Multi-Modal Measurement for Intelligence Analyst Cell Size Optimization

Kristy Kay
Aptima, Inc.
Woburn, MA
KKay@aptima.com

Blake Martin, David Bryant
Defence Research and Development Canada
Toronto, Canada
Blake.Martin@drdc-rddc.gc.ca, David.Bryant@drdc-rddc.gc.ca

Kent Halverson, Lisa Lucia
Aptima, Inc.
Woburn, MA
KHalverson@aptima.com, LLucia@Aptima.com

Lisa Tripp
Air Force Research Laboratory
Dayton, OH
Lisa.Tripp.1@us.af.mil

ABSTRACT

Current Tactical Operation Centers (TOCs) frequently employ forward-deployed or even integrated processing exploitation and dissemination (PED) of full motion video (FMV). As intelligence, surveillance, and reconnaissance (ISR) platforms continue to proliferate and generate even more imagery, outpacing the growth of the intelligence analyst career field, the demand on analysts to PED in a timely manner continues to be a challenge. Considering this, one way of enhancing and optimizing performance is to reconfigure the PED cell positions, roles, and responsibilities to ensure that limited resources are effectively applied. However, PED cell tasks can be dynamic and complex, resulting in both sequential and reciprocal interdependence, which can blur the lines between roles and potentially cause confusion. Analysts in a PED cell must work in close collaboration and with high efficiency, which places pressure on the crew as it completes required tasks, exchanging workload amongst the team.

The present pilot study aims to validate a research approach intended to investigate the effects of PED cell configuration on performance by introducing an innovative measurement approach consisting of state, process, and outcome performance measures leveraging human-computer interaction (HCI), external observations, and self-reports. In this work, we assessed the performance of two- and 3-person PED cells under increasing workload. PED cell participants completed six simulated scenarios of varying complexity. The following constructs were measured: workload, communication, production quality, performance, situation awareness, and task distribution.

This study is unique in that it introduces a novel, multi-modal approach for measuring workload and team performance in complex, dynamic environments. Multimodal data were analyzed to identify configuration parameters that influence PED cell performance. Preliminary results suggest that the multi-modal approach and data obtained will provide the desired insight necessary for making valid comparisons between PED cell configurations in future experiments.

ABOUT THE AUTHORS

Ms. Kristy Kay is an Associate Scientist at Aptima, Inc. Ms. Kay has research experience in human performance measurement, team dynamics and training, training design and evaluation, survey development, and both quantitative and qualitative data analysis. She holds a M.S. in Industrial-Organizational Psychology from San Diego State University, and a B.A. in Psychology from California State University San Marcos.

Dr. Blake Martin is a Defence Scientist in the Psychological Effectiveness Group of the Human Effectiveness Section at Defence Research and Development Canada's Toronto Research Centre, as well as an adjunct professor in York University's Dance Science Diploma program. Dr. Martin's background is in kinesiology, sensory-motor integration, neuroscience, and education. Blake Martin holds a Ph.D. in Kinesiology with a Graduate Diploma in Neuroscience, an M.A. in Dance, a B.E., and a B.F.A in Dance.

Dr. Kent C. Halverson is the Director of the Performance Assessment Technologies Division at Aptima, Inc. Dr. Halverson has research experience in multimodal performance measurement, organizational analysis, social network analysis, leadership

development, training design and evaluation, and quantitative data analysis. He holds a B.S. in Civil Engineering from the Air Force Academy, a M.S. in Civil Engineering from University of Illinois, and a Ph.D. in Business Administration (organizational behavior emphasis) from University of Florida

Dr. Lisa C. Lucia is a Scientist at Aptima, Inc., where she leads projects on the topics of neuroscience-based training, decision support tools, and medical informatics systems. Her research interests span from the brain-bases of perception and memory to visuospatial abilities and skill learning. She holds a Ph.D. in Cognitive Neuroscience from Tufts University, and a B.S. in Biological Psychology from Bates College.

Dr. David Bryant currently works at Defence Research and Development Canada, Toronto Research Centre in the Psychological Effectiveness Group. David does research in Applied Psychology, Cognitive Psychology and Cognitive Science. He holds a B.S. from the University of Toronto, and PhD in Psychology from Stanford University.

