12th International Command and Control Research and Technology Symposium
Adapting C2 to the 21st Century

Assessing Human Performance in a Distributed Virtual Battle Experiment

AUTHORS:
J. M. Crebolder
Defence Research & Development Canada
Dartmouth, Nova Scotia, Canada

Carol S. Cooper-Chapman
Defence Science and Technology Laboratory
Fareham, England

Mark G. Hazen
Defence Research & Development Canada[1]
Dartmouth, Nova Scotia, Canada

Colin Corbridge
Defence Science and Technology Laboratory
Fareham, England

Addresss communication to:

Jacquelyn Crebolder
DRDC Atlantic
Dartmouth, Nova Scotia
Canada, B2Y 3Z7
email: jacqui.crebolder@drdc-rddc.gc.ca
(902) 426-3100 ext.296

ABSTRACT
Human performance and decision making in distributed teams was examined in a multi-national virtual battle experiment conducted to investigate the impact of using an uninhabited aerial vehicle (UAV) to supply intelligence support to a maritime coalition defence force. In a synthetic simulated force defence operation, two allied frigates escorted two neutral high-value units through littoral waters with the threat of a swarm attack by small, fast inshore craft (FIAC). The ships were manned by navy command teams geographically located in their respective countries and nations were networked together for interactive play and collaboration. Performance was assessed in scenarios with and without the UAV. Subjective measures of workload and situation awareness of command team members, as well as within- and between-ship communication patterns were collected. Objective measures were also recorded and included number of leakers (vessels within range) and response time to classification. This paper reports on the subjective measures of performance and highlights lessons learned in conducting multi-national distributed experiments involving repeated trials. The study uses a synthetic environment coupled with live command teams in a netcentric operation to extend the findings of an operational research study in which command and control issues were identified in force defence against swarm attack.

INTRODUCTION

In today's world military and security authorities continually face the ubiquitous challenge of maintaining security and defence against global terrorism on land, sea, and in the air. In the maritime domain the new age of war necessitates a transition from traditional open-water defence strategies to those centered in coastal waters. The littoral arena is distinct from blue water, with constrained waterways where pleasure craft, fishing boats, and merchant shipping mix with military and maritime security vessels, and the proximity of nearby shipping, port and coastal facilities. As such, maritime transportation becomes highly vulnerable in the busy sea-lanes of coastal waters not only to terrorism directed at military warships, but to threats aimed at creating economic disruption through assault on land facilities, commercial shipping, such as oil tankers and cargo ships, cruise liners and ferries (Greenberg, Chalk, Willis, Khilko, & Oritz, 2006; Wood, 2005).

A significant and rising littoral threat are fast inshore attack craft (FIAC). These small boats often resemble pleasure craft or fishing boats allowing them to move indiscriminately amongst similar-looking vessels. FIAC are also highly maneuverable and, coupled with swarming tactics where a large number simultaneously assault a target, they pose a substantial and ever-increasing threat in the maritime environment (Galligan, Galdorisi, & Marsland, 2005). Their effect has been seen in recent incidents, including the single small boat attack on the navy destroyer USS Cole in 2000, attack on the French oil supertanker Limburg in 2002 and a swarm offence by six small powerboats on a US Military Sealift Command tanker in the Persian Gulf in 2002 (Clarke, 2005).

By its very nature asymmetric threat is covert. Consequently, military and security vessels are frequently in a state of force defence while traveling through populated waterways where the possibility of hostile intent exists but may be well hidden. In homogenous surroundings like this the enemy is difficult, if not impossible to detect. Furthermore, within the time and space constraints of the littoral environment, detection

and classification of the enemy leaves little time to react and engage an appropriate response.

Any tactic or strategy that can increase the time available for determining and responding to hostile intent, particularly from asymmetric threat, would be of benefit in the maritime force protection domain. One approach is to increase situational knowledge of the surrounding environment through sharing of information across available sources. According to the (US) National Strategy for Maritime Security (Department of Defense and Homeland Security, 2005), the key to establishing a heightened level of defence capability can only be attained through interoperability. Netcentric liaison, connecting multiple resources at the operational and tactical levels, such as air, ship, and land units, linked together by an information network, is seen as one of the most powerful ways of increasing maritime domain awareness. However, the key to the usefulness of information supporting situation awareness is its timely arrival (Galligan et al., 2005). Intelligence, provided through surveillance and reconnaissance platforms for example, must arrive in time to input decision making and make an effective response.

Findings in a recent operational research study investigating the effect of communication in a force defence scenario indicated that low levels of information sharing led to significant risk and that, even with a shared operating picture, information must arrive in a timely manner to be of use against the swiftness of a swarm attack (Galligan et a., 2005). Using a spiral concept development process in which high-level modeling is followed by more detailed study, the intention of the present study was to investigate the impact of shared information on force defence by expanding the operational research work. As such, a synthetic environment was used, human operators were added to carry out the mission, and more detail was introduced into some of the sensor and command and control elements of the operational research simulation. Shared information, in the form of surveillance intelligence, was supplied through an unmanned air vehicle (UAV) flying overhead. The operational research work found that information provided by a surveillance UAV was instrumental in moving the battle space outward, giving the ship more time to evaluate and prepare an appropriate and effective response. Critical to the success of using the UAV was the networking and communication capability between it and the controlling ship (Galligan et al., 2005).

The scenario used in the current study involved two allied ships escorting two high value units (HVU) through a narrow strait busy with commercial and pleasure traffic. Each allied ship was controlled tactically by a command team from a different nation. To simulate the realism of distributed teams taking part in a coalition mission a synthetic environment was assembled in which maritime platforms were physically located in each individual nation. Therefore, as is typical of a real world coalition scenario, the interaction between command teams was through networking and communication systems. A secondary advantage of the geographically distributed nature of this study was that it allowed for multiple runs to be conducted with limited expense and investment of resources. By conducting multiple runs statistical sensitivity to the variables of interest would be maximized.

We expected that the surveillance information provided by the UAV would increase the level of situation awareness available to the coalition team. Consequently, surveillance intelligence would aid in decision making, reduce time to detection, and

increase the overall effectiveness of the mission. For comparison the study included two scenario types, one with the UAV available and one without the UAV.

In each scenario run (a run being 1 experimental trial) the Commanding Officer (CO) of one of the allied ships was designated as Officer of Tactical Command (OTC). The OTC directs the mission and has the authority to command other ships in the coalition task force. Since a team's perception of performance might differ depending on whether or not their ship was OTC we also included OTC as a variable of interest.

This study is relatively unique in that it involved multiple experimental trials using geographically distributed teams, a synthetic environment, and encrypted communication and networking. Finding little existing methodology that could be used for guidance in assessing human performance in this type of distributed simulation one objective of the study was to evaluate the data collection methods used as far as their ability to support human in the loop testing. Results and lessons learned will assist in further development of appropriate methods and metrics for future maritime-based empirical evaluations using distributed teams and synthetic environments.

Objective measures of performance (e.g., number of hostile FIAC within attack range; time from target detection to classification as hostile) were recorded throughout each scenario run (for details see Hazen, Jones, Ping, Macferson, & Kuster, 2007). Objective measures are useful in that they can be compared to ground truth, in this case events generated by the simulation, but they do not tell us much about conditions and cognitive processes leading up to final outcomes. Consequently, measures that might aid in understanding why results occurred and how underlying cognitive processes unfolded were also included. These measures of performance focused on team workload, situational awareness (SA), and interactions between team members.

Since workload may differ depending on the presence, and use, of the UAV, subjective estimates of workload were collected from each team member for Base and UAV scenarios and when the CO of each ship was OTC and when not OTC. Continuous workload measures were taken throughout the experiment, and a NASA TLX (Hart & Staveland, 1998) assessment was taken at the end of each run. Since the presence of the UAV provided an additional source of situational knowledge the Crew Awareness Rating Scale (McGuinness & Foy, 2000) was conducted after each run to provide a measure of each team member's perception of their own SA.

Decision making is, in part, related to situation awareness and attaining situation awareness can be correlated to the degree of interaction and communication between individuals as they share information. Consequently, team communication was recorded to examine interaction patterns. The work by Entin and Entin (2001) was used as a guide and frequency of requests, transfers and acknowledgements between team members was recorded.

METHOD

Scenario description:

Two allied warships are escorting two HVUs through a confined strait of water in force defence operation. The two HVUs will fall in behind the lead ship with the other allied ship pulling up the stern to form a convoy (as shown in Figure 1).

