Real Time Workload Detection in Supervisory Control Applications Using fNIRS

Delaney Lagrew

Advisor: Dr. Mary Cummings

Duke University*

http://hal.pratt.duke.edu
*Duke University Department of Neuroscience, Durham, NC 27708

Abstract:

With increasing use of automation in complex system operations like air traffic or nuclear plant control, the role of human operators is increasingly shifting from hands-on control to remote supervision, known as human supervisory control. During a human supervisory control task there may be sudden changes in taskload. When critical events occur, operators can be overloaded by sudden increases in workload, to the point of making errors. The ability to remotely detect a neurophysiological state that is likely to lead to problematic human performance is especially useful for supervisors of such systems because it allows us to detect when an operator may be bored or overwhelmed by the taskload presented. By combining the measured hemodynamic and metabolic responses of indirect brain and cognitive activity using functional Near-Infrared Spectroscopy (fNIRS), it may be possible to determine a change in a subject's cognitive activity while performing different tasks. Whether such data can be reliably connected to workload and performance in an actual supervisory control setting, where most other neurophysiological efforts have failed, is an open question. Using fNIRS in a human supervisory control experiment (in this case supervising multiple unmanned vehicles), we examined whether measured hemodynamic fNIRS data correlates to an actual change in low and high workload in supervisory control settings. Participants (n=36) controlled unmanned vehicles in low and high mental workload situations. Overall performance scores were significantly higher in low workload scenarios compared to high workload scenarios. However, there were no significant differences in HbO percent change or HbR percent change based on scenario, which indicates that fNIRS may not be suitable for detecting mental workload changes over short periods of time.

TABLE OF CONTENTS

1. Introduction

1.1 Research Question

As technology progresses in society, human operators are being removed from the role of direct controller to take on the responsibility of "supervisory controller" or "system administrator." A supervisory controller is someone who intermittently interacts with a computer interface to complete a task. During human-automation interaction, supervisory controllers are faced with periods of low and high workload. As technology continues to increase, the periods of low workload continue to grow. It is important to understand when a supervisory controller is subjected to high or low mental workload in order to optimize performance of the operator during a human supervisory control task.

This investigation aims to expand the understanding of how a neurophysiological tool, functional near-infrared spectroscopy (fNIRS), can be used to reliably measure mental workload during periods of low and high cognitive task demands. Mental workload is important because it is distinguishable from other cognitive constructs such as attention or memory (Parasuraman et al., 2008). The study of mental workload aims to improve understanding of human interactions with machines in complex environments, such as air traffic control, nuclear power plant operations, military unmanned aerial vehicles, and robotic manufacturing. The studies of this experiment are aimed at elucidating the relationship between measures of brain function and workload in humans using a combination of established workload measures.

Numerous techniques are available for measuring workload in applied and experimental settings, including subjective, performance-based, and physiological measures. From the latter category, neurophysiological measures are a potential method for objectively viewing mental activity at different levels of workload to achieve a better understanding of how humans respond to changes in task loading. This research focuses on understanding the cognitive and neural mechanisms underlying workload, using a combination of subjective measures of individual differences (e.g., self-report questionnaires), behavioral measures obtained from a supervisory control simulation (e.g., response time, accuracy), and non-invasive neurophysiological measures of oxygenated hemoglobin and deoxygenated hemoglobin derived from scalp-recorded fNIRS. More specifically, this research is focused on determining the effectiveness of fNIRS as a tool for measuring workload during task performance.

The following sections will provide background on mental workload, including various techniques for measuring workload, followed by motivation for an experiment that compares performance-based workload measures to measures of cognitive arousal using fNIRS.

1.2 Mental Workload

Early workload research began with Yerkes and Dodson at the beginning of the 20th century. They determined that humans performed greatest at a medium mental workload levels and decreased in performance during very low and very high levels during a training task. This inverse U relationship can be seen in their Yerkes-Dodson Curve (1908; Figure 1). In a more modern version, Hancock and Warm (1989) were able to recreate a similar inverse U during training for a learning task that showed a relationship exists between attentional resource capacity and stress level. They confirmed that as humans are performing tasks, it is important that arousal levels are not too high because a decrease in performance can lead to mistakes while performing a task.

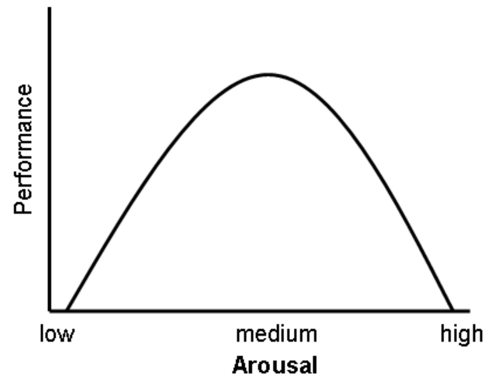

Figure 1: Yerkes-Dodson Arousal vs. Performance Curve

In the mid 1970s, Moray gave the first concrete definition of mental workload:

> A load is something which imposes a burden on a structure, or makes it approach the limit of its performance in some dimension. Go far enough along that dimension and the system will fail in some way. In the case of mental workload, the central concept is the rate at which information is processed by the human operator, and basically the rate at which decisions are made and the difficulty of making the decisions. (Moray, 1979)

Many other definitions have been provided over the years but still follow a similar theme. The main idea is that the operator regulates a limited pool of mental resources to assign to task demands. If task demands are within our resource limits, performance will not be hindered. However, if task demands exceed resources, performance may suffer (O'Donnell & Eggemeier, 1986).

1.3 Taskload and Workload

When discussing mental workload, an important distinction must be made between *taskload* and *workload*. Taskload is the measure of the actions and procedures required of any operator to execute a task and is subject-independent. Workload, in contrast, is the operator's assignment of mental resources to taskload demands. An example of this is two unmanned vehicle operators each responsible for controlling the same number of unmanned aerial vehicles. These two operators may report different levels of workload due to differences in individual training, experience, skill and fatigue, but have the same taskload because they are both controlling the same number of aircraft.

Wickens' model of information processing provides a system chart for demonstrating the transformation of data from the surrounding environment into a response (Figure 2: Wickens & Hollands, 1999). Humans receive stimuli through sensory organs (e.g., nose, eyes, ears), process those signals into functional information through perception (e.g., motion, speech), access their working and long-term memory to better understand the information, and finally use their decision and response selection centers to generate and execute a response. A limited pool of attention resources is used to prioritize the various nodes between perception and response execution. Understanding the neural correlates to a conceptual model of information processing is a growing trend and one of the underpinnings of the present research.

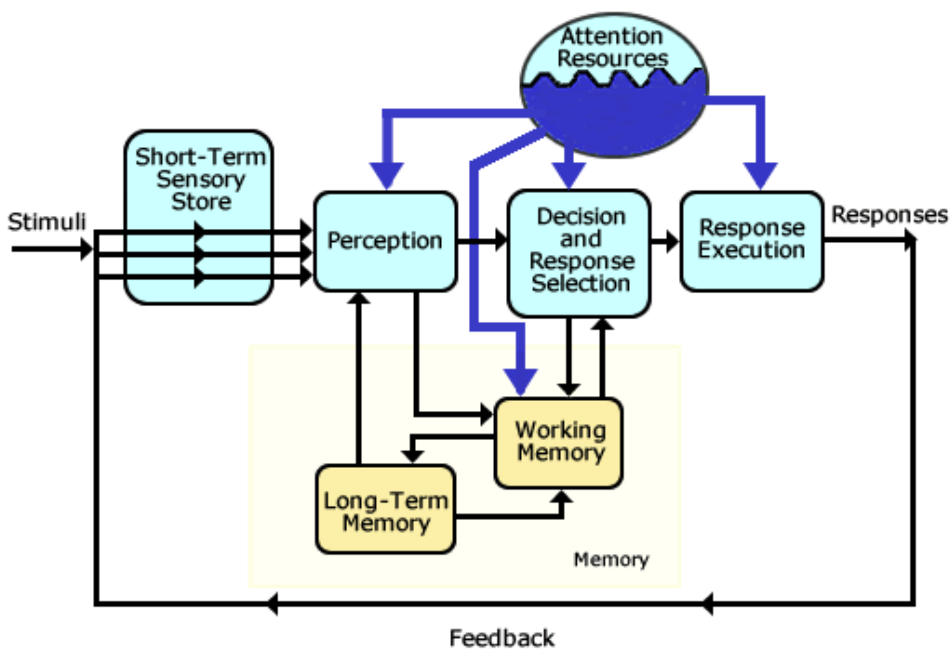

Figure 2: Model of Human Information Processing

Human mental workload has been studied extensively. In the aviation domain, numerous studies have examined military pilots, airline pilots, and air traffic controllers (Battiste & Bortolussi, 1988; Wilson, 2002). In addition to aviation, there have also been a plethora of high workload studies that deal with specialized occupations such as surgeons, astronauts, and missile defense operators (Klein, Riley, Warm, & Matthews, 2005; Berka et al., 2005). The common theme of the previous works is that these tasks deal with high workload in critical situations. Endsley and Rodgers (1997) were able to summarize the ideas that come from all these high workload situations using operational errors in en route air traffic control. They found that there was a positive correlation between workload and operational errors in the high-workload domain. This can be put in simpler terms by saying that as the workload increases, the number of errors that the user makes also will increase.

1.4 Workload Measurement

In order to perform the research on workload discussed in the previous section, it is necessary to identify measures that accurately describe workload levels. The four main methods of measuring workload are (1) user performance, (2) surveys, (3) physiological measures, (4) neurophysiological measures.

The first method of measuring workload is looking at user performance. As Wierwille and Eggemeier (1993) showed, measures of speed and accuracy can be used to represent an objective task performance metric. The use of task performance to measure workload requires the assumption that speed and/or accuracy of performance will decrease as workload increases beyond a critical value or threshold for unimpaired performance. The main drawback to this idea occurs when the user consistently has excess resources available to keep task performance at a high level. This could occur during tasks where the user is asked to change from a low workload task to a medium workload task and can easily offset the increases in task demands. The issue here is that performance-based measures have been shown to be effective at high workload levels, but they may not be reliable when the operator is performing well at low or medium workload levels. Therefore, we need a measuring technique that is effective across a broader array of workload levels, including medium and low.

The second method of measuring workload uses subjective data captured using workload surveys. Some of the most popular surveys include the Cooper-Harper Rating Scale (Cooper & Harper Jr., 1969), the NASA Task Load Index (TLX; Hart & Staveland, 1988) and the Subjective Work Index Test (SWAT; Reid & Nygren, 1988). These surveys incorporate a series of questions to measure workload. The main advantage of these surveys is that they are easy to complete and can be compared to previously completed research. The limiting factor in these surveys is that they fail to specifically pinpoint areas of high or low workload fluctuations throughout the task.

