Multimodal Assessment of Workload for Predator Training Research

Chantale Wilson
Air Force Research Laboratory
Dayton, OH
chantale.wilson.1@us.af.mil

Lisa Ramsey, Jill Bloomer
L3 Communications
Dayton, OH
lisa.ramsey.ctr@us.af.mil, jill.bloomer.ctr@us.af.mil

ABSTRACT

The concept of mental workload has received much attention when it comes to multidimensional evaluation of human performance in complex environments. This is especially important in military settings as modern technology and increasing battlefield demands increasingly impact operators' mental workload. Military personnel training must reflect cognitive demands of these real-world environments and trainee experiences of workload should match on-the-job experiences. Most current training systems, including virtual simulations, default to subjective workload measures as the singular data source assessing training effectiveness. Conversely, other emerging research focuses on objective workload measures (i.e., physiological) as indicators of performance and training effectiveness. Our research examined both approaches to measuring trainee workload while operating Predator Research Integrated Networked Combat Environment (PRINCE), a remotely piloted aircraft training simulator. Twelve student pilots from the School of Aviation at the University of North Dakota participated in novel training scenarios within PRINCE as pilot and sensor operators. Brain wave activity was recorded using Advance Brain Monitoring's head-mounted B-Alert X-24 and X-10 systems, capturing electroencephalogram, electrocardiogram, and electrooculography activity throughout training missions. Cardiac rhythms and eye movements were captured with B-Alert as measures of engagement, distraction and fatigue. Intific's Neurobridge software captured and attempted integration of objective workload data with objective simulation performance data collected using the Performance Evaluation and Tracking System. Self-report measures of workload were also gathered immediately following each training scenario. Subjective ratings of workload were significantly different between pilot and sensor operator roles and mission difficulty. Engagement, distraction, and fatigue were significantly different for scenarios, with the most difficult scenario showing the lowest distraction rate, highest workload, and lowest performance success. Results shed light on the congruency between subjective and objective measures of workload. These measures offer insight into capturing, synthesizing, and characterizing multifaceted workload data to better understand how workload relates to performance during training.

ABOUT THE AUTHORS

Chantale Wilson is a Research Psychologist at the Air Force Research Laboratory's Continuous Learning Branch, Warfighter Readiness Research Division at the 711th Human Performance Wing, Human Effectiveness Directorate. Her research centers on the development and evaluation of tools to enhance performance and learning for individual, team, and team-of-teams training for military operations. This includes the integration of physiological monitoring capabilities with current simulation-based technology and performance measurement methods used in warfighter training research. Her work involves generating innovative performance management solutions and testing multidimensional approaches to training evaluation. She manages multiple research efforts on game and simulation-based training, as well as international collaborations for joint and coalition training research. She holds a Master's degree in Industrial/Organizational Psychology from the University of Akron and is completing her doctorate degree.

Lisa Ramsey is a Research Scientist and Manager for Link Simulation and Training. She received her M.S. and B.S. degrees in Applied Psychology from Arizona State University (2006, 2003, respectively). She has also completed certifications in Project Management from Villanova University (2014). Her research has focused on creating and validating methods, models, and technologies for learning and training warfighter readiness by exploring human behavior through both quantitative and qualitative research methods. This includes usability evaluation of a learning management system for pilot training using user-experience design, developing tools for assessing pilot training against mission essential competencies, training intuitive decision-making in a simulated real-world environment using cognitive research design, and assessing workload of Predator pilots through task analysis, physiological, and cognitive methods. She has also served as adjunct faculty for both Arizona State University and Mesa Community College designing course curriculum including lecture presentation materials, applicable exercises, and knowledge assessments for research methods and physiological psychology courses.

Jill Bloomer is a Research Psychologist for Link Simulation and Training at the Air Force Research Laboratory, Warfighter Readiness Research Division at the 711th Human Performance Wing. She received her B.S. degree in Cognitive Psychology from Wright State University in 2013. Her focus has been evaluation and enhancement of human decision-making and learning based on visual, haptic, and tactile stimuli in diverse, simulated environments for both basic human tasks and military essential operations with an emphasis on intuitive decision-making. She has contributed to the development and validation of training research models, cognitive design, and execution protocols for evaluation of individuals and teams in wide range of military operations for the advancement of warfighter readiness. She has also contributed to development of haptics research and its effectiveness on learning at Wright State University.

Multimodal Assessment of Workload for Predator Training Research

Chantale Wilson
Air Force Research Laboratory
Dayton, OH
chantale.wilson.1@us.af.mil

Lisa Ramsey, Jill Bloomer
L3 Communications
Dayton, OH
lisa.ramsey.ctr@us.af.mil, jill.bloomer.ctr@us.af.mil

INTRODUCTION

In response to continued threats to national security, technology is developing at a rapid pace to enhance the efficiency, safety, and superiority of our warfighters. Consequently, there is a steady and continual increase in perceptual, attentional, and performance load on warfighters. The mounting complexity of operating technology, combined with battlefield demands, make high cognitive workload ubiquitous (Hancock & Szalma, 2008). Many studies have been conducted to assess cognitive workload (also referred to as mental workload) in order to determine ways to reduce performance degradation and fatigue while increasing efficiency and productivity with various attention-saturated tasks (e.g., driving and air traffic control). Over the years there has been particular interest in developing cognitive workload measures for aircraft pilots (Sanders & McCormick, 1987). Certain piloting tasks require attention allocation between routine performance information and new or unexpected events. According to Wickens (2002), the number of piloting tasks are so extensive that checklists serve as reminders of what to do, but fail to capture switching between tasks or unexpected events. These types of tasks require significant attentional resources that can negatively impact cognitive workload and subsequently, performance.

Generally, cognitive workload refers to the level of mental effort one is exerting on a task (Craven et al., 2007). Pinpointing which tasks are taxing, or the number of tasks that produce higher workload, can inform how warfighter roles are tailored and therefore, how training is developed and delivered so individuals can adapt to or optimize their workload on the job. The literature has suggested several ways to characterize workload based on goals of the situation and the environment. From a theoretical perspective, Multiple Resource Theory describes an aspect of workload in which someone is overloaded with tasking, causing a possible breakdown or slowdown in work performance (Wickens, 2002). Based on this theory, workload refers to the potential to perform a task in dynamic, high-demand situations. The reverse is also true. There can be an ‘underload’ of cognitive disruption where task interference is irrelevant but the inattention itself can cause the operator to become passive and less aware. Remotely Piloted Aircraft (RPA) pilots encounter these types of experiences, which include unpredictable, short periods of extreme attentiveness (higher attentiveness) mixed with longer periods of mundane, static activity (lower attentiveness).