Dr. Lisa Tripp is a Research Psychologist at the Air Force Research Laboratory, where she conducts research on training design and evaluation, human performance measurement, teaming, and both quantitative and qualitative data analysis. She holds a Ph.D. in Experimental Psychology from Washington State University and a M.S. in Applied Mathematics from Washington State University.

Multi-Modal Measurement for Intelligence Analyst Cell Size Optimization

Kristy Kay
Aptima, Inc.
Woburn, MA
KKay@aptima.com

Blake Martin, David Bryant
Defence Research and Development Canada
Toronto, Canada
Blake.Martin@drdc-rddc.gc.ca, David.Bryant@drdc-rddc.gc.ca

Kent Halverson, Lisa Lucia
Aptima, Inc.
Woburn, MA
KHalverson@aptima.com, LLucia@Aptima.com

Lisa Tripp
Air Force Research Laboratory
Dayton, OH
Lisa.Tripp.1@us.af.mil

BACKGROUND

PED/C2 Interaction and Integration

Asymmetrical warfare and the emergence of low-contrast enemies which cannot be readily distinguished from non-combatants has increasingly imposed a need to mass intelligence surveillance and reconnaissance (ISR) efforts for a longer durations in more focused areas. This is to create a more complete tactical and strategic picture, which is of greater use for SOF and UAV prosecution (Flynn, Juergens, & Cantrell, 2008). Moreover, in the modern "find, fix, finish" operation, command and control (C2) structures require either forward-deployed or integrated intelligence analysis personnel so that intelligence, operations, exploitation and analysis of effects can be seamlessly integrated (Flynn et al., 2008; Rosenbach & Peritz, 2011). To be most effective, C2 operations must have continuous information regarding the enemy situation, which can no longer be discerned from the placement of conventional weapons of war such as forward operating bases, aircraft, and tanks.

FMV PED Overview

Increasingly, real-time and near real-time full motion video (FMV) imagery is used to observe adversary behavior to understand and eventually predict. The crews that manage and interpret these data must work in close collaboration to deliver appropriate, accurate, and timely information within a C2 framework. The video feed, typically from an unmanned aerial vehicle (UAV) or manned platform, must be continuously monitored to detect cues in the environment, identify, and screen for relevancy, analyze to describe, explain, and predict adversary behavior. Once hypotheses are supported or refuted with analytic rigor, a report must be generated to explain the nature of the information, including supporting images or video clips, and disseminated to the appropriate partners. The process, exploitation, and dissemination (PED) functions must be performed efficiently even as events continue to unfold in real time, placing considerable pressure on the crews who must continually balance the workload and prioritize required tasks regardless of the crew size or configuration. Currently in Canada, FMV PED crewmembers are trained to fill all roles and therefore, are capable of performing all necessary tasks. In practice, during on-the-job training, operators specialize within teams to screen the incoming video and tag suspect clips, to explore the tagged clips, or to write the related reports. However, the workload distribution of tasks and their effective completion is dependent on crew size, as well as the complexity of a given mission.

Canada PED Training Policy

With the recent announcement of Canada's new defense policy[1] and a commitment to UAV technology, the Remotely Piloted Aircraft System program, (formerly Joint Unmanned Surveillance and Target Acquisition System (JUSTAS) is quite certain to advance. As a result, the use of UAVs and related intelligence opportunities will

[1] Canada's new defence policy, *Strong, Secure, Engaged* was released in June 2017. With a long term funding commitment over a 20-year horizon, the policy specifically identifies remotely piloted systems and related training as key priorities for armed, naval, air and joint forces (National Defence, 2017).

increase, however the PED capability in the Canadian Armed Forces (CAF) has relatively few operators and the concept is still under development. The qualification standard (National Defence, 2012) and training plan (National Defence, 2016) for FMV operators do not explicitly address collective training nor the optimal functioning of PED cells, nor how the cell can best manage workload through task division or crew size. It is important therefore to characterize the nature of PED cells through a structured exploration of workload and the advantages and liabilities of two or three person teams. Outcomes such as the number and quality of intelligence products and communications with those external to the team, and the latency of their production in relationship to critical events in the scenario, could provide important understanding into the functioning of these teams. Equally, measures of processes within the team, such as the load imposed by the work, its distribution among team members, and timings and quality of communications within the team would provide insight into team function.