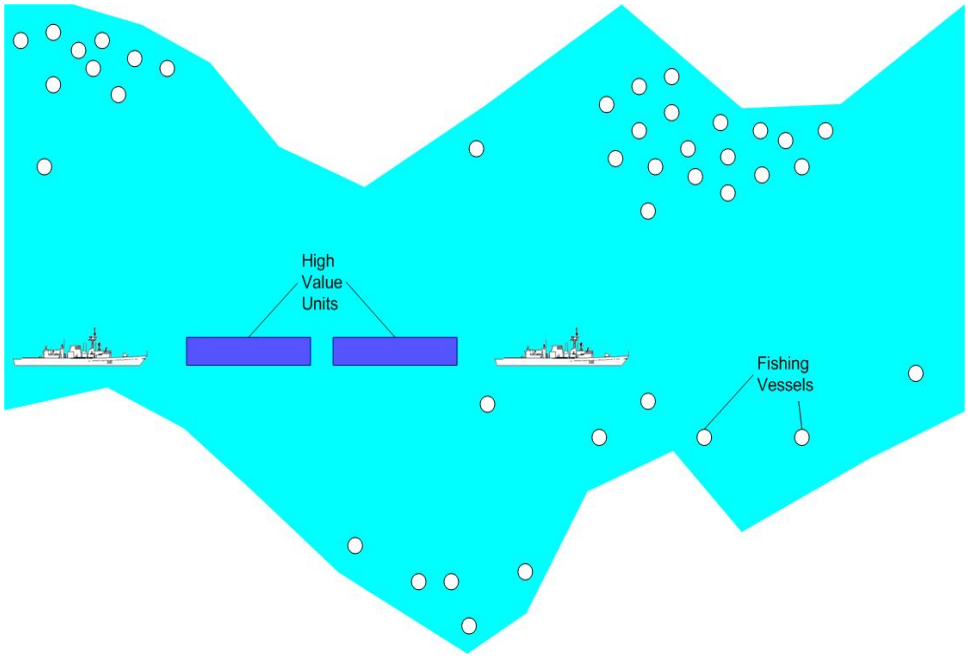


Figure 1. Scenario Description

There are numerous fishing vessels and pleasure craft in the area as well as merchant and commercial shipping. Intelligence reports indicate a possible threat from terrorists in the vicinity. As the convoy transits the strait, a swarm of fast inshore attack craft will form up and attack one of the HVUs.

An attack never took place in the first 30 minutes of a run and attackers and neutral craft were indistinguishable prior to initiation of the attack. Upon attack initiation, FIAC behaviour changed so that they converged at high speed on one of the HVU, and the appearance of persons onboard changed from civilian to persons carrying guns or rocket propelled grenades. Any craft closing on the convoy could be warned off by gun fire, flares or radio communication. Table 1 provides detail of the weaponry and defense capabilities of the platforms.

Blue Forces		**Red Forces**
Frigates	**HVUs**	**FIAC (8-10 small craft)**
No organic air asset 76 mm gun .50 cal machine gun	No defence capability Could maneuver Assumed encrypted communications could be received	Rocket propelled grenades (RPG) 7.62 mm automatic weapons

Table 1. Weapons and Defense Capability Modeled

Each UAV scenario was matched with an identical Base scenario where only ships' sensors and visual data from upper deck sentries provided data on the evolving situation. The UAV and Base scenarios made 1 scenario pairing. To minimize effects of

familiarity across runs onset of attack and number of pleasure craft, fishing fleets, and cross strait traffic in the vicinity was varied between scenario pairings. Each scenario pairing was conducted twice so that each nation ran in each scenario once as OTC and once not as OTC, equaling 4 runs per scenario. The OTC always controlled the UAV.

Participants:
Nations - The study was a collaborative effort involving five nations: United Kingdom (UK), Australia (AS), Canada (CA), United States (US), and New Zealand (NZ). AS, CA, and UK participated directly in the scenario runs.

Positions - Each nation provided the following personnel for each run :

- CO - Commanding Officer (possibly OTC) - Responsible for leading his ship's team, tactical and mission related decision making, directing coalition when OTC.
- ORO - Operations Room Officer - In charge of all systems operated from the Operations Room, responsible for collating information from all sources into a global picture, communicating state to CO.
- SWC - Sensor Weapons Controller - Builds tactical picture, coordinates all surface weapons and sensors controlled from the Operations Room,
- FPO - Force Protection Officer – Builds tactical picture, communicates with sentries, coordinates .50 cal machine gun operators.

These personnel formed one command team for each country. The AS team was a fully worked up team from a single ship. The CO was replaced by a substitute CO in the third scenario run but he participated in all other runs. The UK command team had not worked together and on some runs civilians filled positions when Royal Navy personnel were unavailable. CA's command team had worked together before, with the exception of the FPO who was relatively new to the Canadian Navy and the replacement SWC in the second week of runs.

- Observers - 4 personnel assigned one on one to each position, plus 1 lead observer - Collected human factors and military observations during the runs.

CA's observers were military personnel and defence scientists. The lead observer was a naval training analyst. All observers in the UK were human factors personnel including the lead observer. In AS either defence scientists or human factors personnel were observers although for some scenario runs the lead observer was a military officer.

- UAV Operator - flies UAV using Microsoft Flight Simulator® and provides voice reports to the controlling frigate's FPO. Communicates with ORO of the OTC ship.
- Sentry Interactor – simulates six upper deck sentries and provides voice reports to FPO.

A number of additional personnel were required to conduct the experiment and monitor and control technical aspects. For a full list of support staff refer to Hazen et al. (2007).

Equipment:
Figure 2 shows the typical room layout.

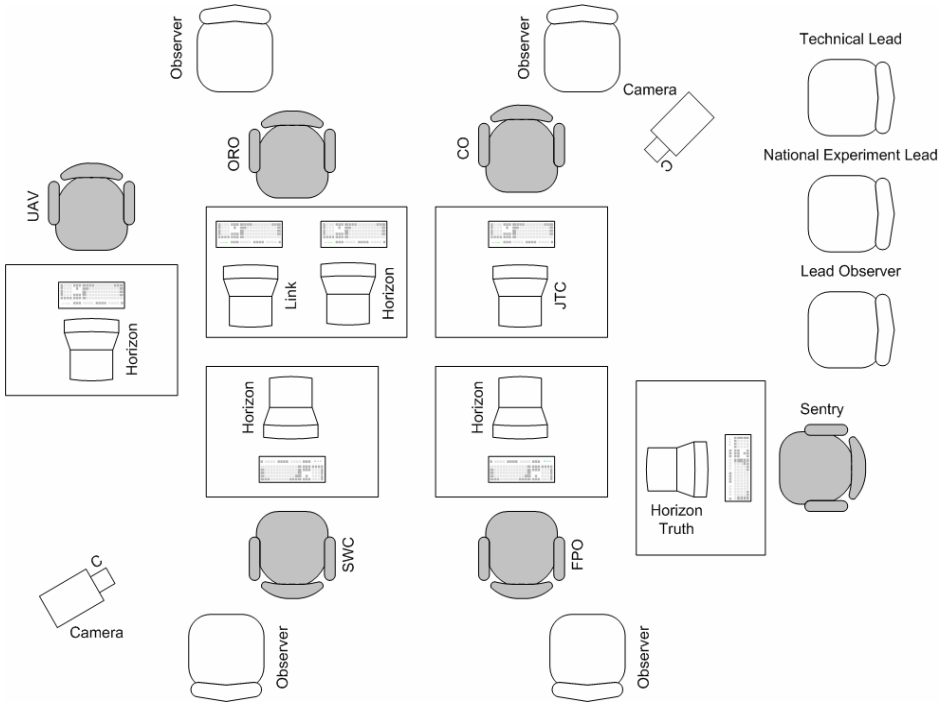


Figure 2. Room layout

Equipment included: Headphone/microphone sets for 2 participants (CO and ORO) as well as their observers; IP Phone Link for execution control and technical discussions; a CFBLNet encrypted network connection; Audio and Video recording equipment; Personal Digital Assistant or equivalent for each observer (4 per team), compiled with Interaction Recorder® software (DSTO, 2005) (for a full list and details on technical equipment refer to Hazen et al., 2007). Observer Pre-scenario Questionnaire package (paper); Observer Post-scenario Questionnaire package (paper); Participant Pre-scenario Questionnaire package (on-line); Participant Post-scenario Questionnaire package (on-line); two versions of the NASA TLX (Hart & Staveland, 1988), both in electronic form (DRDC, 2002).

Pilot Study:
Six weeks prior to the experiment a pilot study was conducted using civilian participants and observers from CA and UK. The primary objective of the pilot work was to test and troubleshoot technical equipment, networking, and procedures.

Training:
Officers and observers completed a one week training period to familiarize with roles, tools and displays, and experimental and data collection procedures. The rationale behind the various human factors measures and data collection techniques was explained to participants and observers. During this week the technical team ensured networking and other aspects of the distributed simulation were in place and functioning appropriately.

Experimental Runs:

Experimental runs were conducted over a two week period and were scheduled as closely as possible to cover normal working hours in each country. Two nations participated in each run, each controlling one of the two allied ships. One CO was assigned as OTC on each run and the OTC ship always had control of the UAV. During the first week 5 successful and 3 partial runs were conducted between pairs but a number of technical problems resulted in insufficient data at the end of the week. It was decided at that point to make the best use of our naval personnel and abandon some of the networking issues by relocating CA's team to Sydney, Australia. In Australia the two teams (AS and CA) were in separate rooms. Ten runs between AS and CA were successfully completed.