The third method of measuring workload during a task is physiological measures. There are many physiological measures that have been shown to reliably predict mental workload (Kramer, 1991). Some of these measures, such as heart rate or blood pressure, can be directly traced to an increase demand by the brain, but many are responses that are merely correlated with increased mental workload. Cardiovascular measures are some of the most commonly-used methods for tracking workload over time. Electrocardiogram (ECG or EKG) measures heart rate, heart rate variability, blood pressure, and blood volume. Generally, heart rate, heart rate variability, and blood pressure all increase during periods of high mental workload (Sirevaag et al., 1993). There are also several different measures of eye activity that are associated with mental workload. Pupil dilation, for example, has been found to be a good measure of workload, with increased dilation occurring during periods of high workload

(Beatty, 1982). The third type of physiological workload measurement is through symptoms of the sympathetic nervous system (SNS), which is part of the autonomic nervous system (ANS). The SNS is commonly associated with the “fight-or-flight” response and stimulates many systems in the body when activated. One of the most commonly-used measures is galvanic skin response, which measures sweat produced in certain regions of the skin. Galvanic skin response has been associated with mental workload in several different environments (Wierwille, 1979) and is a relatively low-intrusive technique.

A fourth approach to measuring workload throughout an experiment or task is physiological tracking. Physiological tracking allows for continuous monitoring of subject state, whereas many primary and secondary task measures can only measure the subject’s state at discreet event times. Research by many such as Koechlin, Basso, Pietrini, Pazner, & Grafman (1999) and Miller & Cohen (2001) have shown the firing of neurons to produce electrical signals consumes oxygen and glucose and gives off carbon dioxide as byproducts. Any of these features (electrical signals, oxygen, glucose, etc.) can be used as a method to measure workload.

Neurophysiological measurement is another possible way to objectively measure mental workload. Other forms of workload measurement rely on subjective input from participants, may not be sensitive at low or moderate workload levels, or are only limited to measuring workload at isolated times. The continuous ability to measure workload through direct contact with the scalp of the head makes neurophysiological tracking the most ideal tool to measure workload. The next section will go into further detail about neurophysiological tracking, specifically neurocognitive measurement.

1.5 Noninvasive Neurocognitive Measurement

The main focus of this research is looking at noninvasive neurocognitive workload measurement. Noninvasive neurocognitive measurement, as opposed to invasive measurement, does not require inserting sharp probes into the body. When investigating cognitive behavior, invasive techniques are not a practical approach because they pose a higher risk to the subject and have a very high cost. Therefore, noninvasive techniques are widely used in cognitive neuroscience because of their lower cost, lower risk, and higher reliability in healthy subjects. This section will give a brief overview of the noninvasive techniques that are used in cognitive neuroscience research.

The three main stages of brain activity indicators are (1) supply (what is delivered to the brain), (2) electrical signals (neuron firing), and (3) by-products (waste that is removed) (Pasley & Freeman, 2008). These stages of brain activity are ways in which researchers can successfully track what is happening in the brain.

When looking at the *supply* stage, the two main ingredients required for neuronal activity are glucose and oxygen. Glucose uptake can be measured using Positron Emission Tomography, or PET scan, which uses a radioactive tracer that is analogous to glucose (Buckner & Logan, 2001). Oxygen is able to travel to the brain via hemoglobin and transcranial Doppler sonography, or TCD, measures the velocity of the blood flow to the brain, or cerebral hemovelocity (Warm et al., 2009).

Electrical signals are the most direct method of measuring brain activity because the supply (oxygen and glucose) and byproducts are simply a maintenance mechanism. When normal levels of glucose and oxygen are available, the firing of the electrical signals occur from neuron to neuron. When there is higher mental workload, the nucleus to the neuron consumes larger amount of glucose and oxygen to produce more electrical signals.

Electroencephalography (EEG) decomposes the complex EEG waveform into its constituent frequency bands and quantifies the energy in each band. The standard EEG bands are typically used, which include delta, theta, alpha, beta, and gamma. Increased workload appears to be associated with decreased activity in the alpha band but increased activity in the theta band. Beta and gamma band activity may also increase under conditions of higher mental demand. (Wilson & Eggemeier, 1991) The EEG can provide high temporal resolution (precision of a measurement with respect to time), but very low spatial resolution (distances between different brain regions). Whereas EEG records the electrical activity associated with neuronal depolarization oriented perpendicular to the surface of the brain, the technique of Magnetoencephalography **(**MEG) records the magnetic field produced by this electrical activity oriented parallel to the surface of the brain. Both the MEG and the EEG struggle with noisy signals caused by electrical signals other than the brain (e.g., head movement, blinking, etc.). Fortunately, EEG is widely used enough that many algorithms have been created to filter out some of these noise artifacts.

After supplies have been delivered and the neuron has fired, the third stage of brain activity involves the removal of *by-products* comes in the form of deoxygenated hemoglobin carried away in the bloodstream. This deoxygenated hemoglobin can be measured through its magnetic properties. Functional Magnetic Resonance Imaging (fMRI) takes advantage of these magnetic properties to measure the concentration of deoxygenated hemoglobin with great spatial resolution (Bucker & Logan, 2001; Carr, Rissman, & Wagner, 2010). This spatial resolution is key in seeing which specific areas of the brain are being activated during a task. In recent years there has been a rapid increase in the popularity of fMRI studies. In 1993, the number of published articles citing functional magnetic resonance imaging was fewer than 20. In 2003, that number had increased to almost 1,800 (Berman, Jonides, & Nee, 2006). The rise in popularity comes from fMRI's ability to measure brain activity during various tasks to show activation in different regions of the brain (Cabeza & Nyberg, 2000; Bunge, Gross, & Gabrieli, 2002) and its capacity to map the entire brain at once to show brain regions that work in conjunction with one another (Monchi, Petrides, Petre, Worsley, & Dagher, 2001). fMRI has excellent spatial resolution ($\leq 10mm^3$), but does have some drawbacks including: limited temporal resolution, expensive operation costs, and their noisy, unnatural environment. Subjects have to lie down in a confined space where no metal objects are allowed, which severely limits which tasks can be performed.

Another example of a noninvasive neurocognitive measurement looking at oxygen is functional near infrared spectroscopy (fNIRS). fNIRS will be discussed in much greater detail in the next section, but simply put fNIRS measures both oxygenated and deoxygenated hemoglobin through the absorption of infrared lighting.

As the previous section shows, there are many different techniques to measure workload non-invasively. Each device has its own advantages and drawbacks depending on what the researcher is trying to accomplish. fNIRS is promising for workload research because it allows the user to test in environments that mimic real working conditions. However, its accuracy in predicting workload changes still remains unknown and requires further investigation. Figure 3 below is a visual representation of the different neurocognitive measurement tools with spatial resolution vs. temporal resolution. With moderate temporal and spatial resolution (compared to EEG and fMRI), the biggest advantage of fNIRS as a brain imaging device is its ability to be used outside of tightly controlled research environments. Subjects are not forced sit in an fMRI machine or avoid blinking during an EEG experiment, so the data that is gathered does not need to take these factors into consideration when drawing conclusions from the results.

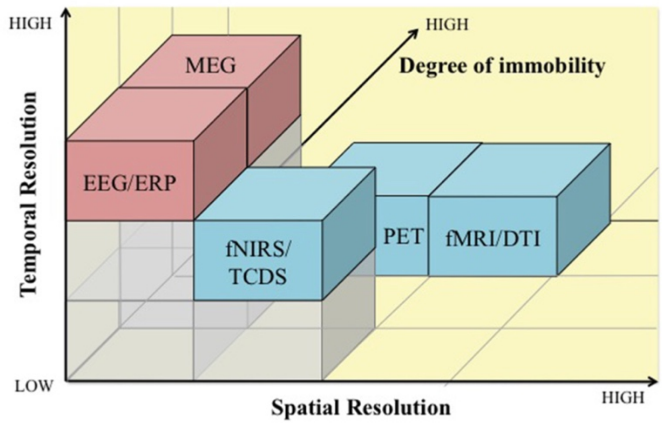

Figure 3: Spatial and temporal resolution of neuroimaging devices (Mehta & Parasuraman, 2013)

1.6 Neurophysiology and functional Near-Infrared Spectroscopy (fNIRS)

At the root of all hemodynamic-dependent studies, such as fNIRS and fMRI, is the blood oxygen level-dependent (BOLD) signal. A quote from the 2011 Encyclopedia of Clinical Neuropsychology:

> BOLD imaging is a version of magnetic resonance imaging that depends on the different magnetic properties of oxygenated versus deoxygenated hemoglobin and, thus, indirectly, on variations in local tissue perfusion. The utility of BOLD imaging for fMRI also depends on the physiological phenomenon by which metabolically active cerebral tissue "demands" more perfusion than less-active tissue. Thus, populations of neurons that are particularly active during a cognitive or motor task actually elicit a surplus of perfusion which, in turn, results in an increase in the ratio of oxygenated to deoxygenated hemoglobin, detectable as a change in the BOLD signal. (Whyte, 2011)

It is generally acknowledged that an increase in neural activity in a certain region of the brain will demand greater blood flow in an attempt to supply more oxygenated hemoglobin, while removing the deoxygenated hemoglobin. Therefore, it must be understood that the blood oxygen level-dependent signal is an indirect measure of what researchers are really interested in looking at, neural activity.

In 1977 at Duke University, Franci Jöbsis reported that relatively high degree of brain tissue transparency in the near-infrared range enables real-time non-invasive detection of hemoglobin oxygenation using transillumination spectroscopy. Transillumination spectroscopy uses nonionizing optical radiation to gain information about tissue properties and is mainly used today as an assessment of breast cancer risk. (Jöbsis, 1977). Jöbsis used this technique to study cerebral oxygenation in sick newborn infants. In 1980, Marco Ferrari began using a prototype near-infrared spectroscopy instrument to measure changes in brain oxygenation in animals (Ferrari et al., 1980) and adult humans for the first time (Ferrari et al., 1982). Subsequent development and refinement of this technique to accurately measure oxygenated and deoxygenated hemoglobin in the brain and other regions of the body led to the first functional near-infrared spectroscopy.

fNIRS functions by injecting near-infrared light from lasers at certain wavelengths (typically 690nm and 830nm) into the region of the brain that the researcher is interested in by applying sensors to head. This infrared light is able to pass through both skin and bone to be absorbed by hemoglobin in the outer cortex of the brain; this can be seen in Figure 4. The increase in oxygenated hemoglobin and the associated decrease in deoxygenated hemoglobin reflects an increase in arteriolar vasodilation (widening of blood cells within arteries), which increases cerebral blood flow and cerebral blood volume. The increased oxygen transported to the area typically exceeds the need of the local neuronal rate of consumption, which causes an excess of blood oxygenation in active areas. Figure 5 shows the shows this increase in oxygenated hemoglobin (red line) coupled with the decrease in deoxygenated hemoglobin (blue line) during cortical activation, represented on the x-axis as Time 0 through 10. The total change in blood flow is represented by the green line.