Study Description

In early 2017, three FMV subject matter experts (SMEs) and a defense scientist from Canada collaborated with American SMEs, scientists, and contractors at the Air Force Research Lab (AFRL) in Dayton, Ohio to develop a series of simulator scenarios of increasing difficulty. Over five days, this team identified 12 mission categories (Helicopter Landing Site (HLS) soak, Route Study, Counter IED, Source Follow, Wide Area Search, Domestic Search and Rescue, Convoy, HLS Infil/Exfil, Troops in Contact (TIC), Weapons Employment, Collateral Damage Estimate (CDE), Battle Damage Assessment (BDA)) employing a set of 11 complexity parameters (Search Geometry, Tasking Type, Recognizability, Specificity of Essential Element of Information, Product Timing, Team Cohesion, Priority Match, Communication Clarity, Weather, Airspace Restrictions, Environment) expected to influence the difficulty of executing a PED mission They then designed seven, 75-minute scenarios consisting of three vignettes representing one of the 12 mission categories. Complexity parameters manipulated were restricted to Search Geometry, Tasking Type, Recognizability, Communication Clarity, Weather and Environment. An eighth script was generated in Toronto to match the minute-by-minute pacing and difficulty of the first, as a comparison to determine training delta. The scenarios were intended to represent ‘crawl’, ‘walk’ and ‘run’ levels of difficulty with scenarios 1 and 8 both containing one vignette at each level.

The scenarios were scripted in the Testbed for Integrated Ground Control Station Experimentation and Research (TIGER) environment, a simulated training testbed consisting of a suite of integrated and interactive hardware and software. The TIGER platform can be configured for different aircraft and crew situations, permitting examination of parameters affecting team integration, function and workload. Six of the eight scenarios were sufficiently developed in the simulation environment to be used in the pilot study.

The present work is the first step in a longer-term research effort intended to characterize the behavior of two and three person PED crews under increasing workload in order to develop PED crew policies. This pilot study involved a very small convenience sample to test the study protocol, validate the scenario content, analyze data, and capture lessons learned to consider for a larger, future experiment.

METHOD

Participants

The participants in this pilot study were recruited from among qualified intelligence analysts in the Royal Canadian Air Force. Due to a small population of PED FMV analysts eligible to participate, the pool of participants was expanded to include individuals with related intelligence analysis expertise such as an Airborne Sensor Operator or an Intelligence Operator. The individuals were selected such that criteria for experience, age, and qualifications were similar to those of other participants. Senior officers from the PED FMV community assisted with selection. Five analysts total were used in this study divided into one team of two (week one) and one team of three (week two). The two-person team had the most experienced and qualified participant, with the greatest amount of training and experience in-theatre. All other participants were relatively equal in experience. All graduated from high school, and two had some college education. Ages ranged from approximately 21-45; to reduce the likelihood of having identifiable information considering the small population the participants are a part of, the researchers asked for age ranges rather than exact ages. Participants ranged between 2 to 11 years of service in the regular Forces. All research was conducted in compliance with DRDC Human Research and Experimentation Committee guidelines.

Manipulation Check

As described previously, researchers attempted to manipulate scenario difficulty by modifying complexity parameters within each scenario. The difficulty levels of the missions were qualitatively validated by subject matter experts (SMEs); for more detail on their development, interested readers are directed to Martin and colleagues (2018). In order to establish evidence that the manipulation functioned as intended, participants evaluated scenario difficulty across a variety of parameters, which will inform the development of adaptive training. Difficulty was assessed using the NASA-TLX, experiential self-report of workload and additional user surveys and compared with external observer reported data and human-computer interaction (HCI) data. This provided user-assessed, SME-assessed, and objective levels of workload for each scenario, which will aid development of progressive training, and will further contribute to understanding of measures as indicators of workload.

Procedure

Researchers administered pretest instruments to obtain demographic information and baseline physiometric data. Participants performed the NASA Multi-Attribute Task Battery II (MATB-II; Comstock Jr & Arnegard, 1992) and the NASA Task Load Index (TLX). The NASA-TLX is a subjective, questionnaire-based tool to assess workload across a number of dimensions, and has been used in literally hundreds of studies (Rubio, Díaz, Martín, & Puente, 2004). The TLX indicates how taxed the rater felt, regardless of the actual toll of the workload. Both paper and electronic versions of the instrument are available. The questionnaire is typically completed at the conclusion of a work session, providing a post hoc summary of user workload. A performance benchmark for the NASA-TLX and other measures can be obtained through a standardized workload baseline task, such as the MATB-II. The MATB-II requires the simultaneous performance of monitoring, dynamic resource management, and tracking tasks.