PROCEDURE

According to each nation's ethical protocol participants were required to read and sign a voluntary consent form prior to the start of the experiment. Biographical information was collected from participants and observers at this time.

Pre-run preparation:

The OTC CO was briefed on the mission and participants were seated at their stations. Typically the CO then briefed his crew and CO on the allied ship as to his proposed strategy. The crews conducted watch handover in which all traffic currently visible on the operations room displays were identified and marked. This procedure primarily involved the FPO, SWC and Sentry Interactor.

Observers were each provided with a PDA, or laptop, and seated next to their assigned participant. Their primary task was to record all communications, both incoming and outgoing, involving the position they were assigned to observe. Figure 3 shows the PDA interface used for collecting communication data.

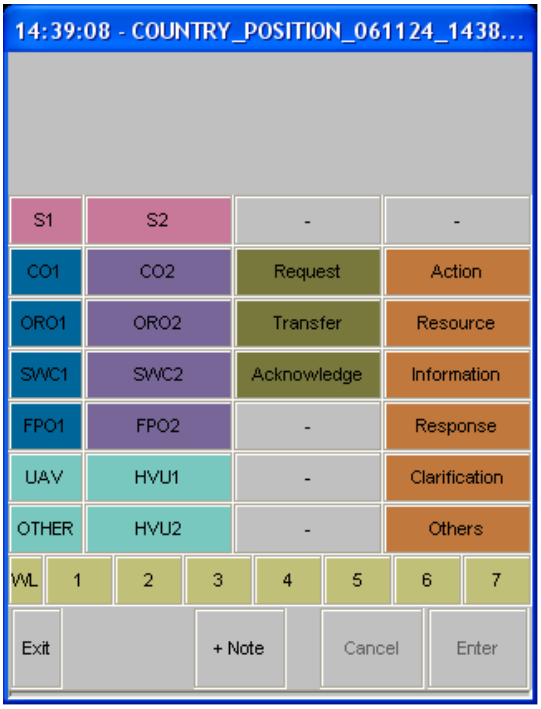


Figure 3. PDA interface layout

The PDA was also used to collect subjective workload estimates from participants every five minutes throughout each scenario run. Responses were collected verbally initiated by a verbal request "Workload" from the lead observer. Response scale was 1 through 7, 1 being lowest (can easily complete tasks) and 7 highest (cannot take on another task).

Observers also collected pencil and paper notes on their participant's behavioural state (e.g., bored; busy) as well as significant events that occurred during the scenario (e.g., hostile target marked; equipment problem).

Scenario run:
Nations were completely networked together for interactive play and collaboration and displays were updated continuously across nations according to input from the teams. COs were connected by Chat (BuddySpace®), and COs and OROs on each ship were networked together over headphones (via TeamSpeak® software). Observers for the CO and ORO were able to listen in using headphones. Continuous recording of display screenshots, BuddySpace and TeamSpeak communications, and overall room video and audio, were recorded for post-scenario analysis.

Post-scenarios:

Termination of a scenario occurred when the enemy had successfully damaged one of the HVUs; when all hostile FIAC had been neutralized; or when technical difficulties interfered with the simulation to the point that is was affecting team performance.

At the conclusion of each scenario several performance measures were collected. Subjective estimates of workload were collected through two electronic versions of the NASA TLX (Hart & Staveland, 1988). In the first, participants estimated their own workload based on six performance-related dimensions (physical demand; mental demand; temporal demand; effort; frustration; and performance). Following the estimate of individual workload a modified version of the NASA TLX was completed in which each participant estimated the workload of the whole team based on the same dimensions. Following the NASA TLX an on-line post-scenario questionnaire was completed by each participant. The questionnaire included an assessment of each team member's perceived SA through the Crew Awareness Rating Scale (McGuinness & Foy, 2000). The CARS considers four aspects of SA, the first three of which are based on Endsley's model (Endsley, 1995): a) Perception – the detection, acquisition and assimilation of information available; b) Comprehension – the understanding and interpretation of information in terms of known schemas; c) Projection – the anticipation and insight into how a situation is likely to unfold; d) Intention – the understanding of what courses of action are available and which is optimal in the current situation. CARS also differentiates between perceived actual knowledge, referred to as cognitive content, as well as the cognitive processes that underlie attaining, maintaining, and using that knowledge, referred to as process. The assessment is subjective and it reflects an individual's confidence in their current level of awareness and the ease with which they feel they could use the information in the context of the current situation. Questions were designed to tap into each of the 4 aspects at each of the 2 levels, and responses were on a scale of 1 to 4, 1 reflecting highest perceived SA. Also included in the post-scenario

questionnaire were questions related to team shared awareness, workload, and scenario realism.

Observers were also required to complete a short post-scenario questionnaire in which they estimated their participant's workload, CO's leadership style, and provided comments on the tools they had used for collecting data (PDA, behavioural data sheet).

RESULTS

Fifteen runs were fully completed during the course of two weeks. Of these, 7 runs were removed from the data set either because technical difficulties may have affected performance (e.g., slow transmission rate, problems with weapons assignment) or because data was missing (e.g., participant did not complete questionnaire). The remaining 8 runs resulted in 4 sets each made up of a UAV scenario and a matching Base scenario, and on which each CO had acted as OTC on 1 UAV and 1 Base. Because of the small number of runs in the final data set statistical analyses were not possible but data patterns were examined for general trends in performance.

Although several measures were collected during the experiment only workload, perceived SA, and communication patterns are reported here. Others are discussed in the Discussion Section with respect to level of usefulness and validity.

Overall, the data showed that CA and AS differed widely in their subjective estimates of workload and perceived situation awareness, and in overall variability within each of these measures. For this reason, rather than collapsing across countries, the data are presented for each nation separately and possible reasons for differences between countries are offered in the Discussion Section.

Workload:

a) Continuous workload:

Subjective estimates of workload were collected from participants every five minutes and recorded by observers. Figure 4 shows mean workload estimates on a scale of 1 ('can easily complete tasks') to 7 ('cannot take on another task') for each nation as a function of Position(CO; ORO; SWC; FPO)/Scenario type (UAV or Base) and OTC (OTC or not OTC).

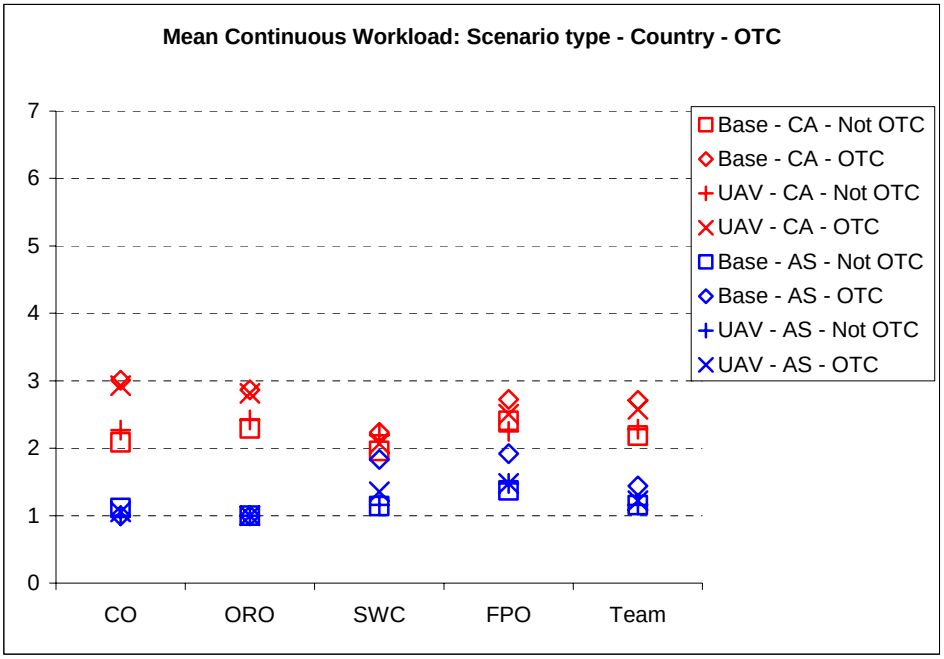


Figure 4. Mean subjective estimate of workload

In general, members of both teams felt they experienced more workload when their CO was OTC. This was particularly true for the CO and ORO of the CA team, although their AS counterparts reported little difference in workload based on being OTC, or as a function of the presence or absence of the UAV. Note that overall AS rated their workload as very low and always lower than CA. AS's CO and ORO show no variability in their estimates of workload. In a review of a videotape of one of the last scenarios the AS CO asked how often a response to workload would be requested. It is unclear why a procedural question like this would be posed at this point in the experiment since several runs had already been conducted, but it suggests that the purpose and procedure of this workload assessment was not clear.

b) NASA TLX:

Two versions of the NASA TLX were completed – one in which a subjective estimate of each individual's workload was collected and the other allowed each participant to provide an estimate of the team as a whole. The former is reported here.