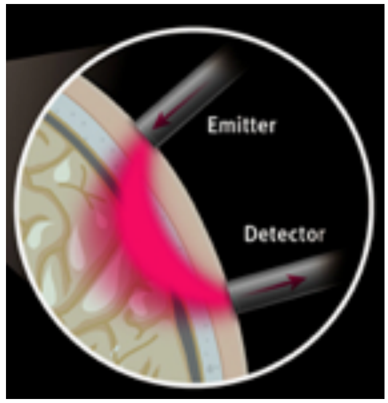

Figure 4: Infrared light traveling through skin and bone to reach the cortex (from ISS, Inc.)

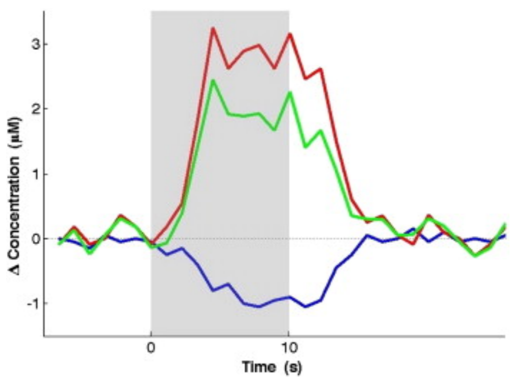

Figure 5: Oxygenated hemoglobin and deoxygenated hemoglobin during cortical activation (Ferrari & Quaresima, 2012)

After the infrared light leaves the cortex it returns to the detector, and is measured through photomultipliers. The photomultipliers are responsible for converting the light into digital signals for post-processing purposes. Post-processing of optical signal utilizes the Modified Beer-Lambert Law. The Modified Beer-Lambert is an algorithm that derives changes in tissue optical properties based on continuous-wave (CW) diffuse optical intensity measurements. In its simplest form, the scheme relates differential light transmission changes to differential changes in tissue absorption (Baker et al., 2014).

Figure 6 shows how the fNIRS sensor is applied to the forehead of a subject with corresponding detectors and light sources. This figure is shown using 8 channels. For fNIRS, a channel is a system in which several independent signals may be sent down an optical fiber link by monitoring them on light-carriers of different wavelengths. The infrared lasers are extremely sensitive to outside light channeling, which can compromise the signal. fNIRS also requires very firm contact with the subject's scalp, with may cause discomfort for participants due to compression on the scalp.

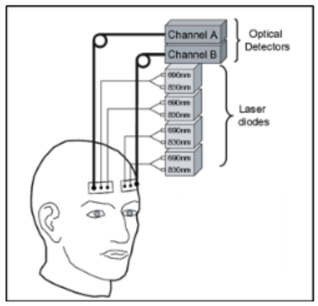

Figure 6: fNIRS sensor diagram for prefrontal cortex (from ISS, Inc.)

1.7 fNIRS and Workload

As a workload measuring device, fNIRS has been used in cognitive neuroscience in addition to other physiological measures of workload such as heart rate, blood pressure, galvanic skin response, and respiration. When looking at workload and how it is to be measured, the main area of the brain that will be focused on is the prefrontal cortex, which plays an important role in the processing of memory and the associated workload. Jelzow et al. (2011) showed the correlation between these physiological responses and fNIRS as a way to measure workload, but only used 15 participants during their study using a semantic continuous performance task (CPT). Findings from Herff et al. (2014) show that measuring hemodynamic responses in the prefrontal cortex with fNIRS can be used to quantify and qualify high mental workload during the *n*-back task (a test of working memory), again using only 10 participants. A study was conducted in a human supervisory control environment (Boyer et al., 2015), which did not detect workload changes with fNIRS using 30 participants. This study used a missile defense response task in a supervisory control setting. From this previous research, it can be seen that there are conflicting results. In previous fNIRS studies (e.g., Herff et al., 2014), it was determined that there was a correlation between fNIRS and changes in mental workload. However, these studies are limited due to the small number of participants and the simplicity of the mental workload tasks. Boyer's research takes a much different approach (increased number of participants and a much more complex workload task) and did not detect changes in mental workload using fNIRS. From this point it is important there is a conflict in previous results that must be further investigated.

1.8 Summary

The study of mental workload in human supervisory control settings is very important to our understanding of how humans are able to interact with technology. More specifically, it is important to understand when a user is not performing at an optimal level. fNIRS is a relatively new method of neurocognitive measurement that may be able to give insight into mental workload during human supervisory control tasks, but still remains unproven.

An experiment was conducted to measure workload while using fNIRS during a human supervisory control task. Participants performed a computer-based task under high and low workload conditions. Simulator task performance was compared to fNIRS output to identify how each techniques responds to changes in workload. It was hypothesized that subjects will have lower overall performance scores and accuracy in high workload scenarios compared to low workload scenarios. It was also hypothesized that there will be an accompanied larger increase in oxygenated hemoglobin (decrease in deoxygenated hemoglobin) during high workload scenarios compared to low workload scenarios.

2. Experimental Methods

2.1 Experimental Framework

This experiment employed a simulation designed to mimic the responsibilities of the Unmanned Vehicle (UV) operator. The Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) simulator was used as a test platform in the experiment. RESCHU has been used in previous studies to model workload impact in multiple unmanned vehicle supervisory control settings (e.g., Donmez, Nehme, & Cummings, 2010). This simulation required a single operator to control a team of UVs composed of unmanned air and underwater vehicles (UAVs and UUVs). All vehicles were engaged in related surveillance tasks, with the ultimate mission of identifying specific targets of interest in urban coastal and inland settings. Participants moved multiple UVs from one part of the screen to various targets while avoiding obstacles. The overall goal of the task was to get as many targets (objects located in a Google Earth image) correct while avoiding threat areas (circular areas on the map that will cause damage to a UV if flown over). In order to achieve this goal, the participants had many tools at their disposal. A computer mouse was used to assign UVs to targets, engage in targets (begin searching the image), and add waypoints (coordinates on the map that would route the UV on a different path) to avoid threat areas.

The RESCHU visual interface consisted of five major sections: map, camera window, message box, control panel, timeline (Figure 7). The map (Fig. 7a) displayed the locations of UVs, threat areas, and areas of interest. Vehicle control was carried out on the map, such as changing vehicle paths, adding waypoints, or assigning a target to a vehicle by selecting the UAV with the mouse. In addition, the map contains yellow circles representing threat areas. When a UV intersected with a threat area (UV moved to a position where a yellow circle is located on the map), RESCHU recorded the length of time the UV spent there. When the vehicles reached a target, a simulated video feed of a Google Earth image was displayed in the camera window (Fig. 7b). The participant had to visually identify a target (automobiles, landmarks, pools) in this simulated video feed. The main events in the mission (i.e., vehicles arriving to goals, or automatic assignment to new targets) were displayed in the message box, along with a timestamp (Fig. 7c). The message box was used by the operator to determine whether or not they had correctly identified a target, or one of the UVs had arrived at its target. There were also additional tools available to the operator such as a control panel (Fig. 7d) and timeline of expected UV arrival (Fig 7e). This gave the operator more information about when a UV would arrive at its target compared to the other UVs in motion.

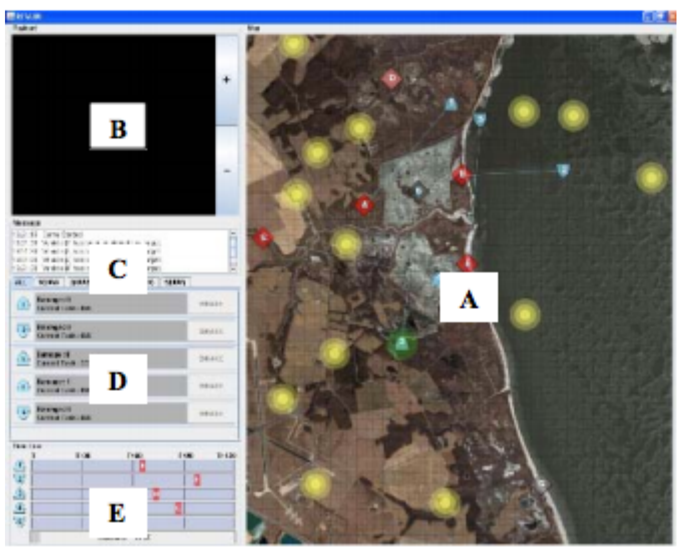

Figure 7: RESCHU interface (A: map, B: camera window, C: message box, D: control panel, E: timeline)

When a vehicle arrived at a target (flying for UAV and traveling underwater for UUV), a visual flashing alert indicated that the operator could engage the target by selecting the UV and clicking "engage". The operator then had to complete a search task by panning and zooming the camera with the mouse until the specified target was located (Fig. 7b). Once the operator submitted the target identification by right clicking on what they believed to be the correct target, the message box (Fig. 7c) notified the operator on the accuracy of response (used to simulate feedback that real operators get from their commanders or teammates as a consequence of their actions), and the vehicle was automatically re-assigned to a new target without controller input. Figure 8 contains images of what the video feed looked like when the operator engaged in a target.

Participants were instructed to maximize their score by avoiding threat areas that dynamically changed (every few minutes the threat areas would move to a new random location) and completing as many of the search tasks correctly as possible. In order to do this, operators could take advantage of re-planning (making corrections, or reassigning UVs to different targets) when possible to minimize vehicle travel times between targets and ensuring a vehicle was always assigned to a target whenever possible by clicking on unassigned UVs and giving them a target. The UVs were not modeled on real UV performance data as this experiment simulated a futuristic system (i.e., there are no operational command and control systems with integrated heterogeneous unmanned operations).

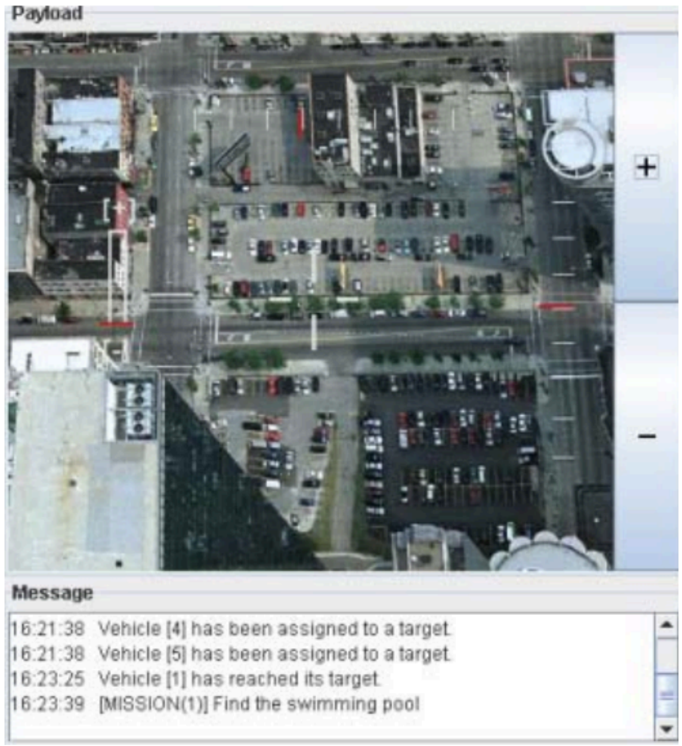

Figure 8: Activated camera view during search task.