Prior to the first scenario, the MATB-II was presented to participants on a laptop computer equipped with a mouse, joystick and headphones. After brief training and practice, participants completed a MATB-II task at high level of difficulty (lasting 5 minutes). For each task, participants reported on their perceived workload using the NASA-TLX to serve as ground truth (or “baseline”) for later use in the analysis of scenario-related data. Finally, they were familiarized with the TIGER set-up and oriented to the study protocol.

All participants performed the six different selected simulator scenarios of varying difficulty over the course of five days. Each scenario began with a pre-mission brief led by a knowledgeable researcher who acted as the Imagery Mission Supervisor (IMS), a superior who has direct command over the PED crew. Next, the PED crew performed the scenario, created products, and finally, the scenario concluded with a debrief. Following each scenario, participants completed the NASA-TLX questionnaire and an additional post-scenario questionnaire asking them to reflect on and evaluate aspects of the scenario. During each simulated scenario, participants performed all tasks related to the PED role to the best of their ability. After completing all scenarios, participants completed a final questionnaire asking about relative difficulty of all scenarios and changes to workload and workload contribution over the course of the week.

Measures

This pilot study applied a multi-modal measurement approach leveraging traditional self-report instruments, subjective assessments using a tablet-based tagging system, and unobtrusive HCI data. HCI logs were recorded and analyzed for dwell time in windows, keystroke and mouse click rate, and application use. Observer recordings of participant behavior were assessed on paper and using a tablet-based subjective assessment tool (MacMillan et al., 2013) and tagged for target behavior, then analyzed for indices of team communication and performance. Table 1 below shows the modalities that were used to measure the constructs included in this study. The operationalization of each construct is described per measurement modality in each of the applicable sections below.

Table 1. Mapping of Constructs by Measurement Modality

	Constructs				
Modalities	Workload	Workload Distribution	Communication	Situational Awareness	Performance
Self-report	X	X	X	X	
Behavioral observation	X				X
Human-Computer Interaction	X		X		

Recall that this pilot study used a very small sample size, and therefore, it is not possible to make inferences about the overall character of the population for most measures. Nevertheless, parametric and non-parametric statistical should yield insight on workload and crew size in this particular instance.

Self-Report

After each scenario, participants completed a questionnaire on their perceived workload, difficulty of the scenario, estimated contribution to the overall work of the team (i.e., workload distribution), team communication, and situational awareness. Finally, as a longitudinal measure of workload, participants were asked to report the workload (high, medium, or low) they were experiencing at that moment every 4 minutes in response to an automated chat window.

Additionally, the NASA-TLX questionnaire was presented through an app on an iPad mini. The questionnaire includes six subscales measuring the dimensions of Mental, Physical and Temporal Demands, as well as levels of Frustration, Effort and Performance, all of which contribute to an overall measure of workload (Hart, 2006). Each subscale is a 21 level gradient with anchors ranging between 'Low' and 'High' except Performance, which ranges from 'Perfect' to 'Failure.' Users tap at the appropriate place on the screen to indicate their rating for a given scale. An additional portion of the questionnaire consists of paired rankings for the relative importance of each subscale, allowing raters to weight the relative contribution of each subscale.

Behavioral Observations

A researcher rated participant behavior using an assessment rubric for both PED configurations over the two weeks. A paper version of the rubric was used for the entire pilot study period, while a tablet-based version was used during the second week to test the platform and usability of the application. Technical difficulties prevented the use of the tablet during the first week. Participants were rated on performance (correct identification and detection), as well as their workload. Prior to data collection sessions, behavioral descriptions for these tags were defined, and raters received training and practice with the technology.