NASA TLX estimate of Individual workload

Figure 5 shows the overall workload scores for each position and each country as a function of the presence or absence of the UAV (UAV/Base) and whether or not the team's CO was OTC (OTC/notOTC).

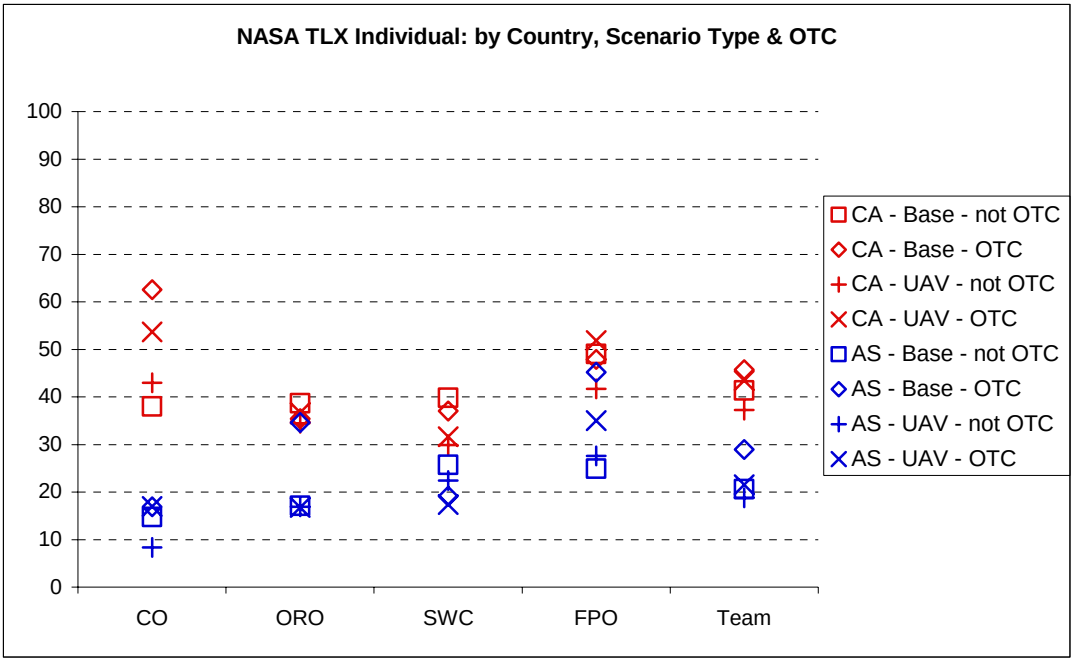
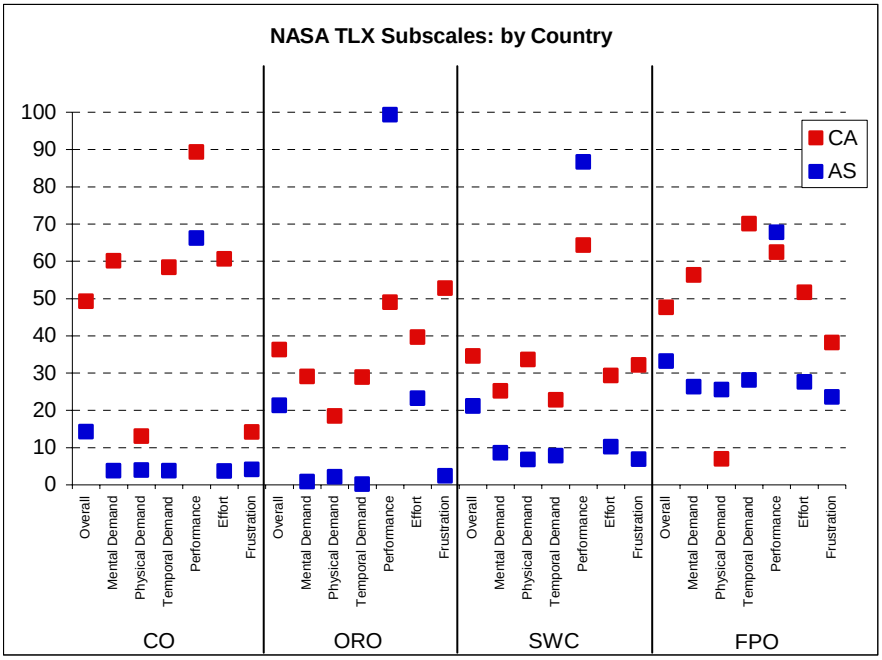


Figure 5. Mean NASA TLX overall scores for Individual

Although the data is variable the general trend appears to be that being the OTC ship was perceived to create more workload. CA's CO and AS's ORO and FPO reported that being OTC and having no UAV produced highest workload, and virtually all positions and nations agreed that more work was involved when the UAV was not available. Like the continuous workload measure shown in Figure 4 AS consistently reported lower workload than CA.

Figures 6 and 7 depict 2 views of the NASA TLX dimension of physical, mental, and temporal demand, effort, frustration, and performance. Figure 6 shows the mean scores for each position and country averaged across Scenario Type and OTC, while Figure 7 shows the data for Scenario Type and OTC averaged across Position.



Mean NASA TLX subscale scores for each Country

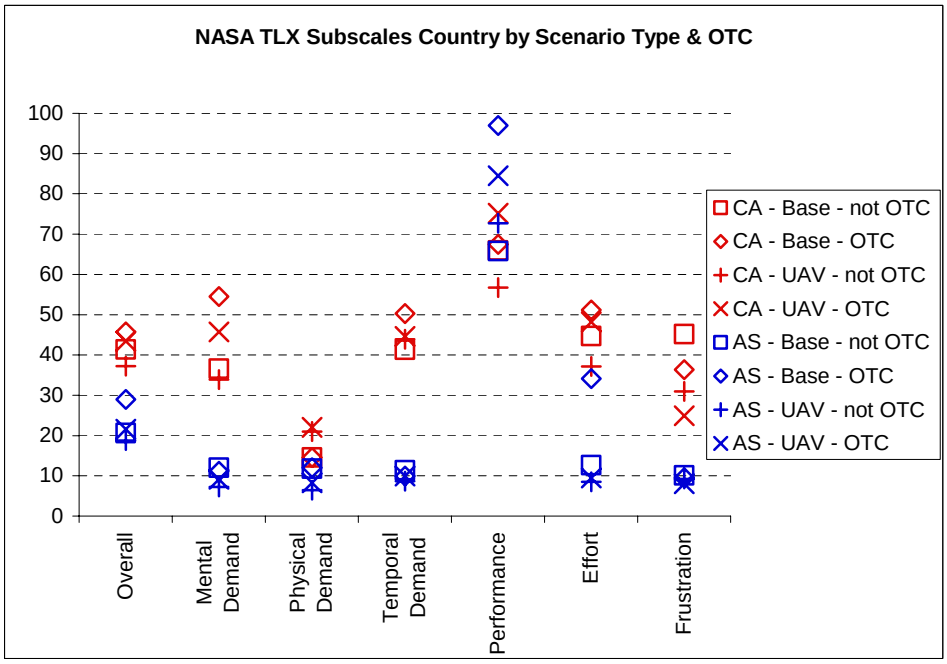


Figure 7. Mean NASA TLX Overall subscale scores for each Country, Scenario Type and OTC

As seen in both graphs AS's scores are generally very low with little variability and, as before, are lower than those reported by CA. AS consistently rated their overall

performance much higher than any of the other 5 dimensions. Although one might expect physical demand to be rated lower than other demands since the officers were seated and motor activity included only minor arm and hand movements, their ratings of physical demand are very similar to scores on the majority of the other subscales. CA on the other hand shows a fairly dramatic drop in scores on physical activity, with the exception of the SWC. Considering the simulated environment and the likelihood that the scenario did not entail physical, psychological, and mental components matching real life, CA workload data reflects what might be expected – low physical demand with intermediate levels of mental and temporal demand, effort and frustration.

Situational Awareness:

Mean scores from the Crew Awareness Rating Scale for each position as a function of Scenario Type and OTC are shown in Figure 8. A score of 1 reflects highest perceived SA, 4 lowest SA.