Moore, Ivie, Gledhill, Mercer, and Goodrich (2014) propose dividing cognitive workload into two categories: parallel sensing and sequential decision making. Parallel sensing refers to perceiving complementary stimuli over different modalities, such as visual, auditory, and haptic stimuli. An example of this would be listening to instructions for a task while simultaneously watching information on a display. If the information coming through each of these channels is competing for one’s attention, or offers conflicting information, this can increase an aspect of cognitive workload known as attentional workload. Sequential decision making refers to making decisions in a chronological order. When it is imperative to make multiple decisions to maintain balance or complete a task, a ‘bottleneck’ can occur, such that multiple channels require a decision to be generated immediately or else the information exceeds the limits of working memory, which induces attentional workload.

Additionally, another aspect of cognitive workload, known as temporal workload, can occur, which refers to the scheduling of prioritized, infrequent, and/or repetitious tasks (Moray, Dessouky. Kijowski, & Adapathya, 1991). Two characteristics influencing temporal workload are: (1) tasks which are constrained by time, order, or cause a scheduling conflict, and (2) operational tempo, which indicates how frequently new tasks occur. Workload increases when managing the rate of arrival and when response to information is required in order to make a decision. Therefore, subjective accounts of workload increase as a function of time pressures. In the current study with RPA operators, temporal workload was manipulated by mission time criteria (creating the time pressure), increasing task complexity (causing a conflict), and number of new tasks presented (affecting the overall tempo). Each mission was broken down into segments. These segments represented a change in workload—whether that change was a new task introduced, an increase in urgency, or a change to the task priority. To accurately measure attentional and temporal workload, both objective and subjective measures were captured and are described in later sections.

Most RPA systems require more than one person to operate them, requiring the allocation of workload between operator roles. During any given mission there may be various tasks requiring user input and attention including: operating or flying the RPA, entering flight data (e.g., entering waypoint coordinates), managing a payload (e.g., camera) or managing mission objectives (e.g., tracking a moving target), (Moore et al., 2014). Understanding how cognitive workload affects a two-man crew is essential for determining automation of tasks and procedures (e.g., path planning), allocation of tasks and functions to operators, as well as understanding how to most effectively train an RPA crew, such as by adapting task difficulty, teaching strategies, or techniques to optimize cognitive workload.

According to Shriram (2013), to successfully measure physiological workload the source must be sensitive (able to detect changes in workload), accurate (mirror changes in workload), valid (dependent on workload and no other factors), and reliable (predict workload each time the measure is used). At the Air Force Research Laboratory (AFRL), cognitive workload has historically been captured using subjective measures such as self-reported surveys or ratings. The NASA Task Load Index (TLX), for instance, is currently one of the more popular subjective workload rating scales used in hundreds of research studies and applied to various occupations, from air traffic control to nuclear power plant operations. The goal of NASA-TLX is to pinpoint empirical factors tied to specific workload experiences (Hart & Staveland, 1988). Assessments of 'perceived' cognitive workload may not provide a complete picture of an operator's workload experiences. Alternatively, studies have implemented physiological measures as more objective methods of assessing workload. This includes heart rate, galvanic skin responses, and blood pressure (Fournier, Wilson, & Swain, 1999; St. John et al., 2004; Wetherell & Carter, 2014; Vogt, Hagemann, & Kastner, 2006). Some studies have also used magnetic resonance imaging (MRI) and functional MRIs, which are usually very expensive, making research less viable for many laboratories. In more recent studies, the use of electroencephalogram (EEG) provided a quantitative measure of workload by measuring brain wave activity in real time. The EEG measures cognitive state by recording electrical activity of the brain across the cerebral cortex, where higher-order cognitive processing takes place. These waveforms are categorized according to their frequency, amplitude, and shape, as well as the locations on the scalp where they are recorded. EEG signals are generally classified into four bands: alpha waves (8–13 Hz), beta waves (> 13 Hz), theta waves (4–7.5 Hz), delta waves (≤ 4 Hz) (Shriram, Sundhararaajan, & Daimiwal, 2013). Beta waves are associated with changes in complexity, whereas delta and alpha waves indicate changes in complexity and volume. An EEG shows that growing task demands and time-on-task is associated with an increase in the frontal theta activity and a decrease in parietal alpha activity (Holm, 1989). In the current evaluation, we used the B-Alert wireless head-mounted physiological recording device detailed later in this paper. This system combines EEG, ECG (electrocardiogram), and EOG (electrooculography) to characterize cognitive workload in a quantifiable format.

The PRINCE system is a modular, modifiable, and deployable mission-rehearsal simulator offering research opportunities to fill training gaps for Air Force Special Operations Command (AFSOC) MQ-1 and MQ-9 operators (USAF, 2010). Human-in-the-loop systems, such as PRINCE used in this evaluation, have become more robust in training capabilities. As a result, these high-tech systems can stimulate higher cognitive workload for the operators. Improving the functionality of these systems, therefore, requires a better knowledge and understanding of workload and cognitive states to determine the threshold of workload saturation for simulated mission-rehearsal tasks. Unlike a pilot's task of being able to physically control an airplane, RPA operators have more visual and cognitive challenges to maneuver the aircraft. Under high cognitive workload conditions, an RPA operator must be able to allocate attention across several tasks and screens, dividing attention among multiple external inputs and make instantaneous decisions without much of the physiological feedback pilots receive in a cockpit (Horst, Mahaffey, & Munson, 1989).

PRINCE is capable of training RPA crews effectively using high-fidelity scenarios; however, the level of cognitive workload these scenarios induce and sustain in trainees has not been previously quantified. There were three primary goals in this study. First, by examining the relationships between PRINCE scenario events and subjective and objective measures of workload, it is possible to better characterize PRINCE scenario events associated with cognitive workload from a multidimensional perspective and pinpoint which types of RPA activities and events predict various levels of workload. This can inform future scenario development and integration with RPA training programs for higher fidelity RPA training experiences. Secondly, comparing objective and subjective measures of workload and their degree of predictive power can inform how subjective and objective workload assessments can and should be used in future research with complex, synthetic learning environments. Lastly, we demonstrate how objective cognitive workload data can be integrated with objective simulation performance data in order to evaluate multiple facets of human performance in realistic and operationally-relevant domains. This has implications for better understanding the impact of technological demands on human performance which, in turn, can be used to develop more effective training (e.g., modifying user schedules and workload protocols).