The tablet-based assessment tool, the Scenario-based Performance Observation Tool for Learning In Team Environments (SPOTLITE), was used to systematically document team-based performance with behaviorally anchored rating scales for pre-defined factors of interest (MacMillan et al., 2013). In this method, a timeline is annotated with tags as appropriate, so that observed behaviors are synced with specific events in training scenarios and linked to a specific participant (or set of participants). Importantly, these data were recorded differently between

the two behavioral observation tools. On paper, the SME provided an overall score per 10-minute period. In the tablet-based tool, assessments were recorded regularly throughout the scenario. These scores are compared by averaging the regular scores in the tablet-based tool for the same 10-minute increments that were used in the paper-based tool.

Human-Computer Interactions (HCI) Logger

Keystrokes and mouse use was recorded for offline analysis using custom HCI logging software from AFRL. HCIs can be recorded automatically and therefore unobtrusively. Using an HCI logger, researchers can monitor and record software application use (e.g., active program windows, accessed sites, folders, and documents) and computer peripheral activity (e.g., mouse clicks, keyboard inputs). Actions are time-stamped to the millisecond enabling precise tracking of timed responses for each participant. Prior work suggests that HCIs can be used to identify boredom and engagement (Bixler & D'Mello, 2013), confidence, hesitance, nervousness, relaxation, sadness, and tiredness (Epp, Lippold, & Mandryk, 2011), psychological distress (Karunaratne, Atukorale, & Perera, 2011), and stress (Koldijk et al., 2014; Rodrigues et al., 2013). HCI logs can be reviewed offline affording careful examination and analysis of data.

For the purposes of this exploration, the researchers assumed every interaction with the computer to be a unit of work, whether a mouse click, keyboard stroke, or interaction with a window. Analyses that are more refined are possible, but were not performed at this time. Thus, when referring to workload, this construct is operationalized as the number of acts. Similarly, communication is operationalized quantitatively, such that the researchers analyzed the number of chat acts.

RESULTS

Descriptive Results

Self-Report

Table 2 shows the average post-scenario reported workload, workload distribution, communication, situational awareness, experiential workload, and workload according to the NASA-TLX of each team configuration across all scenarios. Scenarios 1 and 8 include vignettes at each difficulty level.

Data consistently shows 2-person team had higher workload. However, the 2-person team also reports better workload distribution, communication, and situational awareness than the 3-person team. Scenarios 1 and 8 offer an interesting opportunity for comparison because both include vignettes at each difficulty level (i.e., 1, 2, and 3) throughout the scenarios. Between scenarios 1 and 8, teams have the opportunity to work together throughout the week, so it would be expected that all variables would show improvement within each team. Although the 2-person team showed a decreased workload in scenario 8 compared to scenario 1, the 3-person team experienced the opposite, less-expected outcome. It is possible that this is due to the substantial difference in workload distribution experienced for the two teams during these scenarios. Whereas both teams completed scenario 1 with similar workload distributions, the 2-person team was substantially more effective in distributing their workload than the 3-person team during scenario 8. Thus, some of the 3-person team members rated their workload more drastically in scenario 8, perhaps due to the poor workload distribution, whereas the 2-person team did not indicate the same issue.

Table 2. Self-Report Results

	Scenario	1	2	3	4	6	8	Avg	SD
	Manipulated Scenario Difficulty	**1-2-3**	**1**	**2**	**2.5**	**1.5**	**1-2-3**		
Source: Post-Scenario Survey	**Workload**								
	2-person team	3.25	3.50	2.50	3.50	3.25	3.00	3.17	0.38
	3-person team	2.17	2.00	2.50	2.33	3.33	2.67	2.50	0.47
	Workload Distribution								
	2-person team	2.83	3.33	3.17	3.67	4.33	4.33	3.61	0.62
	3-person team	2.94	3.33	3.22	3.33	3.56	3.44	3.31	0.21
	Communication								
	2-person team	3.83	3.67	4.33	4.33	3.83	4.50	4.08	0.35
	3-person team	3.11	4.33	4.11	3.67	3.67	3.89	3.80	0.42
	Situational Awareness								
	2-person team	4.00	4.00	4.50	4.50	4.50	4.50	4.33	0.26
	3-person team	2.67	3.50	3.33	4.00	3.33	4.00	3.47	0.50
Source: Workload Reported to Automated Chat Prompt	**Experiential Workload**								
	2-person team	1.50	1.56	1.55	1.78	1.63	1.53	1.60	0.10
	3-person team	1.47	1.20	1.28	1.44	1.64	1.91	1.49	0.26
Source: Post-Scenario NASA-TLX Questionnaire	**NASA-TLX Workload**								
	2-person team	55.83	60.00	47.83	52.50	58.83	50.33	54.22	4.82
	3-person team	51.89	46.33	29.56	42.00	46.67	40.33	42.80	7.64