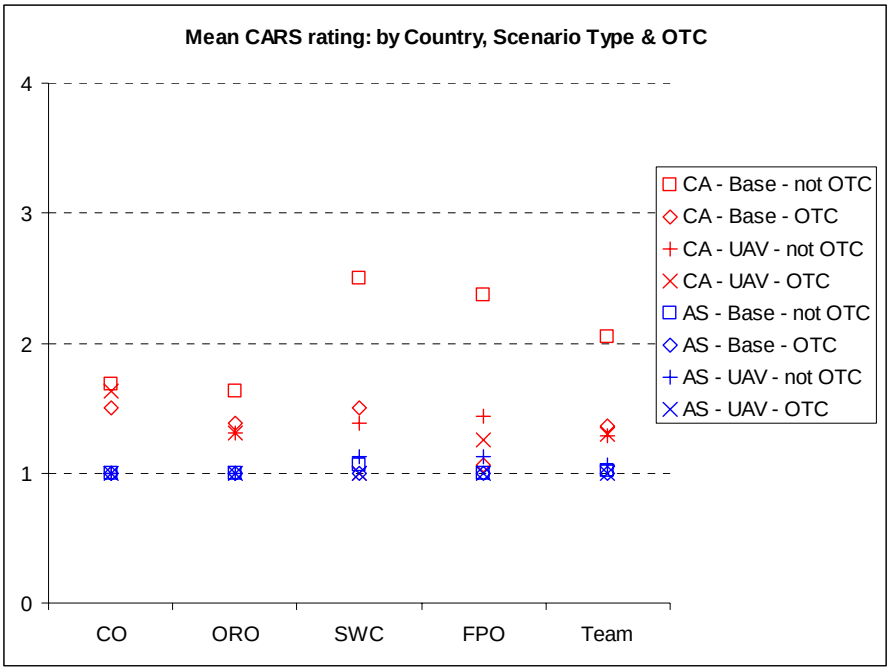


Figure 8. Mean CARS overall scores

Being OTC or having the UAV available had little effect on AS's perceived level of SA according to their responses on this measure. CA reported significantly lower SA when the UAV was not available and their CO was not OTC and in general their scores supported having the UAV for increasing SA.

The CARS scores broken down into aspects of perception, comprehension, projection, and intention are shown in Figure 9. The graph is divided into components of content and process – content indicating perceived knowledge based on information available, and process scores reflecting perceived ease in processing that information.

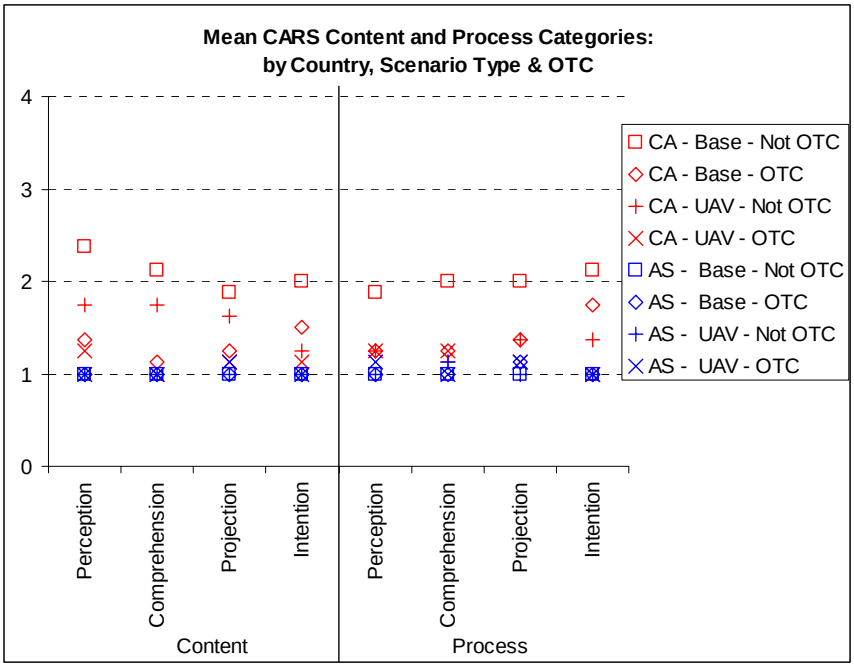


Figure 9. Mean CARS Content and Process scores

The content perception scores for CA suggest that the team felt they were least aware of relevant information when their CO was not OTC and when there was no UAV available to the coalition. If they did not perceive that information was available it is not surprising that scores on subsequent SA aspects of comprehension, projection and intention in this condition are also low. Associated process scores follow the same pattern, again not surprisingly. With the addition of the UAV, CA's perceived SA increased for all aspects of SA.

When CA was OTC their perceived SA was slightly higher when the UAV was available, with an almost linear pattern across all aspects of SA as far as both perceived reliability of information achieved (content) and the ease with which that information could be processed (process).

Communication Frequency:

Frequency and type of requests, transfers, and acknowledgements were recorded by observers throughout each scenario run. Interactions included those between ownship team members as well as to and from the other allied ship's team, sentries and HVUs. Examination of the data has been limited to frequency of out-going interactions only. Reasons for reducing the data in this way are covered in the Discussion Section. The data in Figures 10 -16 have been collapsed across Scenario Type since any changes in frequency of out-going communications were expected to be primarily a function of the OTC variable.

Overall the CA team interacted about twice as often as the AS team. Given the large difference between the number of interactions the scale for the CA plots allows up to 80 interactions for any position, whereas the AS plots allows for 40 interactions. The AS data have been plots are larger for ease of viewing individual data points.

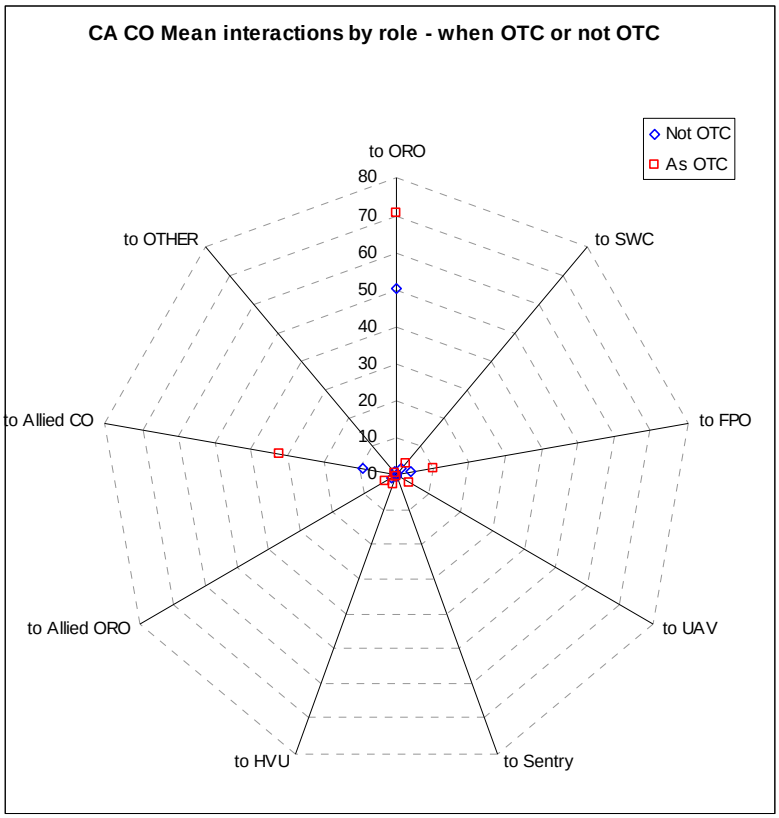


Figure 10. Mean Verbal Interactions of CA CO

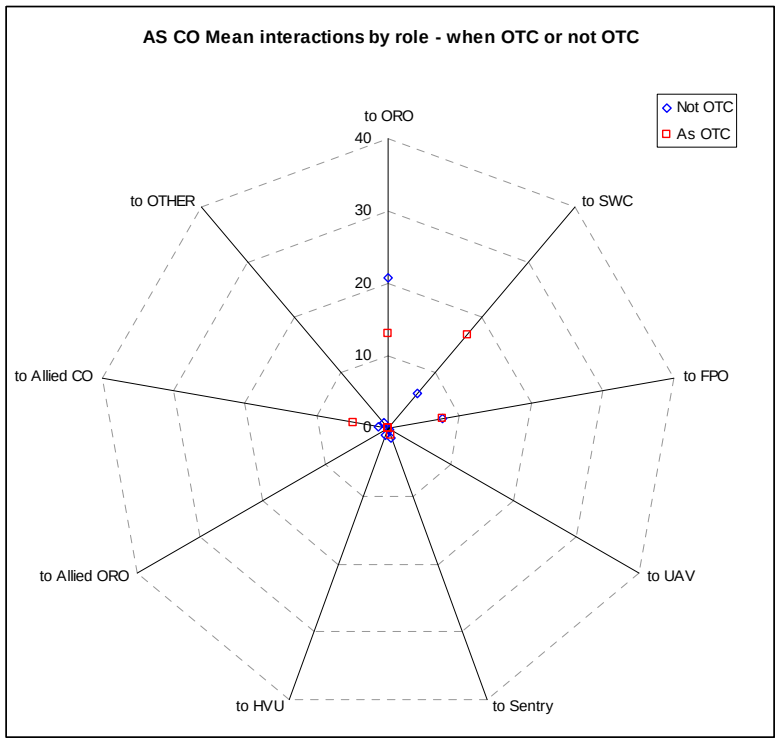


Figure 11. Mean Verbal Interactions of AS CO

Figures 10 and 11 show out-going verbal interactions for the CA CO and AS CO respectively. For CA's CO the number of out-going interactions increased when acting as OTC. This is most notable for interactions with his own ORO and FPO, and with the allied CO. For AS's CO the number of out-going interactions when OTC is not significant. However, who the AS CO interacts with changes when he is not OTC. When not acting as OTC interactions are primarily with his ownship ORO, with a few interactions between his SWC and FPO. On the other hand, when acting as OTC the main focus of his interaction is with his ownship SWC, and there are fewer interactions with his ORO. When he was OTC verbal interactions to the allied ship are low compared to those of CA's CO.