EXPERIMENTAL OVERVIEW

For this evaluation, we assessed workload through a combination of subjective and objective workload metrics. Objective data was measured by brainwave activity collected via physiological methods and subjective data measured by perceived workload collected via survey data. Using physiological methods to measure workload for an RPA pilot and sensor operator (SO) is a relatively less-charted area. While both objective and subjective methods of workload measurement have proven useful, the combination of both methods in data collection has utility for validating results and creating a more comprehensive understanding of the role of workload in complex, simulated environments. Validation of cognitive state measures generally involves experimental manipulation of task demands to induce cognitive state changes, objective measurement of performance metrics (e.g., accuracy, reaction time), and subjective measures that allow participants to describe their perceived level of difficulty, as well as the amount of effort exerted in a given task (Berka et al., 2005). A general task engagement and mental workload study with 80 participants utilized EEG monitoring during a variety of mental tasks such as verbal and image-learning, memory tests, digit-span, recall, and addition tests (Berka et al., 2007). There was a significant physiological increase in engagement and workload during the encoding portion of memory tests compared with recognition/recall portions. Workload also increased with level of difficulty in digit-span, recall, and addition tests. EEG measures correlated with both subjective and objective performance measures, suggesting that these are valid methods to accurately measure cognitive workload. The tasks required of RPA Pilots and SOs vary in complexity, and external circumstances often affect the amount of effort required to successfully complete otherwise routine assignments. EEG measures may prove effective in capturing and predicting workload for various RPA duties in a wide array of situations.

Obtaining a dependable indicator of workload using subject matter expert (SME) interviews is a well-established method. For example, Aldrich, Szabo, and Bierbaum (1989) employed SMEs to obtain valid task time estimates to establish operator workload prediction models, which were useful in the development of new military weapons. Workload was established for each time segment by adding the workload ratings, which rendered a correlation of .74 between predicted workload levels and subjective pilot workload ratings. Lai and Lamoureux (2012) utilized the expertise of SMEs in the development of new systems to support a submarine command team. Experimentation included human-in-the-loop processes in a virtual environment, for which multiple SME interviews were relied upon to validate measures of performance by individuals (or small teams) in system success, measures of operational effectiveness in achieving the overall objective, and methods of quantifying both objective and subjective measurements of the study. Each SME was presented with a selection of functions considered to be most important and asked to rate his agreement with the defining characteristics and the appropriateness of each task.

For the current study, researchers and SMEs developed and validated workload indicators to identify and verify gaps in RPA training with PRINCE. Pilot and SO tasks, enlisting the skills required to navigate an RPA and effectively track targets, were outlined as well as potential obstacles that an RPA might encounter. These tasks and potential obstacles were evaluated and scored on a scale of one to five by individual SMEs for predicted workload indicators, and these ratings were compared against each other. Based on these ratings, six 20-minute scenarios were constructed, creating predicted fluctuations in workload throughout each session. After completion of the blueprint for potential scenarios, SMEs were interviewed again for input regarding the combination of elements and their potential effects on predicted workload. Alterations to scenarios were made accordingly. SMEs were essential to the development of the workload indicators involving various tasks and complexity levels to induce a predicted assortment of workload for each scenario. Table 1 below contains a list of SME-identified tasks for the RPA scenarios.

Table 1. RPA Participant Task Definitions

RPA Task	Definition
Call	Receiving communication for assignment
Copy	Begin preparation and implementation of task
Lull	Awaiting new task
Fly	Flying the aircraft to appointed destination
Search	Attempting to locate the point of interest (POI)
Track	Following movement of the POI
Reached	Task completed

Although past studies have used the simulation of select RPA piloting skills in a limited capacity to assess changes in cognitive workload for novices (Ayaz et al., 2012; Afergan et al., 2014), the unique structure of partnership between the pilot and SO in the field, as well as the workload demand for this specific set of fundamental skills, have received less attention in past research. Each task unique to these positions, at varying levels of difficulty, have yet to be identified in terms of the magnitude of cognitive workload, as research specific to this skill set is virtually nonexistent. The focus of this effort was to develop workload indicators, induce and manipulate levels of workload, and capture workload data for an RPA pilot and SO using both objective and subjective methods. The goal for this evaluation was to inform future RPA training research with regards to manipulating and accounting for cognitive workload as part of evaluating human performance in these highly-complex domains.

Methods

Subjects. A total of twelve student pilots from the School of Aviation at the University of North Dakota participated in the workload evaluation. A total of two females and ten males participated voluntarily. The mean age was 23.7 years and the median RPA flight experience was 40 hours. The median combined manned and unmanned flight experience was 170 hours. All participants completed the evaluation by participating in two separate sessions lasting approximately two hours each, once playing the role of pilot and once in the role of SO. The sessions were not consecutive. Participants were paid $15/hour for their participation in each session.

Technical description of hardware/software. The evaluation was conducted using PRINCE, the AFRL simulator used for training operations that addresses AFSOC mission requirements. PRINCE enables research and development in an unrestricted and high-fidelity environment that supports training effectiveness for RPA crews as well as collaboration between the United States Air Force, joint, and coalition partners. PRINCE delivers a collaborative solution for the warfighter by addressing training gaps through research data collection in an effort to develop adaptive training models, curriculum, and learning management systems. The PRINCE simulator used in this evaluation was built and maintained by the University of North Dakota (UND). UND's PRINCE simulator was constructed in accordance with the AFRL PRINCE team's hardware specifications. A number of Commercial Off-The-Shelf (COTS) and Government Off-the-Shelf (GOTS) software tools were used in this evaluation. The Meta VR Scenario Editor application was used to construct and run the scenarios used during the study evaluation. This allowed the team to add cultural features (e.g., buildings and objects), as well as behavioral features to include moving vehicles, animals, equipment, and persons. Individual behaviors were scripted to occur at specific pre-determined scenario time(s). Brain wave activity was recorded using two head-mounted systems from Advanced Brain Monitoring (ABM), the B-Alert X-24 and the B-Alert X-10. Both are wireless, Bluetooth®-enabled (with a USB receiver dongle) EEG monitoring systems. B-Alert was used to obtain brain wave activity, cardiac rhythm, and eye movements from the participants. The X-10 uses nine channels to monitor EEG activity and one channel to monitor ECG and EOG activity. The X-24 uses twenty channels to monitor EEG activity and four channels to monitor ECG and EOG activity. The B-Alert capture equipment was affixed to all participants during the evaluation and provided raw data on a number of brainwave frequencies. A baseline cognitive assessment included with the X-24 system was first used to provide a baseline of brain and cardiac activity, as well as eye movements from each pilot participant. The X-10 was connected using software from Neurobridge and did not have a baseline cognitive assessment included; therefore, no baseline data was obtained from the sensor operator participants.