Behavioral Observations

Table 3a shows the results from the paper-based behavioral observations across scenarios for both PED cell configurations. Detect and Identify scores were converted into proportions of correct behavior on each dimension, as compared with a scenario timeline established by the researchers and a SME. These two constructs were considered performance indicators. Workload scores are on a 1-3 scale, where 1 = low, 2 = moderate, and 3 = high.

These data indicate that the 2-person team was consistently perceived by the researcher as performing better than the 3-person team on both performance constructs, though had higher workload overall.

Table 3b shows the results from the tablet-based assessment tool method for the second week only used to record performance and behavior. This was used in the second week by a second SME in conjunction with the original SME recording on paper. This method was included in order to compare the utility of this tool.

Across all three common dimensions, the difference in average recorded score was very close for Detect (delta = .03, proportion correct) and Workload (delta = .12, 1-3 scale). However, the difference was larger for Identify (delta = .15, proportion correct).

Table 3a. Paper Behavioral Observation Results

Scenario	1	2	3	4	6	8	Avg	SD
Scenario Difficulty Level	**1-2-3**	**1**	**2**	**2.5**	**1.5**	**1-2-3**	**-**	**-**
Detect								
2-person team	0.83	0.84	0.80	0.97	0.56	0.92	0.82	0.14
3-person team	0.84	0.74	0.79	0.63	0.58	0.78	0.73	0.10
Identify								
2-person team	0.76	0.86	0.83	0.88	0.74	1.00	0.84	0.09
3-person team	0.72	0.67	0.68	0.88	0.69	0.83	0.74	0.09
Workload								
2-person team	3.00	2.88	1.75	3.00	1.63	2.13	2.40	0.64
3-person team	2.00	1.75	2.00	1.88	1.25	1.75	1.77	0.28

Table 3b. SPOTLITE Behavioral Observation Results

Scenario	3	4	6	8	Avg	SD
Scenario Difficulty Level	**2**	**2.5**	**1.5**	**1-2-3**		
Communicate	0.79	0.89	0.89	0.95	0.88	0.06
Detect	0.74	0.74	0.72	0.85	0.76	0.06
Identify	0.89	0.92	0.77	0.98	0.89	0.09
Workload	1.00	1.00	2.42	2.19	1.65	0.76

HCI

Across all scenarios and for each team configuration, Table 4 shows the total count of human interaction (e.g., keyboard and mouse clicks) detected by the HCI listener and the total chat activity. Due to technical difficulties, data during scenario 3 for the 2-person team are missing. Although these data are missing, the impact this has on the interpretation of results is minimal. Generally speaking, reliability of the aggregate would benefit had all scenarios been recorded. However, two primary points assuage this concern: a) 5 total data points are still sufficient to ascertain an accurate average for comparison; and b) the scenario *was* completed by the team. In other words, the data are not missing because they did not occur, rather because the recording device was indisposed.

In general, the 2-person team had fewer total acts (apart from Scenario 2) compared to the 3-person team. Initial communication (operationalized as chat activity count) was initially higher for the 2-person team than the 3-person team. However, as the teams progressed through the scenarios, the 3-person team increased their communications, surpassing the 2-person team. Meanwhile, the communication within the 2-person team remained relatively consistent in quantity across the scenarios.

Table 4. HCI Results

	Mission 1	Mission 2	Mission 3	Mission 4	Mission 6	Mission 8	Avg	SD
Total Activity Count								
2-man team	10440	16019		10074	12214	12451	12239.60	2109.86
3-man team	11327	10733	11035	12785	14214	14819	12485.50	1583.24
Chat Activity Count								
2-man team	6168	7658		7219	7866	7822	7346.60	632.16
3-man team	5987	5790	7018	9628	9932	10357	8118.67	1904.14

