, as it always has, on the human achieving SA appropriate to the task at hand. Interface design must begin with identification of the dynamic information needs or SA requirements of the submarine operator.

# 5. References

Adams, J. A. (2007). Unmanned vehicle situation awareness. *Proceedings of HSIS 2007 ASNE Human Systems Integration Symposium, 19*, 21.

Cassidy, A. M. (2009). *Mental models, trust, and reliance: Exploring the effect of human perceptions on automation use.* Unpublished Masters, Naval Postgraduate School, Monterey, CA.

Durso, F. T., Bleckley, M. K., & Dattel, A. R. (2006). Does situation awareness add to the validity of cognitive tests? *Human Factors, 48*, 721-733.

Durso, F. T., & Dattel, A. R. (2004). SPAM: The real-time assessment of SA. In S. Banbury & S. Tremblay (Eds.), *A cognitive approach to situation awareness: Theory and application* (1 ed.). Great Britain: Ashgate Publishing.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., & Nikolic, D. (1999). Situation awareness as a predictor of performance in en route air traffic controllers. William J. Hughes Technical Center, Atlantic City International Airport, NJ: U.S. Department of Transportation, Federal Aviation Administration.

Ehret, B. D., Gray, W. D., & Kirschenbaum, S. S. (1997). Submariner situation assessment: A cognitive process analysis and modeling approach. *Proceedings of the Human Factors and Ergonomics Society, 1*, 163-167.

Ehret, B. D., Gray, W. D., & Kirschenbaum, S. S. (2000). Contending with complexity: Developing and using a scaled world in applied cognitive research. *Human Factors, 42*, 8-23.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37*, 381-394.

Gaba, D. M., Howard, S. K., & Small, S. D. (1995). Situation awareness in anesthesiology. *Human Factors, 37*, 20-31.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In A. H. Peter & M. Najmedin (Eds.), *Advances in Psychology* (Vol. Volume 52, pp. 139-183): North-Holland.

Jones, D. G., & Endsley, M. R. (2004). Use of real-time probes for measuring situation awareness. *The International Journal of Aviation Psychology, 14*, 343-367.

Kaber, D. B., Omal, E., & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing & Service Industries, 10*, 409-430.

Kirschenbaum, S. S. (2011). Expertise in the submarine domain: The impact of explicit display on the interpretation of uncertainty. In I. K. Mossier & U. M. Fischer (Eds.), *Informed by knowledge: expert performance in complex situations* (pp. 189-199): Routledge.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*, 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*, 153-184.

Loft, S., & Morrell, D. (2011). *Development of Situation Awareness Measurement Techniques - Report prepared for DSTO*: University of Western Australia.

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids. *Journal of Cognitive Engineering and Decision Making, 6*, 57-87.

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. *The International Journal of Aviation Psychology, 3*, 1-23.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making, 2*, 140-160.

Paul, C. L., Rovira, E., & Lee, J. (2011). Automation: Friend or foe? *Ergonomics in Design: The Quarterly of Human Factors Applications, 19*, 31-32.

Pierce, R. S. (2012). The Effect of SPAM Administration During a Dynamic Simulation. *Human Factors, 54*, 838-848.

Salmon, P., Stanton, N., Walker, G., & Green, D. (2006). Situation awareness measurement: A review of applicability for C4i environments. *Applied Ergonomics, 37*, 225-238.

Selcon, S. J., & Taylor, R. M. (1990). *Evaluation of the Situational Awareness Rating Technique (SART) as a tool for aircrew systems design* AGARD.

Stein, E. S. (1985). Air traffic controller workload: An examination of workload probe. Atlantic City International Airport, NJ: U.S: Department of Transportation, Federal Aviation Administration.

Vidulich, M., McCoy, A., & Crabtree, M. (1995). Attentional control and situational awareness in a complex air combat simulation. *Situation Awareness: Limitations and Enhancement in the Aviation Environment*, 18.11-18.15.

Vortac, O., & Manning, C. (1994). Modular automation: automating sub-tasks without disrupting task flow. *Human Performance in Automated Systems: Current Research and Trends, Mouloua M, Parasuraman R (eds). Erlbaum: Hillsdale, NJ*, 325-331.