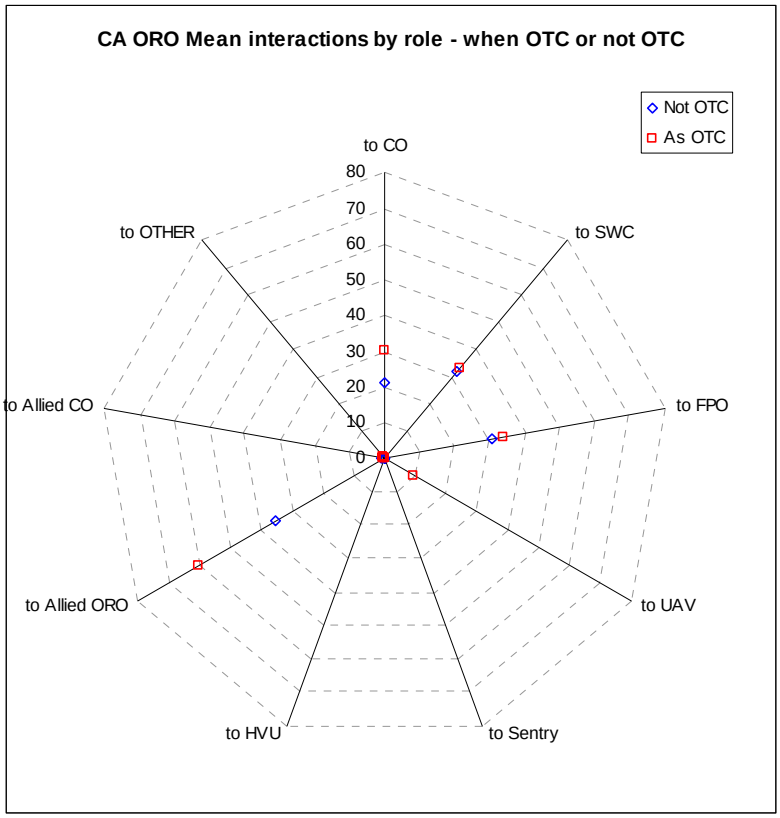


Figure 12. Mean Verbal Interactions of CA ORO

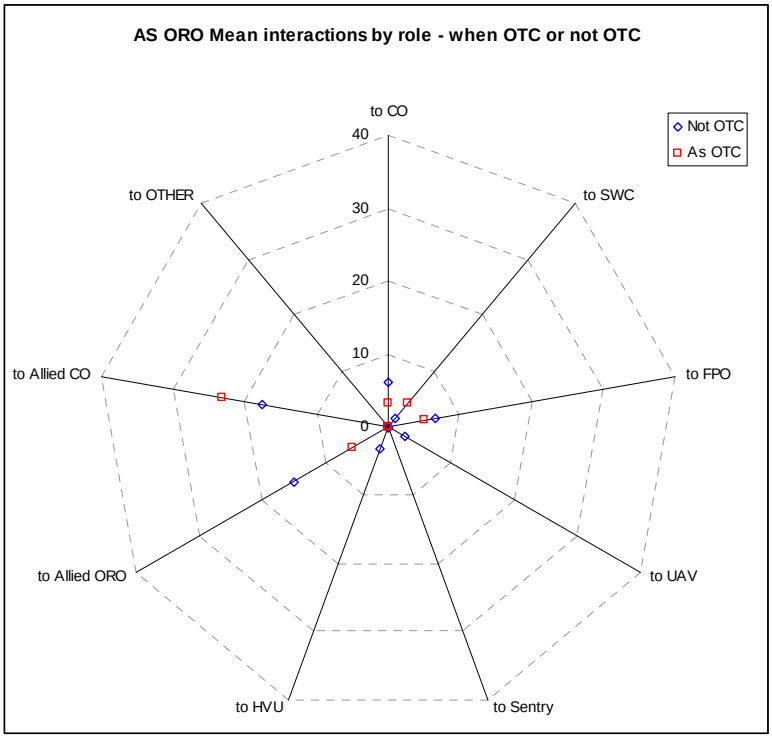


Figure 13. Mean Verbal Interactions of AS ORO

Like his CO, CA's ORO initiates more out-going interactions when his ship is OTC, particularly to his CO and to the allied ship's ORO as shown in Figure 12. The number of interactions he initiates to his ownship SWC and FPO change very little between being OTC and not OTC. Figure 13 shows the interactions for AS's ORO. Overall the ORO initiates few interactions within his own team but he interacts frequently with the allied ORO and CO. Note that the AS ORO communicates more with the allied CO than with his ownship CO. Whether or not this is true or an anomaly of how the data was recorded on the PDA is unclear. Further analysis of the video tapes is required to provide insight.

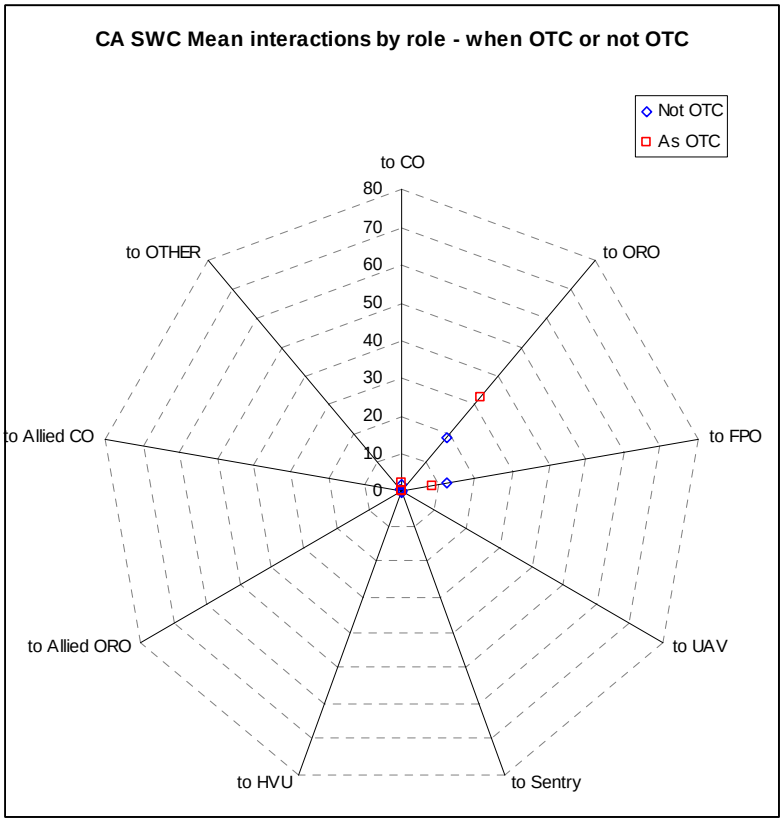


Figure 14. Mean Verbal Interactions of CA SWC

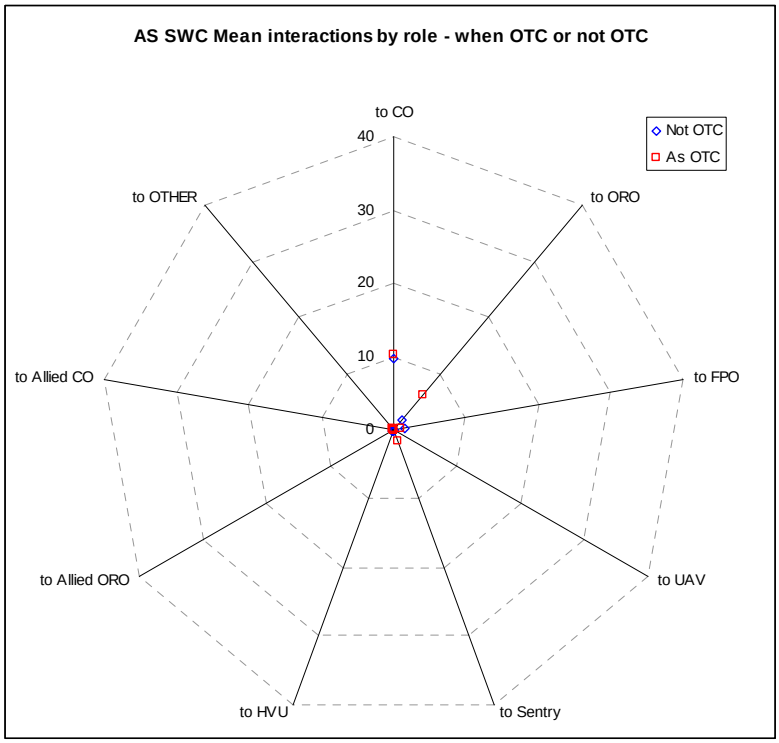


Figure 15. Mean Verbal Interactions of AS SWC

Figures 14 and 15 show out-going interactions for CA and AS's SWCs. CA's out-going interactions appear to be primarily with his ORO, and these increase when his CO is OTC. He interacts to a lesser extent with ownship FPO and little difference is seen whether or not his CO is OTC. Like CA's SWC the AS SWC initiates few out-going interactions. Little difference is seen in the number of interactions when his CO is OTC compared to when he is not OTC, although a greater number of interactions with his ORO is notable when his CO is OTC. However, the pattern of the AS SWC's interactions differs to that of CA, with his interactions being primarily with his CO.