Inferential Results

Self-Report

The researchers observed mean differences over time in self-report measures. It seemed intuitive that workload would be higher for the 2-person team than the 3-person team. In five of the six scenarios, self-reported post-scenario workload was higher for the 2-person team than the three–man team; in the single scenario where this is not true—the difference is minimal (2-person team $\bar{x}$ = 3.17, 3-person team $\bar{x}$ = 2.50). Similarly, it seemed intuitive that the workload distribution among team members would be better for the 3-person team than the 2-person team. Interestingly, the workload distribution between teams was as expected initially in the study. For the first three scenarios, workload distribution was worse for the 2-person team compared to the 3-person team. For the last three scenarios, the 2-person team actually reported better workload distribution compared to the 3-person team (2-person team $\bar{x}$ = 3.61, 3-person team $\bar{x}$ = 3.31). Communication was also reportedly better in the 2-person team in all but one scenario (2-person team $\bar{x}$ = 4.08, 3-person team $\bar{x}$ = 3.80). Additionally, situational awareness was reportedly higher in the 2-person team across all scenarios (2-person team $\bar{x}$ = 4.33, 3-person team $\bar{x}$ = 3.47). Finally, the 2-person team reported slightly higher workload when asked experientially (2-person team $\bar{x}$ = 1.60, 3-person team $\bar{x}$ = 1.49). Thus, despite having slightly higher workload, on average the 2-person team reported better workload distribution, communication, and situational awareness.

In calculating workload with the NASA TLX, the participants completed the NASA TLX after the MATB-II exercise, which took place before any scenarios were completed. Additionally, participants completed a survey associated with the NASA TLX that allows researchers to determine which aspects specified in the NASA TLX actually contributed to workload. An explanation of this process is outside the scope of the present study, but details can be found in Hart, 2006. After applying individual weighting scores, the resulting reported workload was in line with the results thus far, such that across all scenarios, the 2-person team reported higher workload than the 3-person team (2-person team $\bar{x}$ = 54.22, 3-person team $\bar{x}$ = 42.80).

Behavioral Observation

In general, the two PED cell configurations did not differ substantially in their detection and identification of entities during the scenarios, though the 3-person team consistently performed slightly worse (Detect: 2-person team $\bar{x}$ = .82, 3-person team $\bar{x}$ = .73; Identify: 2-person team $\bar{x}$ = .84, 3-person team $\bar{x}$ = .74). Additionally, the researcher performing the observations perceived the workload of the 2-person team to be substantially higher than that of the three man team (2-person team $\bar{x}$ = 2.40, 3-person team $\bar{x}$ = 1.77). Both the higher performance of and perceived higher workload by the researcher for the 2-person team may be attributable to greater focus exhibited by the 2-person team compared to the 3-person team, which is in line with the results of the HCI data.

For exploratory purposes, the researchers used SPOTLITE (in addition to continued paper observations) to record the behavior of the 3-person team during the last four scenarios. In all cases, the resulting ratings were not substantially different when comparing the SPOTLITE results to the paper observation results. Future research should compare these two methods for external behavioral observation between teams for further validation.

HCI

Interestingly, the total activity between the two team configurations did not differ by much (2-person team $\bar{x}$ = 12239.60 acts, 3-person team $\bar{x}$ = 12485.50 acts). The delta between these means is about 2% of their average activity, indicating that workload did not differ substantially between the two teams. However, the 3-person team chatted quite a bit more than the 2-person team (2-person team $\bar{x}$ = 7346.60 chats, 3-person team $\bar{x}$ = 8118.67 chats), with a delta of about 10%. Because there is almost no difference in workload, but more communication among the 3-person team, this indicates that the 2-person team may have had to remain more focused on their individual tasks, thus not communicating quite as much as the 3-person team. Again, consistent with findings from the self-report data and behavioral observation data, the 2-person team seems to perform better overall notwithstanding their higher workload.

CONCLUSIONS

Key Findings

A major strength of this study is the use of a multi-modal measurement approach, allowing for comparison of findings between these different modalities. This allowed the researchers to make some initial conclusions comparing the workload between the two PED cell configurations. First, the average workload is higher for the 2-person team across all measurement modalities. Second, although the 2-person team had less communication via chat, they reported better communication than the 3-person team reported. Two reasons for these seemingly confounding results are most likely: a) the 2-person team, being smaller, had an easier time communicating verbally compared to the 3-person team, and thus utilized this form of communication more often; and/or b) the 2-person team has higher quality communication compared to the 3-person team, resulting in less chat being needed. With the current sample, it is difficult to make overarching inferences. Finally, performance (detect and identify) of the teams in the scenarios unanimously indicates that the 2-person team performed superior to the 3-person team. These findings provide initial evidence that, despite having higher levels of workload, 2-person PED Cell configurations may perform and communicate better than 3-person PED Cell configurations.