The GOTS tools used in this evaluation included several applications from AFRL's 711th Human Performance Wing, the Live Virtual Constructive Network Control Suite (LNCS) and the Performance Evaluation Tracking System (PETS). LNCS is a modular, extensible platform from which a number of simulation and training-related aspects may be controlled. It is implemented as a suite of distributed tools and capabilities used in rehearsal exercises and live-fly domains. For this evaluation, LNCS was used to provide the mission director with a top-down overview of the scenario during the evaluation, to capture user-generated bookmarks for specific events within the scenarios, and to record information on the network to log files for post-test analysis and replay. In addition, LNCS was also used to display waypoints/loiter points to the research team. UND's PRINCE simulator did not contain a LNCS display for the RPA Pilot or SO, so Google Earth was used to push No Fly Zones (NFZ) to these positions as an alternative. PETS is a tracking tool that measures overall objective performance proficiency and provides feedback used for debrief. Here it was used to capture pilot and SO measures of performance by looking at Distributed Interactive Simulation (DIS) data and applying the incoming data to a set of algorithms developed by RPA subject matter experts. For this evaluation, PETS was used to measure times and durations of incursions into restricted operating zones, which was expected to correlate with increased pilot workload. PETS also monitored airspeed, location, and altitude, which was used to benchmark accomplishment of mission goals and constraints. Analyses were not run on performance measurements because each scenario contained a different number of objectives, and failure to meet one objective caused a greater chance of failure in successfully completing the following task. Additionally, periodic technical failure compromised the data collected, making correlation between workload and successful task completion implausible.

Scenarios. Three, 20-minute scenarios were designed (1A, 2A, 3A) using combinations of elements from the RPA task list and corresponding workload indicators derived from SME interviews, with a different dynamic in fluctuation of predicted workload magnitude throughout each scenario. Predicted workload within each scenario increased as more difficult tasks were introduced, time parameters were restricted, and more obstacles interfered with successfully accomplishing the mission (i.e., smoke from smoke stacks, target entering skyscrapers, decoys within the convoy). Scenario 1 was developed with the lowest cumulative level of workload for both pilot and SO, with predicted workload gradually increasing with each new task. Scenario 2 was designed to start with low workload, sharply increase, and then subside during the final portion of the mission. Scenario 3 contained the highest cumulative level of predicted workload, with three spikes in workload intensity throughout the mission. An alternate version of each of these scenarios was created with the intention of duplicating the pattern of workload dynamic within each scenario (1B, 2B, 3B). Each Pilot/SO team was given one scenario series (1A–3A or 1B–3B) during the first session, and when they returned for the second session, the pilot would take the role of the SO and the SO would become the pilot. Having switched positions, they were given the other series (A or B) in the opposite order, mirroring the first session. Both series were administered alternately during afternoon and morning sessions, and scenarios within each series were arranged in different order combinations for each team to account for order effects. Brief descriptions of the scenarios are below in Table 2.

Table 2. PRINCE Scenario Descriptions

Scenario	Pilot Cumulative Complexity	Description
1A/1B	15	Locate waypoint coordinates, loiter, locate POI within time parameter, track slow moving POI, report final destination.
2A/2B	20	Locate waypoint coordinates within time parameter. Avoid No Fly Zone and SAMs in route to new coordinates near Djibouti. Locate/track fast moving POI through visual obstructions, change POI-locate.
3A/3B	24	Locate POI within time parameter, avoiding restricted airspace in route, track fast moving POI through visual obstructions and decoys. Emergency mission to save Blue Force troops in eminent danger, forced to fly outside safety parameters to meet objective.

Note. Pilot Cumulative Complexity consisted of the sum of the complexity scores of each RPA task, on a scale of 1-5, within the scenario.

Subjective ratings. Subjective measures of workload were captured via self-report ratings. These were measured via a modified version of NASA-TLX and Multiple Resource Questionnaires (MRQ). The NASA-TLX assessed workload on six, seven-point scales and asked participants to rate estimates of perceived workload on a sliding scale of 21 gradations from 'very low' to 'very high'. NASA-TLX draws from multiple dimensions to construct weighted average ratings of mental demand, physical demand, temporal demand, performance, effort and frustration levels, and has been used in a variety of fields. The original questionnaire was modified only by eliminating examples from the description of each of these dimensions, thus abbreviating the scale, while maintaining the integrity of each dimension. Repeated measures for reliability have shown a correlation of .77 (Battiste & Bortolussi, 1988). The MRQ measured workload within a specific mental process across seventeen items; participants were asked to rate the extent to which a process was used ranging from 0 (no usage) to 100 (extreme usage). Self-report measures are notoriously variable, and perceived workload also varies across participants. Sometimes, participants rate their workload as high but observable behavior and/or performance is contrary, and vice versa. For example, a subject under extremely high demand may mitigate workload by "shedding tasks, lowering their performance standards, or refusing to exert greater and greater levels of effort…beyond a certain level" (Hart & Staveland, 1988). This potential for subjective ratings to mask or counter true experiences of cognitive workload lends further support for the incorporation of physiological data as alternate measures of cognitive workload.