2.2 Experimental Design

Thirty-six volunteers participated in the experiment, including 20 male and 16 female subjects, with an average age of 25.3 years. All procedures were reviewed and approved by Duke University's Internal Review Board (IRB). All subjects were asked to read and sign a consent form (Appendix A). After finishing the consent form, subjects filled out a demographic survey (Appendix B) and boredom proneness survey (Appendix C). Then subjects were taken to the experiment area where they were trained to use RESCHU. Following the training period, the subjects were seated in front of the monitors used to interact with the system. The training period lasted for 10 minutes and the entire experiment lasted for 60 minutes.

There were two main scenarios used during the experimental session. The first scenario was a low workload setting in which the user controlled six UVs. During these low workload test periods subjects were engaged in controlling the UVs, but had relatively little difficulty managing the UVs and successfully identifying targets. The second scenario was a high workload setting in which the user controlled twelve UVs. During these high workload test periods, subjects again were engaged in controlling the UVs, but were overloaded in completing tasks of avoiding threat areas while also correctly identifying targets. The number of UVs assigned for low and high workload tasks was determined through pilot testing. After testing was completed, subjects completed a post experiment questionnaire (Appendix D) and subjects were compensated $45.

All computer interactions were collected using Camtasia® recording software. RESCHU recorded all performance data automatically. The simulation logged all interactions, such as engaged targets, correct or incorrect responses, and damage to UV vehicles. The subjects stayed seated in the testing room for the entire hour long experiment. The subject sat at the end of a 5'x10' room in front of a computer which was on a desk. The experimenters were able to view the subject through a one-way glass window from an observation room directly next to the subject. Subjects could communicate with the experimenters and receive instructions through an intercom system between the experiment and observation rooms. The fNIRS device was placed centered on the subject's forehead directly between the eyebrows and hairline.

In addition to the computer, data were also collected using a fNIRS measurement device. The fNIRS sensor was kept in place using self-adhesive wrapping that went twice around the subject's head. A stretchable cloth cap was placed over the self- adhesive wrapping. On the inside of this cap was a piece of suede leather that covered the sensors for additional light blocking. Once securely attached, the fNIRS was calibrated and tested for good connection. The fNIRS device employed in this research was the Imagent™ system, developed and manufactured by ISS, Inc. This device is a "non-invasive tissue oximeter for the absolute determination of oxygenated and deoxygenated hemoglobin concentration, oxygen saturation and total hemoglobin content in tissues" (ISS Imagent, Inc). The overall process for data collection can be seen below in Figure 9. The data were collected using the Boxy software package created by ISS. A low pass filter of 0.15 Hz was applied to remove physiological changes such as heart rate and blood pressure. After the low pass filter, a discrete wavelet transform was run in MATLAB to remove motion artifacts in the data. The Modified Beer-Lambert Law was applied to produce HbO and HbR. These values can be seen in Table 1.

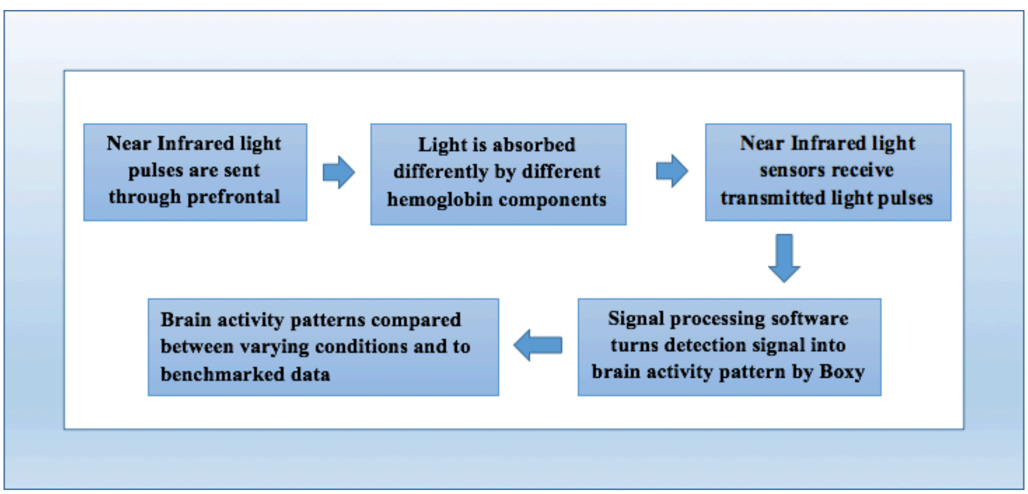

Figure 9: fNIRS Data Collection Method

Table 1: fNIRS Parameters

Parameter	Description
Modulation Frequency	110 MHz
Sources Spacing (distance between emitters and detectors)	16 emitters, 2 detectors. The fiber length is 2.5 m. The emitter-detector distances are 2.5, 3.0, 3.5, 4.0 cm for both right and left hemispheres.
Source Laser	Fiber coupled laser diodes Wavelengths: 690 nm, 830 nm
Light Detectors	Photomultiplier tubes
Sensors	Selected side-on photomultiplier tubes
Low Pass Filter	0.15Hz

2.3 Experimental Procedure

The fNIRS sensor was placed on the subject's forehead at the beginning of the testing period and was worn throughout the experiment. The subjects were instructed to try to avoid moving the sensors in any way and to refrain from furrowing the brow to maintain consistent data collection from the device (Solovey et al., 2009). While the system is resistant to minor movement, it was imperative to closely monitor the data and the subject during the experiment to determine if the sensor was moved from its original position.

The session contained four total test modules alternating between low and high workload settings (four modules, total), which each lasted 8 minutes. After each scenario was completed, the simulation would end and the subjects would relax during a four-minute break. During this break, participants were not allowed to use the computer, a cell phone, or anything else that would distract them. Subjects were allowed to close their eyes and focus on their breathing to prevent focusing on something mentally strenuous. Table 2 contains a basic breakdown of experimental session.

Table 2. Breakdown of experimental session by minutes

Event	Duration (minutes)
Tutorial	10
Break with fNIRS fitting	6
Test Module “Low” (Low 1)	8
Break (without distractions)	4
Test Module “High” (High 1)	8
Break (without distractions)	4
Test Module “Low” (Low 2)	8
Break (without distractions)	4
Test Module “High” (High 2)	8
Total	**60**

The primary independent variable (IV) in this experiment was the two workload levels (low and high), determined by the number of UVs controlled by the subject. The primary dependent variable was the fNIRS data, specifically oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR). For fNIRS, the HbO and HbR levels during specific scenarios were compared to its previous rest period (for example: Low 1 was compared to Rest 1, High 2 was compared to Rest 4) and a percent change was taken. The formal can be seen in HbO percent change (Equation 1) and HbR percent change (Equation 2).

Equation 1: HbO percent

$$HbO\ percent\ change\ = \frac{HbO\ level\ during\ scenario - HbO\ level\ during\ previous\ rest\ period}{HbO\ level\ during\ previous\ rest\ period}$$

Equation 2: HbR percent

$$HbR\ percent\ change\ = \frac{HbR\ level\ during\ scenario - HbR\ level\ during\ previous\ rest\ period}{HbR\ level\ during\ previous\ rest\ period}$$

The other dependent variables were assessments of performance, including overall performance score (Equation 3) and accuracy (Equation 4), as well as assessments of demographics, including boredom-proneness, age, or video game experience. The overall performance score of the UV simulation was using the system to maximize target identification while avoiding threat areas. If an operator's UVs spent ample time in threat areas and did not get many targets correct, the subject would have a negative overall performance score.

Equation 3: Overall Performance Score

$$Overall\ Performance\ Score = \left(\frac{total\ correct}{total\ number\ of\ UVs}\right) - \frac{\sum \frac{total\ time\ UV\ spends\ in\ threat\ area}{maximum\ time\ UV\ spends\ in\ threat\ area}}{total\ number\ of\ UVs}$$

Equation 4: Accuracy

$$Accuracy = \frac{total\ number\ correct}{total\ number\ of\ UVs}$$

2.4 Analysis

It was hypothesized subjects will have lower overall performance scores and accuracy in high workload scenarios compared to low workload scenarios. There was also interest in determining whether there was any impact of demographic data on overall performance score in low vs. high workload scenarios. A repeated measures ANOVA was run to identify the main effect of overall performance score across all four scenarios as well as interaction effects between the overall performance score and BPS, gaming, age, gender. A pairwise comparison was then run to identify specifically which workload scenarios differed from one another.

It was also hypothesized that there will be an accompanied larger increase in oxygenated hemoglobin (HbO) (decrease in deoxygenated hemoglobin (HbR)) during high workload scenarios compared to low workload scenarios. There was also interest in determining whether there was any impact of demographic data on HbO/HbR percent change in low vs. high workload scenarios. A repeated measures ANOVA was run to identify the main effect of HbO/HbR percent change for all four scenarios as well as interaction effects between the HbO/HbR percent change and BPS, gaming, age, gender. A pairwise comparison was then run to identify specifically which workload scenarios differed from one another.

2.5 Summary

This section describes the experiment conducted to measure the effect of time in low task load and situation difficulty on workload. Thirty-six participants were recruited to take part in a UV simulation representing supervisory control of multiple vehicles. The hemodynamic response was recorded throughout the entire hour long experiment using fNIRS. Subject tasks during low task load included controlling UVs to avoid threat areas while correctly identifying targets. Each subject completed the low and high workload simulations twice in the same order for each participant. Subjects filled out several surveys including a demographic survey, the Boredom Proneness Index, the post-experimental survey. Video recording was also performed for each subject.

3. Results

This section first introduces data collected from RESCHU and fNIRS. It also explores fNIRS data trends in the data relating to changes in workload. The detailed statistical data of the results can be seen in Appendix E.

3.1 Behavioral Results

The average overall performance score (Equation 3) and average accuracy (Equation 4) was computed for each subject under each scenario condition. The table below shows the average scores and standard deviation (SD) values for each scenario.

Table 3: Average, standard deviation, maximum, and minimum for overall performance score

	Average	Standard Deviation	Maximum Score	Minimum Score
Low 1	0.485	0.781	1.781	-1.434
High 1	-0.155	0.640	1.040	-1.362
Low 2	0.774	0.963	2.2382	-1.714
High 2	0.069	0.668	1.172	-1.280

Table 4: Average, standard deviation, maximum, and minimum for average accuracy

	Average	Standard Deviation	Maximum Accuracy	Minimum Accuracy
Low 1	1.328	0.547	2.333	0.166
High 1	1.178	0.386	2.083	0.333
Low 2	1.547	0.568	3.167	0.166
High 2	1.492	0.478	2.333	0.333

As Table 3 shows, the highest average score occurred in the Low 2 condition, and the lowest occurred in High 1. A similar trend occurred with accuracy (Table 4), with Low 2 again representing the best mean accuracy and High 1 the lowest. Performance scores ranged from -1.714 to 2.238. Accuracy ranged from 0.167 to 3.167. A visualization of the overall performance scores can be seen in a side-by-side boxplot in Figure 10.