Future Research and Limitations

As a pilot study, this project is primarily limited by its inability to make statistically inferences. Additionally, a more balanced and appropriately experienced sample would have been preferable. Specifically, the two-person team had the most experienced and qualified participant, with the greatest amount of training and experience in-theater. The results could also reflect that the other participants simply did not do as well because they lacked the requisite knowledge to identify and perceive targets and workload. Further, an interaction effect (experience X team configuration) may be at play, but this is untestable with the small sample in this study. Future research endeavors aim to repeat the present study methodology with a larger sample of PED analysts. Additionally, future research could benefit from validation of the difficulty of the scenarios created for the present study. Future research may also wish to use the tablet-based behavioral assessment tool. This tool allowed for more regularly behavioral recording. If used, it would be pertinent that constructs are very clearly defined and understood by all raters.

ACKNOWLEDGEMENTS

We would like to thank the Air Force Research Lab for supporting this work, particularly Dr. Lisa Tripp and Dr. Christine Covas. Additionally, this work could not have been completed successfully without the tireless efforts of the staff at Defence Research and Development Canada, namely: Dr. Stuart Grant, Dr. David Bryant, Nada Pavlovic, Leo Boutette, John Townsend, James Newton, Sheridan Torres, and Elaine Maceda.

REFERENCES

Bixler, R., & D'Mello, S. (2013). *Detecting boredom and engagement during writing with keystroke analysis, task appraisals, and stable traits*. Paper presented at the Proceedings of the 2013 international conference on Intelligent user interfaces.

Comstock Jr, J. R., & Arnegard, R. J. (1992). The multi-attribute task battery for human operator workload and strategic behavior research.

Epp, C., Lippold, M., & Mandryk, R. L. (2011). Identifying emotional states using keystroke dynamics. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.

Flynn, M. T., Juergens, R., & Cantrell, T. L. (2008). Employing ISR SOF Best Practices: National Defense Univ Washington Dc Inst For National Strategic Studies.

Hart, S. G. (2006). *NASA-task load index (NASA -TLX); 20 years later*. Paper presented at the Proceedings of the human factors and ergonomics society annual meeting.

Karunaratne, I., Atukorale, A. S., & Perera, H. (2011). *Surveillance of human-computer interactions: A way forward to detection of users' Psychological Distress*. Paper presented at the Humanities, Science and Engineering (CHUSER), 2011 IEEE Colloquium on.

Koldijk, S., Sappelli, M., Verberne, S., Neerincx, M. A., & Kraaij, W. (2014). *The SWELL knowledge work dataset for stress and user modeling research*. Paper presented at the Proceedings of the 16th International Conference on Multimodal Interaction.

MacMillan, J., Entin, E. B., Morley, R., & Bennett Jr, W. (2013). *Measuring team performance in complex and dynamic military environments: The SPOTLITE method*. Military Psychology, 25(3), 266.

Martin, B.C.W., Halverson, K., Keeney, M., Wilson, C., & Tripp, L., (2018), *PED Simulator Scenarios*, Defence Research and Development Canada Report DRDC-RDDC-2018-L058

National Defence. (2012). Motion Imagery Intelligence Analyst qualification standard. Ottawa, Canada.

National Defence. (2016). Motion Imagery Intelligence Analyst training plan. Ottawa, Canada.

National Defence. (2017). Strong, secure, engaged: Canada's defence policy. Ottawa, Canada.

Rodrigues, M., Gonçalves, S., Carneiro, D., Novais, P., & Fdez-Riverola, F. (2013). Keystrokes and clicks: Measuring stress on e-learning students Management Intelligent Systems (pp. 119-126): Springer.

Rosenbach, E., & Peritz, A. (2011). The New Find-Fix-Finish Doctrine. *Joint Force Quarterly*(61), 94.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology, 53*(1), 61-86.