Procedure

Each participant completed a demographics questionnaire requesting information about background and experience in order to analyze demographic trends, but there was an insufficient number of participants to draw any significant conclusions. Next, participants were fitted with the B-Alert wireless head-mounted data recording device. The device was secured over the head of the participant using a liberal amount of conducting gel applied to each non-invasive electrode pad to ensure the electrodes made good contact with the scalp. Electrode impedance was tested using the B-Alert software to confirm information was being transmitted from each electrode. Because we had two different models of the B-Alert device, we opted for the X-24 for the pilot participants and the X-10 for the SO participants. The pilot participant was fitted with the B-Alert X-24. The X-24 software included a 15-minute cognitive baseline task designed to account for individual differences between each participant with regard to cognitive state and workload metrics. The SO participant was fitted with the B-Alert X-10 and did not receive a baseline cognitive workload assessment. The Neurobridge system used for SO data collection did not provide a cognitive baseline task. Electrode impedance was tested for both B-Alert systems using the B-Alert software to ensure the head-mounted unit was adequately tracking EEG metrics of the participants and that they were synchronized correctly. Next, participants were given instructions in the form of a mission brief from the Mission Director about the training tasks. The mission brief contained an overview of the session, along with details about the task, and was provided before the start of each session. Each scenario lasted approximately 20 minutes. At the end of the session, participants were debriefed, asked about overall reactions as a team (pilot and SO) and were paid for their participation.

Statistical Design and Analyses

Between-subjects multivariate analyses of variance (ANOVAs) were performed using SPSS to compare subjective ratings of perceived workload based on scenario. Bivariate correlations were conducted to examine relationships between participant ratings and demographics. Correlations were also run between performance measures to determine if there was a significant relationship between scenarios and success in adhering to the mission objectives, and univariate ANOVAs were conducted to determine if there were significant differences in performance between scenarios. Physiological workload data was analyzed using between-subjects multivariate ANOVAs. Scenarios were developed to reflect varying levels of workload such that scenarios 1A and 1B produced the least workload, progressing to scenarios 3A and 3B, which were designed to produce the highest workload, and scenarios 2A and 2B producing a medium level of workload. Specific hypotheses were not made since this was a preliminary evaluation and scenarios were not previously validated. All data were therefore analyzed in an exploratory fashion.

RESULTS

Subjective Measures

The NASA-TLX survey data was analyzed in a 2 x 6 x 6 (Position x Scenario x TLX Categories) between-subjects design. A summary of the results can be found in Tables 3 and 4 and Figures 1 and 2 below. A between-subjects

ANOVA found that the SO position was rated significantly higher in overall frustration than the Pilot position, *F*(1,56) = 3.93, *p* < .05. A 3 x 6 between-subjects ANOVA comparing scenarios 1–3 by workload categories showed that the Scenario 3 produced significantly higher temporal demand, *F*(2,65) = 4.09, *p* < .05. All six individual scenarios and TLX categories were then analyzed by session number in a 6 x 6 x 2 ANOVA, showing a significantly higher rating of performance success for Scenario 2A, (*M* = 11.38, SD = 4.22), *F*(5,56) = 2.45, *p* < .05, with the lowest ratings in performance success for Scenarios 2B (*M* = 6.67, SD = 3.39) and 3B (*M* = 6.67, SD = 3.09).

Two-tailed independent t-tests were conducted on MRQ workload measures to determine if there were differences between Pilot and SO in perceptions of workload. SOs perceived significantly greater overall workload for several processes (see Tables 3 and 4 below).

Table 3. Significant MRQ Cognitive Workload Results by Position

MRQ Process	Scenario	Pilot *M (SD)*	SO *M (SD)*	df	*t*	*p*
Spatial Attentive[1]	Pilot-3B, SO-1A	79.7 (20.5)	89.7 (11.5)	66	-2.48	.016
Spatial Emergent[2]	Pilot-3B, SO-1A	64.1 (25.4)	80.1 (22.1)	66	-2.76	.007
Spatial Positional[3]	Pilot-3B, SO-2A	60.4 (25.8)	76.3 (22.6)	66	-2.70	.009
Vocal[4]	Pilot-1A, SO-3B	51.1(28.6)	67.1 (24.3)	66	-2.48	.016

Note. 1 = Visual attention on location. 2 = Visually picking out an object from a highly cluttered background. 3 = Visually differentiating a precise location from other locations. 4 = Use of voice.

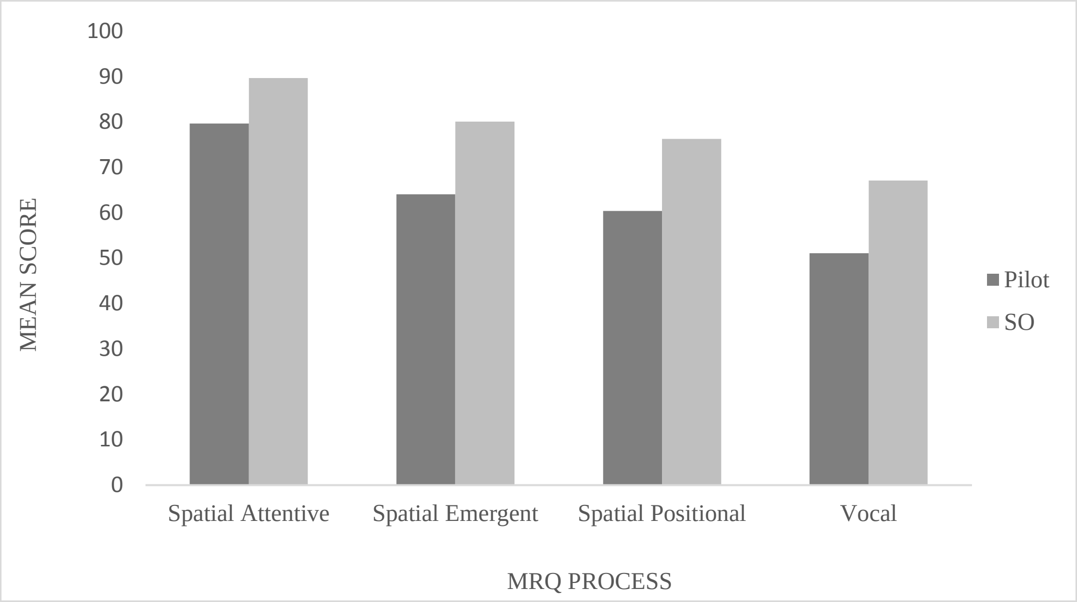

Figure 1. MRQ Results by RPA Role

Table 4. Significant NASA-TLX Cognitive Workload Results

NASA-TLX Process	Least	Greatest	df	*F*	*p*
Temporal Demand (Series)	1	3	2	4.09	.021
Performance Success	3B, 2B	2A	5	2.45	.045