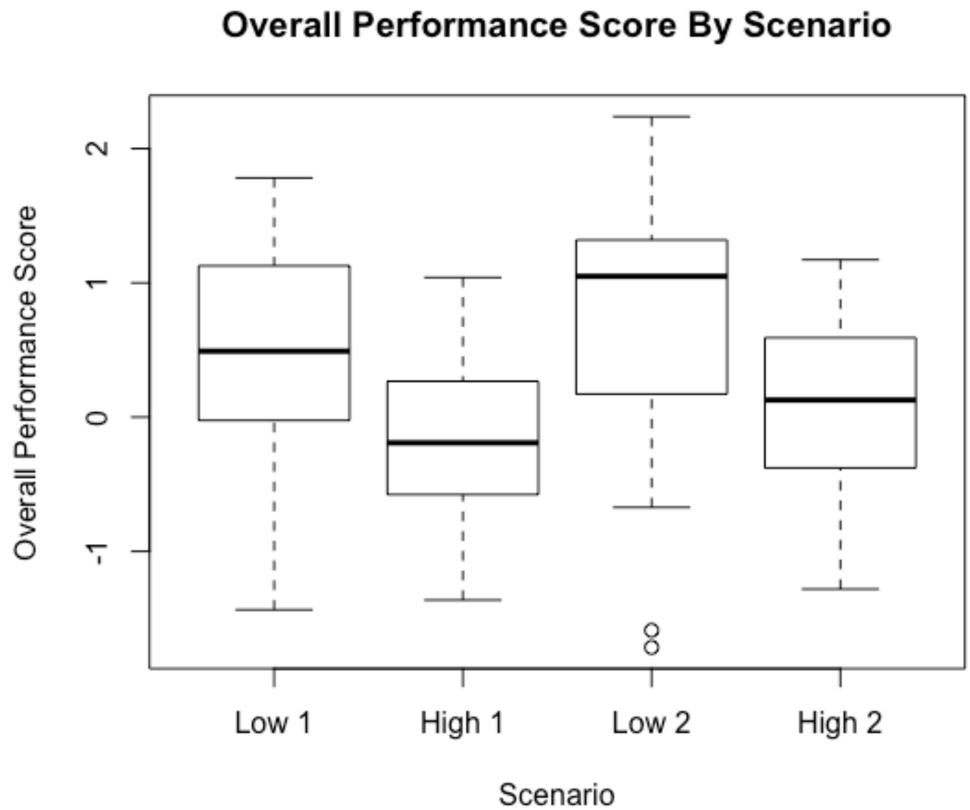

Figure 10: Side-by-side boxplot of comparing overall performance score by scenario

From Figure 10 we can see that on average, overall performance scores were both higher in low workload scenarios compared to high workload scenarios, which confirms the first hypothesis of this experiment, which was that subjects will have lower overall performance and accuracy scores in high workload scenarios as compared to low workload scenarios . This shows good internal validity of the experiment because subjects had a more difficult time with the high workload scenarios (High 1 and High 2) as compared to the low workload scenarios (Low 1 and Low 2). There is a slight learning effect among subjects. The effect is visible in Figure 10, because the average score in scenario Low 2 is higher than the average score in scenario Low 1. Also, the average score in scenario High 2 is higher than the average score in High 1, suggesting improvements over time.

A general linear model was run to compare overall performance scores using subjects' demographic data as covariates. The test was run to identify differences between subjects due to BPS, gaming experience, age and gender. The results showed a significant between subject effect for score and age ($p=0.028$). There were no significant effects between score and BPS ($p= 0.652$), gaming experience ($p=0.209$) or gender ($p=0.142$).

While there was no significant overall effect of gaming, BPS, gender, these were still included as potential interaction factors in a repeated measures ANOVA comparing the overall performance scores. The results of a repeated measures ANOVA, described in section 2.4, revealed no significant differences in performance scores based on scenario ($p=0.192$) under the full model. A significant interaction was discovered between score and age ($p=0.005$). There were no significant interaction effects between score and BPS ($p= 0.838$), gaming experience ($p=0.103$) or gender ($p=0.559$).

A repeated measures ANOVA was run again without any demographic factors. It was found that there was a significant difference in overall performance scores based on scenario ($p<0.001$). Pairwise comparison were run between scenarios for overall performance score to identify significant differences between scenario pairs. This test found that there was a significant difference in overall performance score between both low workload scenarios compared to both high workload scenarios.

Taking a deeper looking into age and its relationship to overall performance score (identified as significant by the GLM), a correlation test was run comparing age and overall performance score in each scenario. The test revealed a significant correlation between age and scenario Low 2 ($r=-.528$, $p=.001$). However, there were no significant correlations between age and Low 1 ($r=-.233$, $p=.178$), High 1 ($r=-.090$, $p=.607$), or High 2 ($r=-.277$, $p=.107$). Figure 11 shows a distribution of subjects' age.

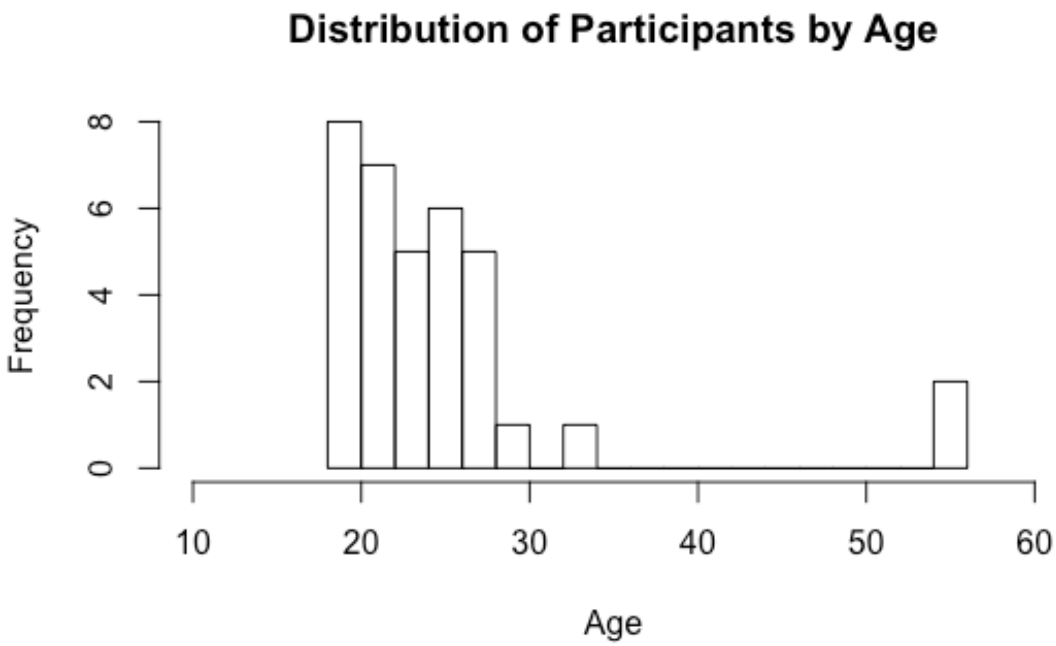

Figure 11: Histogram of Subjects Age

A general linear model was run on the accuracy score, again accounting for the subjects' demographic data. The test was run to identify differences in subject performance due to BPS, gaming experience, age and gender. The results showed a significant between subject effect between score and age (p=0.032) and gender (p=0.045). There were no significant effects between score and BPS (p=0.942) or gaming experience (p=0.155).

While there was no significant effect of gaming experience or BPS, these were still included as potential interaction factors in the repeated measures ANOVA with accuracy. The results of repeated measures ANOVA, described in Section 2.4, revealed no significant differences in performance scores based on scenario (p=0.551) under the full model. There were no significant interaction effects between score and age (p=0.311), BPS (p= 0.205), gaming experience (p=0.667) or gender (p=0.697).

A repeated measures ANOVA was run again without any demographic factors. It was found that there was a significant difference in accuracy based on scenario (p<0.001).

Pairwise comparisons were run between scenarios for accuracy to identify significant differences between scenario pairs. The tests found that there was a significant difference in accuracy between High 1 and Low 2 scenarios ($p=0.002$).

A correlation test was run comparing age and accuracy. There was a significant correlation between age and scenario Low 1 ($r=-.381$, $p=.024$) and High 2 ($r=-.443$, $p=.008$). However, there were no significant correlations between age and High 1 ($r=-.310$, $p=.070$) or Low 2 ($r=-.227$, $p=.190$).

3.2 fNIRS Results

The HbO percent change (Equation 1) and HbR percent change (Equation 2) was computed for each subject under each scenario condition. Below, Table 5 and Table 6 shows the average and standard deviation values for each scenario for HbO percent change and HbR percent change.

Table 5: Average, standard deviation, maximum, and minimum for average HbO percent by condition

	Average	Standard Deviation	Max Average	Min Average
Low 1	-0.062	0.142	0.117	-0.680
High 2	0.023	0.028	0.109	-0.023
Low 2	0.013	0.015	0.053	-0.024
High 2	0.021	0.038	0.174	-0.022

Table 6: Average, standard deviation, maximum, and minimum for average HbR percent by condition

	Average	Standard Deviation	Max Average	Min Average
Low 1	0.017	0.031	0.126	-0.032
High 2	0.019	0.018	0.088	0.001
Low 2	0.042	0.104	0.5836	-0.014
High 2	0.017	0.017	0.062	-0.017

As Table 5 shows, the highest average HbO percent change occurred in the High 1 condition, and the lowest occurred in Low 1. A similar trend occurred with HbR percent change (Table 6), with High 1 again representing the best mean HbR percent change and Low

1 the lowest. HbO percent change ranged from -0.68 to 0.174. Accuracy ranged rom -0.032 to 0.583.

A general linear model was run to compare HbO percent change using subjects' demographic data as covariates. The test was run to identify differences in subjects due to BPS, gaming experience, age and gender. The results showed no significant effects between HbO percent change and age ($p=0.409$), BPS ($p=0.209$), gaming experience ($p=0.477$), or gender ($p=0.143$).

While there was no significant overall effect of age, BPS, gaming experience or gender, these were still included as potential interaction factors in the repeated measures ANOVA with HbO percent change. The results of repeated measures ANOVA, described in Section 2.4, revealed no significant differences in HbO percent change based on scenario ($p=0.483$) under the full model. There were no significant interaction effects between score and age ($p=0.789$), BPS ($p= 0.653$), gaming experience ($p=0.655$) or gender ($p=0.500$).

Pairwise comparisons were run between scenarios for HbO percent change to identify significant differences between scenario pairs. The tests found that there was a significant difference in HbO percent change only in Low 1 compared to High 1 ($p=0.023$), Low 2 ($p=0.019$), and High 2 ($p=0.029$). A visualization of this pairwise comparison can be seen in Figure 12.