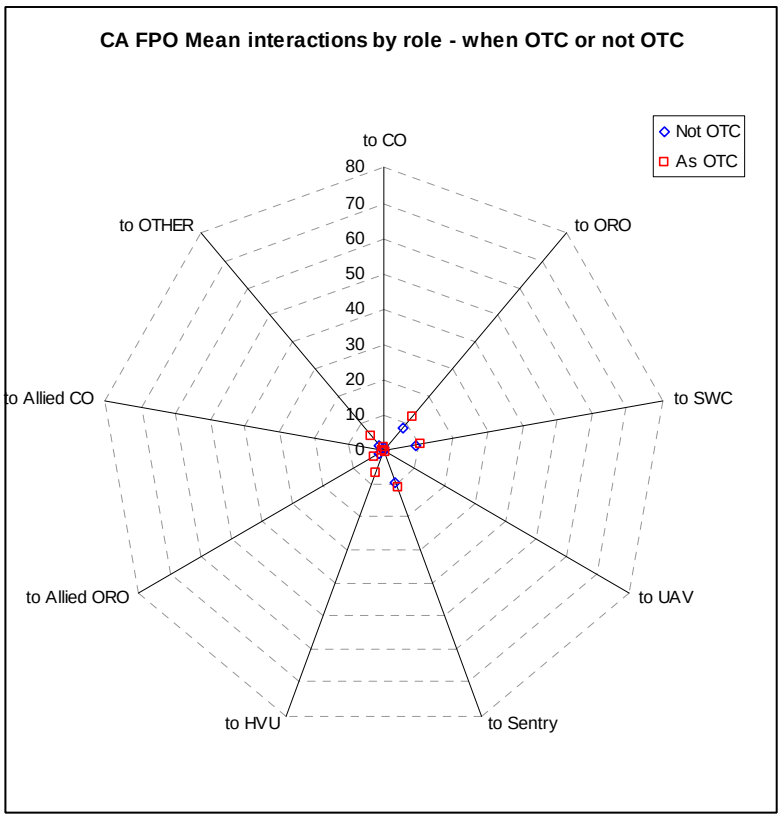


Figure 16. Mean Verbal Interactions of CA FPO

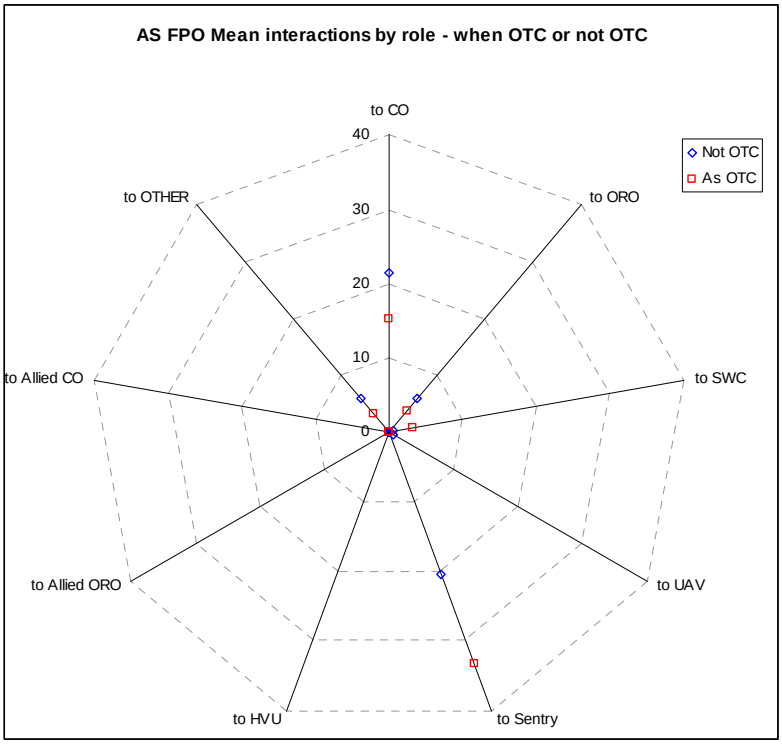


Figure 17. Mean Verbal Interactions of AS FPO

CA's FPO also initiates few interactions and those he does initiate tend to be primarily with the Sentry, ORO and SWC. Little difference is observed between the conditions where his CO is OTC and not OTC. The AS FPO initiates slightly more out-going interactions than his CA counterpart, but more notably, there is a very different pattern to his interactions. Whilst the CA FPO very rarely initiates an interaction with his ownship CO, the AS FPO frequently initiates this type of interaction. The AS FPO interacts more with his CO when his CO is not OTC, but most frequently with the Sentry. (Also notable is the AS Sentry who initiates more interactions than the CA Sentry, means 51 and 23 respectively.)

DISCUSSION

From the large amount of data collected workload, perceived SA, and team interaction were chosen for closer inspection. In general, the findings support the operational research study (Galligan et al., 2005) in that intelligence information supplied by a UAV improved performance in this force defence scenario. Perceived SA was highest for the OTC ship when the UAV was available, at lowest for the Canadian team. Additionally, officers reported significantly lower SA when their ship was not OTC and the UAV was not available. This observation is perhaps not surprising since the OTC ship had control of the mission in general and controlled the UAV. Information shared through communication might have increased SA of the non-OTC team and the data does suggest that, when OTC, both countries COs and CA's ORO increased interactions with the other ship.

Being OTC also increased workload, more so for the CO and ORO but also for the entire team, and workload was highest in the absence of the UAV. The fact that subjective estimates of workload increased when the UAV was not available is interesting because it points to a legitimate benefit to having UAV technology. Technology that is not useful can, in fact, increase a user's workload. When the UAV was present the OTC team was responsible for controlling, monitoring, interpreting, and communicating information from it - no doubt creating a substantial amount of additional work. Therefore, had the UAV not been perceived as a significant advantage reports of increased workload when it was present might have been expected, at least from the OTC team. Furthermore, post-scenario comments captured in the questionnaire clearly voiced support for having the UAV.

CA's responses form a predicted pattern on the NASA TLX subscales, scoring low on the physical dimension, with intermediate scores for the mental and temporal components, as well as for effort. Some positions also reported frustration, possibly in response to technical difficulties that continued to occur to some extent. The exception to scoring low on the physical dimension for the CA team was the SWC whose scores were relatively high in comparison to his teammates. One well-established concern with subjective responding is that individual interpretation may differ, leading to bias in an individual's ratings. The SWC's higher score may simply be a reflection of a different criterion level to which internal states are compared. Individual differences must be considered when interpreting these results since all of the human factors measures in this study rely, to a degree, on subjective assessment either from participants or observers, and the data collected from a very small pool of participants.

The interaction data collected by observers was analysed for frequency of out-going communications from each position's standpoint. Not surprisingly, both teams increased communication with the other allied ship when they were OTC, at least for the CO and ORO positions. Some interesting differences were observed between the two nations. The CO and ORO of the CA team communicated more when their ship was OTC than when not, whereas their counterparts on the AS ship generally did not increase communication frequency but tended to redirect the focus of their communication. For example, the AS CO interacted more with his SWC when he was OTC and less with his ORO, but more with the ORO when he was not OTC. Differences in team communication patterns like these could reflect naval procedural differences or a willingness to depart from standard procedure in this simulated environment where the configuration of the workstations was perhaps conducive to direct communication.

Underlying the analyses is the consistent observation that both teams differed significantly in their subjective scores and in score variability. AS's estimate of workload was always lower than CA, and was frequently very low with little or no variability, regardless of whether the UAV was present or whether or not they were OTC. The general feeling that workload was relatively low for the AS team was supported by responses in the post-scenario questionnaire. For example, in response to the question "What is your estimate of your overall workload over the whole scenario?", AS gave an average score of 3 on a scale of 1 to 10 (1 being lowest) whereas CA's mean score was 6. However, when asked how close the overall level of workload was when compared to workload experienced during a 'real-life' mission, there was relatively little difference between countries scores (CA mean = 5; AS mean = 4). It is not clear why lower

estimates of workload were observed but one explanation might be that the AS team subjectively experienced less workload because they were a well worked up team, having worked together prior to participating in this study, whereas at least two CA members were new to the team. Although this might explain why the AS workload scores were lower than CA it does not shed light on why they reported no, or little difference in workload or SA for the different UAV and OTC conditions.