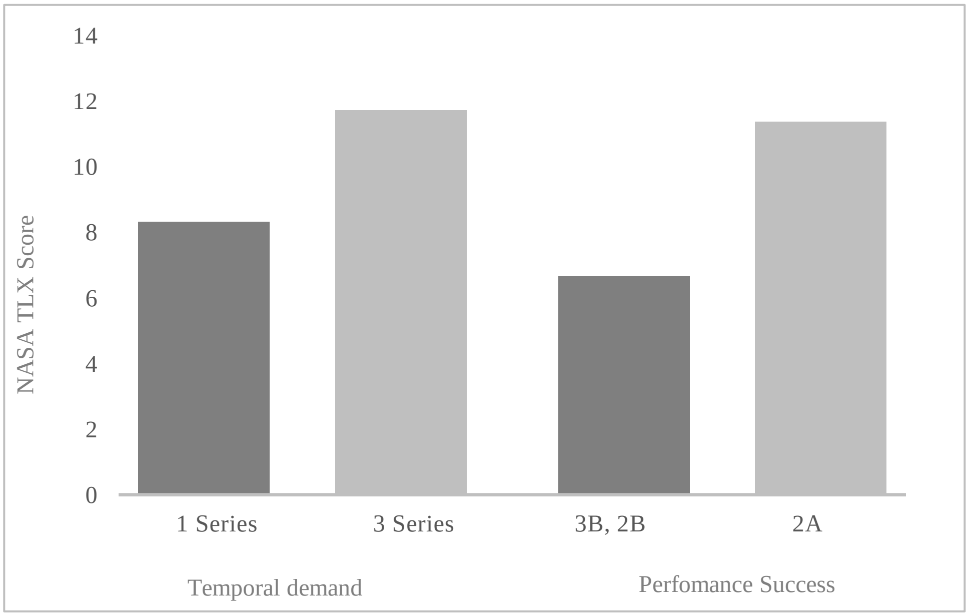

Figure 2. TLX Results by Scenario

Physiological Measures

An assumption of this evaluation is that increasing cognitive workload during a scenario can elicit a physiological response. Workload is typified by multiple, sometimes competing, tasks for cognitive attention. Multitasking requires contributions from prefrontal cortical regions that control attention functions. Generally, as mental workload increases, theta increases in the frontal lobe and alpha decreases in the parietal lobe. Additionally, alpha waves disappear and are replaced by beta waves (Sabbatini, 1997). In EEG studies, growing task demands, as well as time-on-task demands, increase frontal theta activity and decrease parietal alpha activity (Holm et al., 2009). EEGs can potentially be used to identify changes in cognitive load, typically characterized by spikes in waveforms and increases in numeric values, for cognitively demanding tasks such as operating an unmanned aerial system (UAS). However, Berka and associates (2007) found that the former method of cognitive state assessment using simple combinations of alpha and theta band activity from various brain regions to determine association with decision making, motor control, and visuoperceptual demands was an oversimplification of cognitive state assessment. In an effort to avoid misclassification of brain activity, variables sensitive to sleep deprivation were incorporated into a more intricate model of computing cognitive state assessment, which included workload, engagement, and drowsiness classifiers derived from complex combinations of EEG variables. These three classifications measures were used in this study to determine High Engagement Workload (HiEng, a measure of choice vigilance), Low Engagement Workload (LoEng, a measure of auditory psychomotor vigilance), Distraction, and Sleep Onset. The majority of the population (85%) is most sensitive to the Forward/Backward Digital-span (FBDS) measure of workload, the cognitive workload measure used for data analysis (Advanced Brain Monitoring, 2009). The following results are derived from data collected from the Pilots on the B-Alert software during the second session of data collection only. A summary of the findings can be found in Table 4 and Figures 3 and 4.

A 6 x 5 (Scenario x Engagement Metrics) ANOVA showed a significant difference between scenarios for all four cognitive workload metrics. Scenario 3B, (*M* = 55.8, *SD* = 51.2) had the highest mean for HiEng, *F*(6,449) = 61.10, *p* < .001, and for LoEng, *F*(6,449) = 26.78, *p* < .001. This indicates that both measures of vigilance were highest for this scenario. Scenario 1A, (*M* = -0.11, *SD* = 0.39), had the lowest mean for HiEng, and 1B, (*M* = 0.05, *SD* = .22), had the lowest mean for LoEng, showing Scenario 1A demanded the least auditory psychomotor vigilance while Scenario 1B demanded the lowest decision-making vigilance. Significant effect was also evident for Distraction, *F*(6,449) = 27.95, *p* < .001, and for SleepOnset, *F*(5,449) = 6.17, *p* < .001, with Scenario 2A having the highest distraction rate (*M* = 23.71, *SD* = 34.58), and 3B having the lowest distraction rate (*M* = -1.93, *SD* = 2.84), suggesting that participants were most distracted during Scenario 2A and least distracted during Scenario 3B. Scenario 3B had the lowest mean for Sleep Onset (*M* = 378.52, *SD* = 1200.56), and Scenario 3A had the highest mean for sleep Onset (*M* = 25674.49, *SD* = 66997.05). This reveals that Scenario 3B produced the least monotony or boredom while 3A showed the highest. Task analysis showed significant differences between tasks for Distraction, *F*(7,449) = 5.32, *p* <.001, with the highest mean for searching, (*M* = 13.78, *SD* = 28.74) and the lowest mean for tracking (*M* = .87, SD = 6.38). This suggests that searching tasks generated a significantly higher level of distraction while tracking the point of interest generated the least distraction.