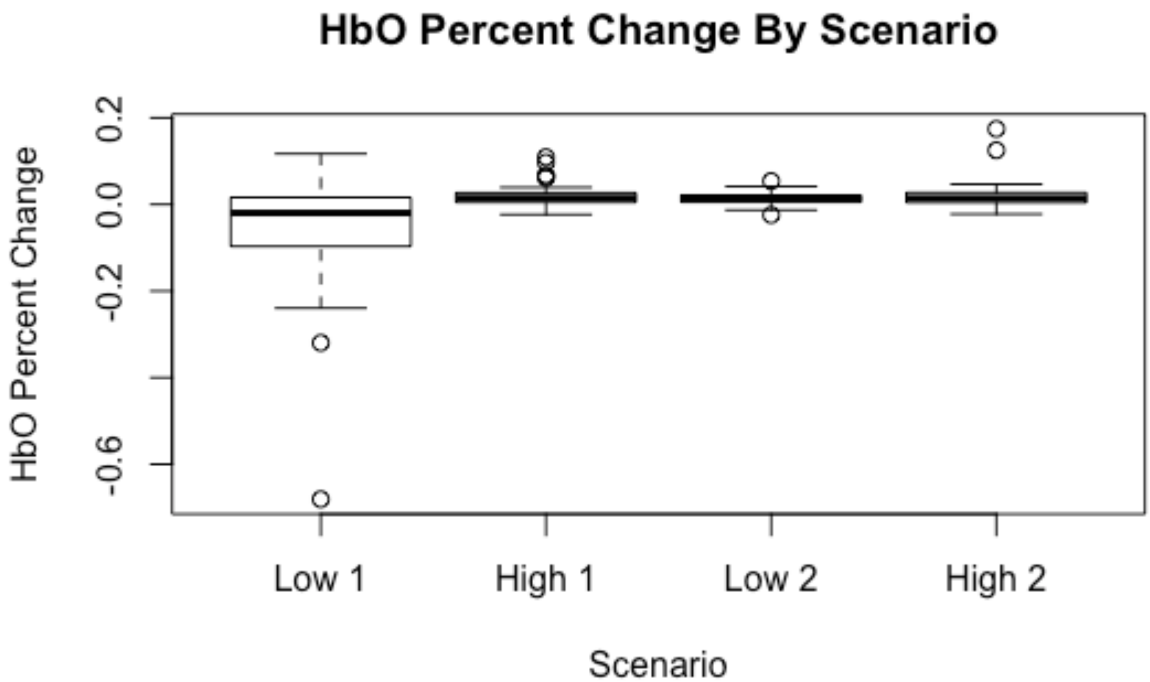

Figure 12: Side-by-side boxplot of HbO percent across all scenarios

A general linear model was run for HbR percent change based on subjects' demographic data. The test was run to identify differences between subjects due to BPS, gaming experience, age and gender. The results showed a significant between subject effect between HbR percent change and BPS (p=0.025). There were no significant effects between HbR percent change and age (p=0.341), gaming experience (p=0.167), or gender (p=0.260).

While there was no significant overall effect of age, gaming experience or gender, these were still included as potential interaction factors in the repeated measures ANOVA with HbR percent change. The results of repeated measures ANOVA revealed no significant differences in HbR percent change based on scenario (p=0.901) under the full model. There were significant interaction effects between HbR percent change and age (p=0.037).

Pairwise comparisons were run between scenarios for HbR percent change to identify significant differences between scenario pairs. The tests found that there were no significant difference in HbR percent change across all scenarios.

Taking a deeper looking into BPS and its relationship to HbR percent change, a correlation test was run comparing BPS and HbR percent change. There was a marginal correlation between BPS and scenario Low 1 ($r=-.316$, $p=.069$) and High 1 ($r=-.341$, $p=.056$). There were no significant correlations between age and Low 2 ($r=-.224$, $p=.234$), High 2 ($r=-.216$, $p=.251$).

3.3 Models

Multiple linear regressions were used to identify how the independent variables of the experiment were related to dependent variable. The first model created looked at the overall performance score during the Low 2 scenario, which immediately followed the first high workload scenario. The Low 2 scenario was selected because this was the time immediately following the first high workload scenario. Therefore, the model was expected to indicate which factors determine how a person will perform after a large shift in workload. The initial predictors were Low 2 HbR percent, Low 2 HbO percent, age, BPS, and gender. The multiple linear regression model of overall performance score during the Low 2 scenario was created and the significant factors included Low 2 HbO percent and age (R=0.682, R Square = 0.465, Adjusted R Square = 0.425, Std. Error of the Estimate=0.76302), and the model is shown below.

$$\textit{Overall Performance Score During Low 2} = 0.295(\textit{Low 2 HbO}) - 0.570(\textit{Age}) + C$$

The second model created looked at accuracy during the High 2 scenario which was the final scenario of the test period. This scenario was critical in our understanding of how people are able to perform at the end of a long task. The initial predictors were Low 1 HbO percent, High 1 HbO percent change, Low 1 HbR percent change, Low 2 HbO percent change, High 1 HbR percent change, Low 2 HbR percent change, age, and BPS. A multiple linear regression model of accuracy during the High 2 scenario was created and the significant factors included: age, BPS, and Low 1 HBR percent (R=0.606, R Square = 0.368, Adjusted R Square = 0.262, Std. Error of the Estimate=0.615) and the model is shown below.

$$\text{Accuracy High 2} = 0.364(\text{BPS}) - 0.498(\text{Age}) - 0.296(\text{Low 2 HbO Percent}) + 0.596(\text{Low 1 HbR Percent}) + C$$

4. Discussion

This section examines how the results of this experiment fit into the broader context of mental workload and fNIRS detection of oxygenated and deoxygenated hemoglobin levels. It also identifies possible confounding variables and limitations of this experiment. Finally, it provides recommendations for future work using functional brain imaging for workload detection.

4.1 Experiment Discussion

There are several conclusions that can be drawn from the experiment. First, the two dependent behavioral variables measured (overall performance score and average accuracy) showed significant differences between low workload scenarios and high workload scenarios. As expected, participants had higher overall performance scores and higher average accuracy when controlling 6 UAVs compared to 12 UAVs. This confirms the first hypothesis that subjects would perform worse on high workload scenarios compared to low workload scenarios based on overall performance score and accuracy.

Age proved to be a significant predictor of overall performance score. These is precedent for this result, as previous studies (e.g., Deaton & Parasuraman, 1993) have linked increasing age to decreases in performance during memory and vigilance tasks. In addition, older subjects may not be as comfortable operating a computer interface and dealing with complex scenarios, so it is not surprising that higher age was found to be a predictor of performance. Interestingly, there was a significant correlation between age and the second low workload scenario (Low 2). It is possible that subjects were mentally fatigued after completing the high workload scenario, in general, but older subjects took longer to recover, which affected performance on the next scenario.

Next, the performance data was compared to the resulting oxygenated and deoxygenated hemoglobin levels recorded by the fNIRS. It was expected that this increase in objective difficulty would correspond to a physiological difference between the low and high workload scenarios. However, the results indicate that the physiological differences in mental workload were not significantly different between the low and high mental workload scenarios. Unlike other studies (e.g., Jelzow et al., 2011; Herff et al., 2014), this study did not show fNIRS was capable of distinguishing different levels of mental workload.

There are many possible reasons for why this experiment did not detect any significant HbO or HbR differences with respect to degree of difficulty. First, the volume of data was much greater for this fNIRS experiment. There were 36 subjects studied (which is significantly more than most previous fNIRS research (Jelzow et al., 2011; Herff et al., 2014)), and each subject received each workload scenario twice for a total experimental runtime of 60 minutes. Another difference between this research and previous work is that this test scenario represented a dynamic task in a more realistic environment, compared to the simple matching tasks cited in previous work. This is consistent with another study conducted in a missile control environment (Boyer et al., 2015), which also did not detect workload changes with fNIRS.

The second factor that could explain lack of sensitivity in the fNIRS results is that the subjects in this experiment were inexperienced operators. Even though subjects were provided a tutorial, experimental guidance, and a knowledge check, it is possible that subjects were still dealing with some uncertainty in using the interface during the first scenario. Subjects could have reached instantaneous peaks of mental workload for the 6 UAV scenario that were equivalent to the subsequent 12 UAV scenario because these subjects were not familiar with the RESCHU interface.

However, analysis of the impact of beginning of the experiment showed some noteworthy results. The most significant result was that the average difference in HbO percent was much more varied during the first workload scenario Low 1, as seen in Figure 11. During second, third, and fourth workload scenario there was no significant difference in HbO percent. The best explanation for this observed outcome is that during the first scenario the brain is assigning resources to meet the demands of the simulation environment. Once it has reached the necessary flow of oxygenated hemoglobin, HbO levels begins to level off for the rest of the experiment.

Interestingly, the timeline of when we observed the increase in HbO change is in line with previous research that analyzes the ability to maintain concentration and attention over prolonged periods of time. This phenomenon, termed the “vigilance decrement” has been studied extensively. According to the experimental timeline, scenario vigilance decrement would have occurred around the 20-minute mark in the experiment. This timeline has been seen in previous vigilance decrement research (See et al., 1995) that shows the 20 to 35 minutes is in window where the vigilance decrement traditionally occurs.

Looking back at Section 1.1, the goal of this experiment was to use behavioral measures obtained from a supervisory control simulation (e.g., response time, accuracy), and non-invasive neurophysiological measures of oxygenated hemoglobin and deoxygenated hemoglobin derived from scalp-recorded fNIRS to determine if fNIRS can be an accurate tool for measuring mental workload. This study showed that fNIRS failed to accurately predict changes in mental workload over short periods of time, but it did detect an initial change in oxygenation levels that could be associated with the vigilance decrement. This is a promising result that should be studied further.

Our initial goal of conducting this research was to determine a way of remotely detecting a workload that could be hindering human performance. By combining the measured hemodynamic and metabolic responses of brain and cognitive activity using fNIRS, our goal was to determine a change in a subject's cognitive activity while performing different tasks. Unfortunately, fNIRS is not capable of detecting this change in cognitive activity during the supervisory control tasks in this study and further models cannot be generated to increase training and performance evaluations.

4.2 Limitations

While every effort was made to control confounding variables and generate effective results, there are some limitations that should be discussed for this work. First, the fNIRS needed to be applied to subjects' forehead and wrapped to prevent light from interfering with the infrared lasers. This cap system did cause some subjects discomfort near the end of the experiments which could have contributed unexpected rises in mental strain. Second, the experiment was conducted under only a single blind condition, with the experimenter having knowledge of the experimental condition at all times. This was done in order to properly monitor the subject and ensure the simulation was working correctly, but may have introduced a bias into the experiment. The experimenter tried to avoid entering the room in before the event and avoid interpersonal interaction, but this was not always possible and may have resulted in decreasing boredom before an event. Third, increased experience in running the experiment may have slightly modified the experimenter conduct over time, especially in regards to addressing questions by the subject about how to best utilize the interface.