As already mentioned, the small number of participants in this experiment and the minimal number of runs that we could include in the data set are of obvious concern Thus, due to a lack of variability in the data, no persuasive conclusions can be drawn nor generalizations made. Despite this, or perhaps because of it, the study offers an important contribution with respect to the work's objective centered around evaluating the human factors measures constructs and their ability to support human in the loop testing in this sort of distributed team setting. The following discussion focuses on some of the lessons learned and observations made.

Keep in mind the unique characteristics of this experiment. It involved contribution and collaboration between several countries, multiple experimental trials using geographically distributed teams, a synthetic environment, and encrypted communication and networking. These points are important because they are fundamental to lessons learned and to defining ways for improving this type of study.

Beginning with data collection – data was collected at various times (before, during and after a run) and measures included workload, perceived SA, shared awareness, communications, behavioural state, scenario realism, tool effectiveness, and biographical data. The overriding lesson learnt is that shorter is better. This may not necessarily be true for a study consisting of a single trial but in this experiment runs were lengthy and repeated numerous times and data collected directly from participants during and after every session. Input requested from individuals should be as little as is feasible and the majority of responses should be scaled with a minimum of qualitative comments required. Over-taxing participants will simply result in unreliable data. This conclusion is not earth shattering of course. One would hope that the design of any study ensures that adequate amounts of data are collected with minimal toll on participants. In studies like this one however, where repeated, long, intense trials were followed by lengthy data collection sessions, the challenge of finding a balance between collecting sufficient, valid data and participants' limits is increased. Through this discovery exercise we were able to identify the kinds of measures and methods most suitable for future work using a similar synthetic environment, multiple experimental runs, and distributed teams. Thus, an objective of this multi-national study was met.

One of the most informative and unobtrusive measures used in this experiment was the continuous workload measure which, being a scaled response, was also straightforward to analyze. The electronic version of the NASA TLX was also easy to complete and the data useful. Since both measures produced similar results future studies might rely on one or the other, depending on the experiment and/or constraints of the environment. However, having both measures of workload in this study will allow for future work to examine the relationship between a uni-dimensional moment-to-moment scale and the multi-dimensional NASA TLX, along the lines of Van Orden's work (2001). The team version of the NASA TLX, on which individuals estimated the workload of the team as a whole, was not particularly informative, although it could be in

other work if a study was designed to focus specifically on shared awareness amongst team members. According to the responses in the post-scenario questionnaire team members were not able to reflect on how busy their peers were. Participants were likely too involved in their own tasks to think about how other team members were coping. Thus, responses on the team NASA TLX are possibly a reflection of an individual's own workload rather than the team's.

The CARS, another scaled response, was relatively easy to administer and to analyze, as well as being informative, although, as with any subjective measure, estimating situation awareness after the scenario is finished is subject to misrepresentation since responses may be biased by the outcome of the task. Ideally SA should be assessed regularly and objectively throughout a scenario but, since this can require interrupting the event or having observers record specific content of dialogue, it would have been difficult to implement in the present study. A subjective measure of SA, like the CARS, provides a self assessment of how an individual feels with respect to their level of SA, but an objective measure of how much knowledge of the current situation the individual actually has is equally important. The communication data collected electronically by observers was also unobtrusive and relatively easy to format for analysis.

One of the most important issues was training, for participants, but perhaps more so for observers. Our original intention was to examine the incoming and out-going flow of communications. However, it became apparent as the experiment progressed that, for most observers, this instruction was difficult to follow due to high workload and confusion sometimes between out-going and incoming dialogue. Adequate training would have resolved this misunderstanding. Sufficient training could also have identified the difficulty observers were having interpreting labels on the PDA – how does "clarification" differ from "information"? Modifications had been made after the pilot study but further refinement and testing of the PDA interface was obviously necessary. Devoting most of their time to recording interactions observers also had to note behavioural state and significant events during the runs. Those comments captured were useful to the data analysts. All in all, observer tasks need to be very clear and structured and, to ensure inter-rater reliability, observers must be trained to a common criterion level – difficult at the best of times but incredibly challenging when individuals are located around the world. Training with a test set of data through video conferencing is a viable possibility. Added to this, some members of the observer and participant teams had to be replaced to accommodate other commitments. For participants, bringing a new recruit on board brought the challenge of training and integrating new members into the team. Teams should be formed prior to the start of an experiment, and ideally should drawn from an environment in which members are already working together professionally on a regular basis. For observers, new members, especially novices, should be placed in the least demanding position if possible.

The tactical display used in this experiment was new to all participants, thus it provided crews with identical platforms from which to work. Time for sufficient training is imperative to overcoming learning effects before the experiment begins and we believe participants had reached that level. Responses to: "To what extent do you feel the practice and preparation time before beginning the scenario was adequate?" gave an average score of 5.5, with 7 being 'very adequate'. Adequate training was necessary for

officers to become accustomed to their roles which were not identical to those performed in real life. Positions used in the study (CO, ORO, SWC, FPO) were based on the Canadian Navy and all nations do not necessarily have an equivalent positions.

Analyzing all the audio/video recordings, BuddySpace (chat), and TeamSpeak (headphone comms), would be overwhelming but this kind of qualitative data was invaluable for examining specific events or general team behaviour in more depth. All three methods were unobtrusive, and the latter two were in fact part of the suite of tools used by participants to perform their tasks.

Technical and logistical problems continually produced set-backs and, like the human factors component, a primary goal of this endeavour was to learn from the experience. However, if the possibility of technical problems exists in future studies, the experimental design should be modified accordingly to reduce the impact. For example, scenarios were paired so that each scenario pairing included one run with the UAV and one without (Base). The two scenario runs within a pairing were conducted on different days and separated by other scenarios in an effort to ensure that participants did not become familiar with features in the scenario that might affect performance, such as time of attack onset. If technical problems make it likely runs could be aborted it would be beneficial for the sake of consistency in data analysis to conduct paired scenarios back to back. This procedure would also reduce any effect of changes in team membership and team configuration that might occur over the course of the experiment.

CONCLUSIONS

This multi-national distributed team experiment was designed to evaluate the effect of supplying intelligence information from a UAV to a maritime coalition team while testing several methods of data collection. The small number of participants and minimal number of trials limit the evaluation but trends in the data suggest that the presence of the UAV increased situation awareness and decreased workload for the non-OTC ship. Scaled measures were found to be the most informative and straightforward to format and analyze. Those included measures of workload and perceived situation awareness, and measure of communications. Questionnaires, if used, should focus on a key construct and be designed to include questions that are relevant to information contained within each individual scenario.

REFERENCES

Clarke, R. A., (2005). LNG Facilities in urban areas. *A security risk management analysis for the Attorney General Patrick Lynch*, GHC-RI-0505A, 32.

Defence Research & Development Canada (DRDC) (2002). Evaluation Program (Workload.exe). University of Toronto Institute for Aerospace Studies Flight Simulation Laboratory. DRDC Toronto CR 2002 – 123.

Defense Science & Technology Organization (DSTO) (200). *Interaction Recorder Software Version R1.0.1*: Innovation Science Pty Ltd.

Department of Defense and Homeland Security. (2005). *National Strategy for Maritime Security.*

Endsley, M. R.. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37(1),* 32-64.

Entin, E. B., and Entin, E. E. (2001). Measures for evaluation of team processes and performance in experiments and exercises. Aptima Inc.

Galligan, D., Galdorisi, G and Marsland, P. (2005). Net centric maritime warfare – countering a 'Swarm' of fast inshore attack craft. Paper presented at the 10th International Command and Control Research and Technology Symposium – The future of C2.

Hart, S. G., and Staveland, E. (1998). Development of NASA-TLX (Task Load Index). Results of Empirical and Theoretical Research. In Hancock, P. A. & Meshkati, N. (Eds), *Human Mental Workload.* Amsterdam: North Holland Press.

McGuinness, B., and Foy, L. (2000). A subjective measure of SA: The Crew Awareness Rating Scale (CARS). Paper presented in the Human Performance, Situation Awareness and Automation Conference Proceedings, Savannah, Georgia.

Van Orden, K. F. (2001). Monitoring moment-to-moment operator workload using task load and system-state information. San Diego, CA, Technical Report 1864.

Wood, D. I. (2005). Terrorism fears diver navy supply ships from Suez Canal. Newhouse News Services.

LIST OF TABLES

LIST OF FIGURES

Figure 14. Mean number of verbal out-going interactions for CA SWC when CA CO was OTC and when not OTC.

Figure 15. Mean number of verbal out-going interactions for AS SWC when AS CO was OTC and when not OTC.

Figure 16. Mean number of verbal out-going interactions for CA FPO when CA CO was OTC and when not OTC.

Figure 17. Mean number of verbal out-going interactions for AS FPO when AS CO was OTC and when not OTC.