A 6 x 5 (Scenario x Cognitive Workload Metrics) ANOVA showed significantly greater Forward/Backward Digital-Span (FBDS) Workload for 3A (*M* = 0.24, *SD* = 0.62) by Scenario, *F*(6,449)=8.65, *p* < .001, and the lowest mean for 2A, (*M* = -.29, *SD* = .54). This indicates that the 3-Series scenarios generated the greatest overall workload while Scenario 2A generated the least overall workload. Heart Rate was also significant by Scenario, *F*(6,449) = 14.21, *p* < .001, with Scenario 1A (*M* = 0.67, *SD* = .39) showing the highest heart rate and Scenario 2B (*M* = .06, *SD* = .28) with the lowest heart rate. This indicates that the heart rate responded with greater intensity to Scenarios 1A and 3A. It has been considered that this anomaly, in relative predicted heart rate for Scenario 1A, is due to the vast amount of missing data for this particular measure of workload caused by complications during portions of data collection, as Scenario 1A (*n* = 31) had less than 50% of the average data samples for all scenarios (*n* = 75.7) for this measure, with all other scenarios having no less than 60 data samples.

Table 4. Summary of B-Alert Workload Results by Scenario

Workload Metric	Highest Workload	*M (SD)*	df	*F*	*p*	95% C.I.	
						Lower	Upper
High Eng	3B, *M=55.8*	13.87 (32.68)	5	73.49	.000	8.50	13.42
Low Eng	3B, *M=6.7*	1.85 (4.97)	5	32.17	.000	1.22	2.09
Distraction	2A, *M=23.7*	5.77 (19.15)	5	33.61	.000	3.87	7.19
Sleep Onset	3A, *M=25674.5*	5396.45 (29492.0)	5	7.42	.000	3052.69	8806.8
FBDS	3A, *M=.25*	-.06 (.61)	5	9.11	.000	-.09	.03

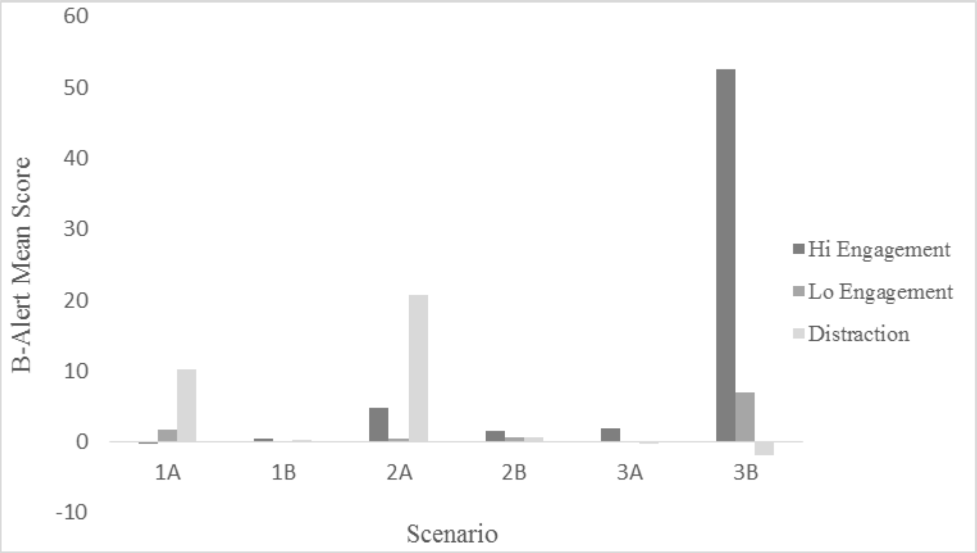

Figure 3. Workload Elements by Scenario

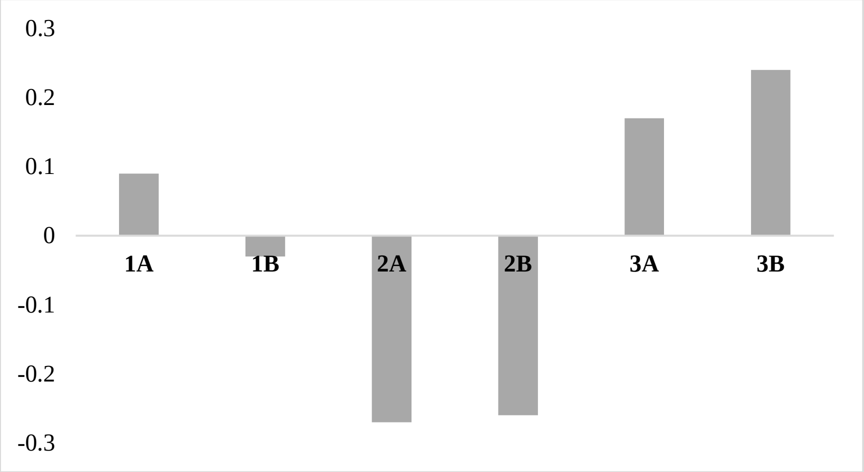

Figure 4. FBDS Workload Means by Scenario

A comparison of all of the significant measures of workload unanimously revealed that Scenario 3B had the highest cognitive demand and mental engagement and lowest performance success, distraction, and sleep onset, as expected. The scenarios shown to be lowest in cognitive demand and lowest in high mental engagement (decision-making vigilance) were 1B and 2A, and highest significant distraction and performance success was 1A. These results were congruent with the prediction that the lower scenario numbers would be less cognitively demanding than the higher numbered scenarios. Subjective results did show Scenario 2A as highest in perceived performance success and lowest in perceived visual temporal demand rather than a 1- or 3-Series scenario, which would reflect a neutral result for those elements. The only significant anomaly to this prediction was Scenario 3A, which showed the highest physiological sleep onset and least low-engagement mental workload. It is possible that Scenario 3A surpassed the threshold of complexity for participants, resulting in diminished effort. For example, in the 3-Series scenarios, participants were required to locate a target within a specific time frame and track its movement. If that deadline was missed, the target would be lost and the participant was considered to have failed that portion of the mission. This failure resulted in a gap in tasking, with participants having to stand by for several minutes for the next task.

Conclusions

We believed that physiological data would consistently reflect higher participant cognitive workload for the more difficult scenarios. To provide a more comprehensive picture of workload, objective (performance, EEG), and subjective (survey) data were analyzed and compared. Through consistency of several methods of measurement in resulting workload outcomes, the analysis revealed that it is possible to successfully capture and measure physiological changes in novice trainees when workload is manipulated in a synthetic learning environment. Brain activity captured throughout a series of RPA training tasks with varying levels of difficulty, as well as subjective survey data from participants immediately following each set of tasks, suggested that the projected workload was indeed measured by physiological measures.