Age proved to be a significant factor in predicting overall performance score, but the overall distribution of age was very skewed towards a younger population with only two subjects in their 50s. It would be helpful to repeat the experiment with more subjects with an even distribution across ages to fully investigate the age correlations.

4.3 Future Work

Since workload research is important for understanding human-supervisory control, there are many avenues to explore that are suggested by this work. The first area for further research would involve studies similar that had larger breaks between tasks. While this study did show that fNIRS was unable to detect workload changes over short periods of time, it might be more accurate if longer breaks are taken between tasks. This would allow for the brain to return baseline levels of cognitive function before beginning a task again. Other permutations of this experiment could modify the overall environment, the amount of training time, the amount of low workload vs. high workload time, or the complexity of the task.

The second area to explore is looking at using fNIRS to specifically focus on measuring the initial change in HbO levels during a given task. This research showed that fNIRS was able to detect an initial spike in HbO, but it would be interesting to explore if this differs with varying tasks. Adding in a dedicated vigilance task such as monitoring a process or video feed could help to elucidate the vigilance findings and increase the similarity to many real-world environments.

References:

Baker W. B., Parthasarathy A. B., (2014). Busch D. R., Mesquita R. C., Greenberg J. H., Yodh A. G., “Modified Beer-Lambert law for blood flow,” Biomed. Opt. Express 5(11), 4053–4075 10.1364/BOE.5.004053

Battiste, V., & Bortolussi, M. (1988). Transport pilot workload: A comparison of two subjective techniques. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Anaheim, CA.

Berman, M., Jonides, J., and Nee, D.E. (2006). Studying mind and brain with fMRI Social, Cognitive, and Affective Neuroscience, 1, 158-1561

Berka, C., Levendowski, D. J., Ramsey, C. K., Davis, G., Lumicao, M. N., Stanney, K., . . . Stibler, K. (2005). Evaluation of an EEG workload model in an Aegis simulation environment. Paper presented at the Defense and Security.

Boyer, M., Cummings, M. L., Spence, L. B., & Solovey, E. T. (2015). Investigating Mental Workload Changes in a Long Duration Supervisory Control. TaskInteract.Comput., 27 (5): 512-520

Buckner, R. L., & Logan, J. M. (2001). Functional neuroimaging methods: PET and fMRI. Handbook of functional neuroimaging of cognition, 27-48.

Cabeza, R., & Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. Journal of cognitive neuroscience, 12(1), 1-47.

Carr, V. A., Rissman, J., & Wagner, A. D. (2010). Imaging the Human Medial Temporal Lobewith High-Resolution fMRI. Neuron, 65(3), 298-308. doi:http://dx.doi.org/10.1016/j.neuron.2009.12.022

Cooper, G. E., & Harper Jr, R. P. (1969). The use of pilot rating in the evaluation of aircraft handling qualities. Neuilly-sur-Seine, France: NATO Advisory Group for Aerospace Research and Development.

Deaton, J. E., & Parasuraman, R. (1993). Sensory and cognitive vigilance: Effects of age on performance and subjective workload. Human Performance, 6(1), 71–97.

Donmez, B., Nehme, C., & Cummings, M. L. (2010). Modeling Workload Impact in Multiple Unmanned Vehicle Supervisory Control. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans 40, 6, 1180–1190.

Endsley, M. R., & Rodgers, M. D. (1997). Distribution of attention, situation awareness, and workload in a passive air traffic control task: Implications for operational errors and automation. Air Traffic Control Quarterly, 6(1), 21-44.

Ferrari, M., Giannini, I., Carpi, A., Fasella, P., Fieschi, C., Zanette, E., (1980). Non invasive infrared monitoring of tissue oxygenation and circulatory parameters. XII World Congress of Angiology, Athens, September 7–12, abs. 663.

Ferrari, M., Giannini, I., Carpi, A., Fasella, P., (1982). Near I.R. spectroscopy in non invasive monitoring of cerebral function. In: Bleifeld, W., Harder, D., Leetz, H.K., Schaldach, M. (Eds.), Proc. World Congress on Medical Phys. and Biomed. Eng. 1982, Hamburg, September 5–11. MPBE 1982 e. V., Hamburg, abs. 22.17.

Ferrari M. & Quaresima V., (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. Neuroimage 63, 921–935.

Hancock, P. A., & Warm, J. S. (1989). A Dynamic Model of Stress and Sustained Attention. Human Factors: The Journal of the Human Factors and Ergonomics Society, 31(5), 519- 537. doi: 10.1177/001872088903100503

Hart, S. G., & Sheridan, T. B. (1984). Pilot workload, performance, and aircraft control automation. Moffett Field, CA: National Aeronautics and Space Administration Ames Research Center.

Hart, S. G. & Staveland, L. E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati (Eds.) Human Mental Workload. Amsterdam: North Holland Press.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., Schultz, T., (2014) Mental workload during n-back task – quantified in the prefrontal cortex using fNIRS. Frontiers in Human Neuroscience. Vol 7. DOI=10.3389/fnhum.2013.00935

Huey, B. M., & Wickens, C. D. (1993). Workload Transition:Implications for Individual and Team Performance: The National Academies Press.

Jelzow, A., Tachtsidis, I., Kirilina, E., Niessing, M., Brühl, R., Wabnitz, H., . . . Macdonald, R. (2011). Simultaneous measurement of time-domain fNIRS and physiological signals during a cognitive task. Paper presented at the European Conferences on Biomedical Optics.

Jöbsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. Science (New York, N.Y.), 198(4323), 1264-1267.

Klein, M. I., Riley, M. A., Warm, J. S., & Matthews, G. (2005). Perceived mental workload in an endocopic surgery simulator. Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting, Orlando, FL.

Kramer, A. F. (1991). Physiological metrics of mental workload: A review of recent progress. Multiple-task performance, 279-328.

Koechlin, E., Basso, G., Pietrini, P., Panzer, S., & Grafman, J. (1999). The role of the anterior prefrontal cortex in human cognition. Nature, 399(6732).

Mehta R. K., Parasuraman R.(2013) Neuroergonomics: a review of applications to physical and cognitive work.Front Hum Neurosci. doi: 10.3389/fnhum.2013.00889.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. Annual review of neuroscience, 24(1), 167-202.

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin Card Sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. The Journal of Neuroscience, 21(19), 7733-7741.

Moray, N. E. (1979). Mental workload: Its theory and measurement. New York: Plenum Press. Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), Varieties of Attention (pp. 63-102). Orlando, FL: Academic Press.

Ochsner, K. N., Bunge, S. A., Gross, J. J., & Gabrieli, J. D. (2002). Rethinking feelings: An fMRI study of the cognitive regulation of emotion. Journal of cognitive neuroscience, 14(8), 1215-1229.

O'Donnell, R., & Eggemeier, F. T. (1986). Workload assessment methodology. Measurement Technique, 42, 5.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2008). "Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs," *Journal of Cognitive Engineer- ing and Decision Making*, Vol. 2, pp. 140–160.

Pasley B. N., & Freeman R. D. (2008). Neurovascular coupling. Scholarpedia, 3(3):5340.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. Systems, Man and Cybernetics, IEEE Transactions on(3), 257-266.

Reid, G. B., & Nygren, T. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. Human mental workload, 185, 218.

See, J., Howe, S., Warm, J., & W. Dember. Meta-analysis of the sensitivity decrement in vigilance. *Psychological Bulletin*, Vol. 117, No. 2, 1995, pp. 230-249.

Sirevaag, E. J., Kramer, A. F., Reisweber, C. D. W. M., Strayer, D. L., & Grenell, J. F. (1993). Assessment of pilot performance and mental workload in rotary wing aircraft. Ergonomics, 36(9), 1121-1140. doi: 10.1080/00140139308967983

Solovey, E. T., Girouard, A., Chauncey, K., Hirshfield, L. M., Sassaroli, A., Zheng, F., &

Jacob, R. J. (2009). Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines. In Proceedings of the 22nd annual ACM symposium on User interface software and technology (pp. 157-166). ACM.

Warm, J.S., Matthrew, G., & Parasuraman, R. (2009). Cerebral hemodynamics and vigilance performance. Military psychology, 21, S75-S100

Whyte, J. (2011). Blood Oxygen Level-Dependent Encyclopedia of Clinical Neuropsychology (pp. 423-426): Springer.

Wickens, C., & Hollands, J. (1999). Engineering Psychology and Human Performance (3rd Edition): Prentice Hall.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. Human Factors: The Journal of the Human Factors and Ergonomics Society, 35(2), 263-281.

Wierwille, W. W. (1979). Physiological measures of aircrew mental workload. Human Factors: The Journal of the Human Factors and Ergonomics Society, 21(5), 575-593.

Wilson, G. (2002). An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. The International Journal of Aviation Psychology, 12(1), 3-18. doi: 10.1207/s15327108ijap1201_2

Wilson, G. & Eggemeier, F.T., (1991). Physiological measures of workload in multi-task environments. In Damos (ed.) *Multiple-Task Performance* (London: Taylor & Francis), pp. 329–60.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit- formation. Journal of comparative neurology and psychology, 18(5), 459-482.