Limitations

The current evaluation was limited by time, funding, and classification constraints, hence the measurement of workload with scenarios that had not been validated prior (or current operational training scenarios for PRINCE). We were also unable to use the same two B-Alert systems, which would be valuable in future research, especially for baselining data and ensuring the same number of data points are being captured per role. Future research using validated training scenarios, the same physiological monitoring systems for all participants, and effective performance metrics for comparison will allow for a more comprehensive understanding of the impact of workload on training effectiveness and consequences for designing and adapting training systems. Nevertheless, the current evaluation was a necessary first step to examine cognitive workload for an operational military synthetic training environment such as PRINCE for the first time.

DISCUSSION

In summary, physiological and subjective data both demonstrated that scenarios with the highest projected levels of cognitive workload had the lowest levels of distraction and sleep onset, as well as the greatest mental engagement and demand. These findings provide initial validation of the training scenarios and produced results consistent with the prediction that workload can be captured and accurately measured for tasks within the RPA skill set. This supports our first research goal of associating RPA scenario events with levels of workload. In general, our findings also support our second and third research goals of comparing the informative power of subjective and objective cognitive workload measures on the effectiveness of a synthetic learning environment like PRINCE and the human performance in these types of simulated training environments.

It should be noted that this report represents limited objective and subjective data from only a small sample of the RPA community and may or may not reflect actual workload demand. Rather than providing definitive answers to research questions, this data should serve as a foundation for future research linking training events with workload. These results could be used to target specific training objectives and measurement considerations in need of improvement or development for future cognitive workload studies using PRINCE or similar synthetic learning environments.

ACKNOWLEDGEMENTS

We would like to acknowledge the University of North Dakota School of Aerospace Sciences, Department of Aviation's Dr. John Bridewell, Mr. Zachary Waller and Mr. Scott Kroeber, for assistance with the UND IRB process, participant recruitment, access to their facility and support during the planning and data collection effort. We would also like to acknowledge Eric Watz and Peter Neubauer, who provided engineering support for this project.

REFERENCES

Advanced Brain Monitoring, Inc. (2009). B-Alert Live Software User Manual.

Afergan, D., Peck, E.M., Solovey, E.T., Jenkins, A., Hincks, S.W., Brown, E.T., Chang, R., & Jacob, R.J.K. (2014). Dynamic difficulty using brain metrics of workload. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3797-3806). ACM.

Aldrich, T. B., Szabo, S. M., & Bierbaum, C. R. (1989). *Applications of human performance models to system design* (Vol. 2, Defense Research). New York: Plenum Press. doi:10.1007/978-1-4757-9244-7_5

Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage, 59,* 36-47.

Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage, 59,* 36-47.

Battiste, V., & Bortolussi, M. (1988, October). Transport pilot workload: A comparison of two subjective techniques. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 32, No. 2, pp. 150-154). SAGE Publications

Berka, C., Levendowski, D. J., Ramsey, C. K., Davis, G., Lumicao, M. N., Stanney, K. (2005). Evaluation of an EEG workload model in an Aegis simulation environment. *Biomonitoring for Physiological and Cognitive Performance during Military Operations*. doi:10.1117/12.598555

Berka, C., Levendowski, D., Lumicao, M., Yau, A., Davis, G., Zivkovic, V., Olmstead, R., Tremoulet, P., Craven, P. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine, 78*, 231-244.

Craven, P.L., Belov, N., Tremoulet, P., Thomas, M., Berka, C., Levendowski, D., & Davis, G. (2007). *Cognitive workload gauge development: Comparison of real-time classification methods.* Retrieved from http://www.atl.lmco.com/papers/1379.pdf on 10/27/2015

Fournier, L. R., Wilson, G. F., & Swain, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task difficulty and training. *International Journal of Psychophysiology, 31,* 129-145.

Hart, S. & Staveland, L., Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, 52, 139-153, doi:10.1016S0166-4115(08)62386-9.

Hancock, P. A., & Szalma, J. L. (Eds.). (2008). *Performance under stress*. Aldershot, England: Ashgate.

Horst, R.L., Mahaffey, D.L., & Munson, R.C. (1989). Brain-wave measures of workload in advanced cockpits: The transition of technology from laboratory to cockpit simulator, (*NASA Contractor Final Report 4240 Document ID: 19890015426*). Retrieved from http://ntrs.nasa.gov

Lai, G., & Lamoureux, T. (2012). *Development of measures of effectiveness and performance from cognitive work analysis products*. Ottawa, ON.

Moore, J. J., Ivie, R., Gledhill, T. J., Mercer, E., & Goodrich, M. A. (2014, March). Modeling human workload in unmanned aerial systems. In *AAAI Spring Symposium Series: Formal Verification and Modeling in Human-Machine Systems* (pp. 44-49).

Moray, N., Dessouky, M., Kijowski, B.., & Adapathya, R. (1991). Strategic behavior, workload, and performance in task scheduling. *Human Factors, 33*, 607–629.

Sanders, M.S., McCormick, E.J. (1987). *Human factors in engineering and design*. New York, NY: McGraw-Hill.

Shriram, R., Sundhararaajan, M., Daimiwal, N., (2013). EEG based cognitive workload assessment for maximum efficiency. *IOSR Journal of Electronics and Communication Engineering*, pp 34-38. Retrieved from http://www.iosrjournals.org on 4/18/2016.

St. John, M., Kobus, D. A., Morrison, J. G., & Schmorrow, D. (2004). Overview of the DARPA augmented cognition technical integration experiment. *International Journal of Human-Computer Interaction*, 17, 131-149.

USAF, Secretary of the Air Force. (2010). *MQ-1B remotely piloted aircraft.* Retrieved from: USAF.

Vogt, J., Hagemann, T., Kastner, M. (2006). The impact of workload on heart rate and blood pressure in en-route and tower air traffic control. *Journal of Psychophysiology*, *20*(4), 297-314.

Weiland, M. (2009). *Workload measurement: Peeking into the brains of system operators*. Retrieved from http://www.mitre.org/publications/project-stories/workload-measurement-peeking-into-the-brains-of-system-operators

Wetherell, M. A., & Carter, K. (2014). *The multitasking framework: the effects of increasing workload on acute psychobiological stress reactivity. Stress and Health, 30*(2), 103-109.

Wickens, C. D. (2002). Situation awareness and workload in aviation. *Current directions in psychological science, 11*(4), 128-133.