Appendices

Appendix A: Consent to Participate in Non-medical Research

CONSENT TO PARTICIPATE IN NON-BIOMEDICAL RESEARCH

Detecting Long Distance Driver Cognitive Disengagement
You are asked to participate in a research study conducted by Professor Mary Cummings, Ph.D. from the Department of Mechanical Engineering and Materials Science at Duke University.. By signing below, you agree to participate in this research. You may keep a copy of this consent form for your records.
PARTICIPATION AND WITHDRAWAL
Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may quit at any time. The investigator may withdraw you from this research if necessary. A withdraw will occur if the investigator is having difficulty fitting fNIRS or EEG to head in order to gain necessary data. You should not participate if you have a history of neurological disorders, seizure disorders, or head injury.
PURPOSE OF THE STUDY
This study is being conducted as part of a project studying next-generation human-computer interaction techniques in low workload/high workload simulations. We are investigating the use of non-invasive brain sensor data as supplemental data to compare two different brain sensing devices (fNIRS and EEG).
PROCEDURES
If you volunteer to participate in this study, we will follow a given protocol:
1. You will be given an overview of the experiment.
2. You will be asked to review and sign an approved informed consent form (this form).
3. You will be asked a number of questions to ensure your eligibility for the study.
5. Sensors (functional near-infrared spectroscopy, fNIRS) will be placed on the forehead. This device is non-invasive and measures blood oxygen levels in the front part of the brain, which can indicate brain activity. When the sensors are applied, the lasers will be turned off to prevent eye exposure, and the system is certified as eye-safe. In addition, we will collect electroencephalography (EEG), which records electrical activity in the brain. This will entail wearing a nylon cap with 20 electrodes embedded in it. We will place some water-soluble gel in each electrode that will allow the EEG sensors to detect your brain's electrical activity. At the end of the session both types of sensor will be removed and you will be given a towel and shampoo to clean your hair. Both fNIRS and EEG have been widely used in thousands of studies to investigate cognitive functions and these non-invasive approaches have no known harmful side effects.
6. You will be asked to complete the Demographic Survey
7. You will be instructed to complete the workload tasks. The workload tasks entail two levels of supervisory control that will last for 8 minutes each. You will use the RESCHU simulation to complete this task. You will be given a 15 minute training period to learn how to use the RESCHU simulation.
8. The lasers will be turned off to prevent eye exposure and then the sensors will be removed.
9. You will then be paid for your participation and any final questions can be answered.
The total experiment will take no more than two hours. During task performance (step 7), measures of your performance will be recorded in terms of your position on the road as well as your blood oxygenation measures from the sensors.
POTENTIAL RISKS AND DISCOMPORTS
There are no major risks anticipated from participation in this study. The blood flow-monitoring device utilizes Class 3B lasers, which presents a minimal risk of eye exposure to laser light during the application and removal of the helmet containing the sensors. To prevent such a risk we will turn off the lasers during the application of the headband. Please inform the experimenter at the first sign of any discomfort. Should you wish to stop or delay the experiment, you are free to do so at any time. If you feel nauseous or sick, please let the experimenter know and the experiment will be stopped; we will offer you some water and rest time. You may then choose to continue or end the experiment.
POTENTIAL BENEFITS
While you will not benefit directly from this study, the results from this study will assist in the design for future interactive systems.
PAYMENT FOR PARTICIPATION
You will be paid $45 to participate in this study or $15/hour if the study is not completed. This will be paid upon completion of the experiment. Should you elect to withdraw in the middle of the study, you will be compensated for the hours you spent in the study.
CONFIDENTIALITY
Upon signing this consent you will be assigned a subject number that will be used to track data collected during this study. The Duke Humans and Autonomy Lab will maintain a copy of this consent in a secure location for a minimum of three years as a record of your participation. Your name, social security number, address and phone number will be provided to Duke accounts payables as a record of payment for participation. There will be no key connecting the payment information to the data. Data recorded during this experiment has the potential to answer research questions beyond the initial scope of this work. Therefore, the deidentified data may be maintained in perpetuity.

IDENTIFICATION OF INVESTIGATORS

If you have any questions or concerns about the research, please feel free to contact the Principal Investigator, Mary L. Cummings, through phone: (919) 660-5306, e-mail: mary.cummings@duke.edu, or mailing address: P.O. Box 90300, 144 Hudson Hall, Duke University, Durham, NC 27708.

RIGHTS OF RESEARCH SUBJECTS

If you have questions regarding your rights as a research subject, you may contact the Duke Office of Research Support: Suite 710 Erwin Square, 2200 W. Main Street, Durham, NC 27705, Phone (919) 684-3030

SIGNATURE OF RESEARCH SUBJECT OR LEGAL REPRESENTATIVE

By signing below, you agree to participate in this research. If you would like a copy of this consent form, one will be provided upon request

Name of Subject

____

Signature of Subject

____

Date:

Appendix B: Demographic Survey

1. Subject number:_____

2. Age:_____

3. Gender: *M* *F*

4. Color Blindness: *N* *Y*

If yes, type:________________

5. Occupation:________________________________

if student, (circle one): *Undergrad* *Masters* *PhD*

expected year of graduation:_________

7. Have you used detailed procedures before (e.g. checklists, model-building)?

No *Yes*

If yes, please briefly explain:_______________________________________

8. How often do you play computer games?

Rarely *Monthly* *Weekly* *A few times a week* *Daily*

Types of games played:____

9. Rate your comfort level with using computer programs.

Not comfortable *Somewhat comfortable* *Comfortable* *Very Comfortable*

10. What is your attitude toward Unmanned Aerial Vehicles (Drones)?

Intense dislike *Dislike* *Neutral* *Like* *Really Like*

Appendix C: Boredom Proneness Survey

1. It is easy for me to concentrate on my activities.	T | F
2. Frequently when I am working I find myself worrying about other things.	T | F
3. Time always seems to be passing slowly.	T | F
4. I often find myself at "loose ends," not knowing what to do.	T | F
5. I am often trapped in situations where I have to do meaningless things.	T | F
6. Having to look at someone's home movies or travel slides bores me tremendously.	T | F
7. I have projects in mind all the time, things to do.	T | F
8. I find it easy to entertain myself.	T | F
9. Many things I have to do are repetitive and monotonous.	T | F
10. It takes more stimulation to get me going than most people.	T | F
11. I get a kick out of most things I do.	T | F
12. I am seldom excited about my work.	T | F
13. In any situation I can usually find something to do or see to keep me interested.	T | F
14. Much of the time I just sit around doing nothing.	T | F
15. I am good at waiting patiently.	T | F
16. I often find myself with nothing to do-time on my hands.	T | F
17. In situations where I have to wait, such as a line or queue, I get very restless.	T | F
18. I often wake up with a new idea.	T | F
19. It would be very hard for me to find a job that is exciting enough.	T | F
20. I would like more challenging things to do in life.	T | F
21. I feel that I am working below my abilities most of the time.	T | F
22. Many people would say that I am a creative or imaginative person.	T | F
23. I have so many interests, I don't have time to do everything.	T | F
24. Among my friends, I am the one who keeps doing something the longest.	T | F

Appendix D: Post Experiment Survey

1. How confident were you about the actions you took?

Not Confident   Somewhat Confident   Confident   Very Confident   Extremely Confident

Comments:

2. How did you feel you performed?

Very Poor   Poor   Satisfactory   Good   Excellent

3. How stressed did you feel during the difficult situation?

Not Stressed   Somewhat Stressed   Stressed   Very Stressed   Extremely Stressed

4. How busy did you feel during the more difficult situation?

Idle   Not Busy   Busy   Very Busy   Extremely Busy

5. Do you feel that the training sufficiently prepared you for the test? *No   Yes*

Comments:

6. How well do you feel you understand simulation operation?

Very Poorly   Poorly   Satisfactory   Well   Very Well

7. Were the procedures easy to understand? *No   Yes*

Comments:

8. Other comments:

Appendix E: Statistical Tables from Results

Repeated measures ANOVA for overall performance score

Effect		Value	F	p-value
OverallPerformanceScore	Wilks'Lambda	.847	1.690	.192
OverallPerformanceScore * BPS	Wilks'Lambda	.971	.282	.838
OverallPerformanceScore * gaming	Wilks'Lambda	.805	2.260	.103
OverallPerformanceScore * age	Wilks'Lambda	.641	5.216	.005
OverallPerformanceScore * gender	Wilks'Lambda	.930	.701	.559

Repeated measures model for overall performance score

Source	Type III Sum of Square	df	Mean Square	F	p-value
Intercept	.315	1	.315	.273	.605
BPS	.239	1	.239	.208	.652
Gaming	1.899	1	1.899	1.651	.209
Age	6.115	1	6.115	5.317	.028
Gender	2.615	1	2.615	2.274	.142
Error	34.502	30	1.150		

Pairwise comparison for overall performance score

(I) Overall Performance Score	(J) Overall Performance Score	(I-J) Mean Difference	p-value
Low 1	Low 2 High 1 High 2	-.264 .637 .427	.689 <.001 .022
High 1	Low 1 Low 2 High 2	-.637 -.901 -.210	<.001 <.001 .846
Low 2	Low 1 High 1 High 2	.264 .901 .691	.689 <.001 <.001
High 2	Low 1 Low 2 High 1	-.427 -.691 .210	.022 <.001 .846

Repeated measures ANOVA for accuracy

Effect		Value	F	p-vaule
Accuracy	Wilks'Lambda	.929	.716	.551
Accuracy * age	Wilks'Lambda	.882	1.248	.311
Accuracy * BPS	Wilks'Lambda	.851	1.630	.205
Accuracy * gaming	Wilks'Lambda	.947	.527	.667
Accuracy * gender	Wilks'Lambda	.951	.483	.697

Repeated measures model for accuracy

Source	Type III Sum of Square	df	Mean Square	F	p-value
Intercept	3.394	1	3.394	5.552	.025
BPS	3.082	1	3.082	5.041	.032
Gaming	.006	1	.006	.009	.924
Age	1.298	1	1.298	2.123	.155
Gender	2.664	1	2.664	4.357	.045
Error	18.341	30	.611		

Pairwise comparison for accuracy

(I) Accuracy	(J) Accuracy	(I-J) Mean Difference	p-value
Low 1	Low 2 High 1 High 2	-.208 .144 -.165	.068 .272 .162
High 1	Low 1 Low 2 High 2	-.144 -.352 -.309	.272 .002 <0.001
Low 2	Low 1 High 1 High 2	.208 .352 .043	.068 .002 1.000
High 2	Low 1 Low 2 High 1	.165 -.043 .309	.162 1.000 <0.001

Repeated measures ANOVA for HbO percent

Effect		Value	F	p-value
HbO percent	Wilks'Lambda	.892	.849	.483
HbO percent * BPS	Wilks'Lambda	.927	.551	.653
HbO percent * gaming	Wilks'Lambda	.927	.548	.655
HbO percent * age	Wilks'Lambda	.952	.351	.789
HbO percent * gender	Wilks'Lambda	.896	.841	.500

Repeated measure model for HbO percent

Source	Type III Sum of Square	df	Mean Square	F	p-value
Intercept	.008	1	.008	2.516	.126
Age	.002	1	.002	.706	.409
BPS	.005	1	.005	1.670	.209
Gaming	.002	1	.002	.522	.477
Gender	.007	1	.007	2.304	.143
Error	.071	23	.003		

Pairwise comparison of HbO percent

(I) HbO Percent	(J) HbO Percent	(I-J) Mean Difference	p-value
Low 1	Low 2	.109	.019
	High 1	.100	.023
	High 2	.111	.029
High 1	Low 1	-.100	.023
	Low 2	.009	1.000
	High 2	.011	.725
Low 2	Low 1	-.109	.019
	High 1	-.009	1.000
	High 2	.002	1.000
High 2	Low 1	-.111	.029
	Low 2	-.002	1.000
	High 1	-.011	.725

Repeated measures ANOVA for HbR percent

Effect	Value	F	p-value

HbR percent	Wilks'Lambda	.973	.192	.901
HbR percent * age	Wilks'Lambda	.673	3.401	.037
HbR percent * BPS	Wilks'Lambda	.951	.363	.781
HbR percent * gaming	Wilks'Lambda	.978	.161	.921
HbR percent * gender	Wilks'Lambda	.708	2.886	.060

Repeated measures model for HbR percent

Source	Type III Sum of Square	df	Mean Square	F	p-value
Intercept	.003	1	.003	2.950	.099
Age	.001	1	.001	.944	.341
BPS	.006	1	.006	5.763	.025
Gaming	.002	1	.002	2.037	.167
Gender	.001	1	.001	1.336	.260
Error	.023	23	.001		