Optical brain monitoring for operator training and mental workload assessment

Hasan Ayaz a,⁎, Patricia A. Shewokisa,b, Scott Bunce c, Kurtulus Izzetoglua, Ben Willems d, Banu Onarala

a School of Biomedical Engineering, Science & Health Systems, Drexel University, USA
b College of Nursing and Health Professions, Drexel University, USA
c Department of Psychiatry, M.S. Hershey Medical Center, Penn State College of Medicine, USA
d Atlantic City International Airport, Federal Aviation Administration William J. Hughes Technical Center, USA

A R T I C L E   I N F O

Available online 17 June 2011

Keywords:
Human performance
fNIR
Mental workload assessment
Training
Air traffic control
Unmanned aerial vehicle

A B S T R A C T

An accurate measure of mental workload in human operators is a critical element of monitoring and adaptive aiding systems that are designed to improve the efficiency and safety of human–machine systems during critical tasks. Functional near infrared (fNIR) spectroscopy is a field-deployable non-invasive optical brain monitoring technology that provides a measure of cerebral hemodynamics within the prefrontal cortex in response to sensory, motor, or cognitive activation. In this paper, we provide evidence from two studies that fNIR can be used in ecologically valid environments to assess the: 1) mental workload of operators performing standardized (n-back) and complex cognitive tasks (air traffic control — ATC), and 2) development of expertise during practice of complex cognitive and visuomotor tasks (piloting unmanned air vehicles — UAV). Results indicate that fNIR measures are sensitive to mental task load and practice level, and provide evidence of the fNIR deployment in the field for its ability to monitor hemodynamic changes that are associated with relative cognitive workload changes of operators. The methods reported here provide guidance for the development of strategic requirements necessary for the design of complex human–machine interface systems and assist with assessments of human operator performance criteria.

© 2011 Elsevier Inc. All rights reserved.

Introduction

Human mental workload plays a critical role in many complex command and control systems. A key feature of the concept of mental workload – which reflects how hard the brain is working to meet task demands – is that it can be dissociated from performance output. Experienced human operators can maintain performance at required levels for a while through increased effort and motivation (Matthews et al., 2000) or strategy changes (Sperandio, 1978), even in the face of increased task challenge. Sustained task demands, however, eventually leads to performance breakdown, and increased mental workload in the intervening period can be predictive of subsequent performance failure. Consequently, it is important to assess mental workload independently of performance measures. Neuroergonomic approaches based on measures of human brain hemodynamic or electromagnetic activity can provide for sensitive and reliable assessment of human mental workload in complex work environments (Parasuraman, 2011).

It is particularly important to assess and measure operator mental workload in situations where performance failures could result in catastrophic losses (e.g., military command and control, air traffic control (ATC), etc.). Accurate assessment of mental workload could help in preventing operator error and allow for pertinent intervention by predicting performance decline that can arise from either work overload or understimulation (Hancock and Parasuraman, 1992, Hancock and Verwey, 1997; Meshkati et al., 1995; Parasuraman and Rizzo, 2007; Parasuraman and Wilson, 2008).

To date, investigators have worked on adaptive aiding schemes to facilitate optimal performance in critical mission systems by dynamically matching the momentary mental capabilities of the operator to the imposed task demands (Young and Stanton, 2002a,b). To be viable, such systems must improve performance above the levels possible with unaided and fully automated systems (Hancock and Verwey, 1997; Parasuraman and Riley, 1997). Further, adaptive aiding systems should provide aid only when required (Scerbo, 1996), as providing unnecessary intervention can lead to performance errors as readily as not providing aid when it is required (Hancock and Parasuraman, 1992; Hancock and Verwey, 1997). An accurate and reliable measure of the operator’s mental workload is a critical component of any such adaptive aiding system (Parasuraman, 2003).

Among the many methods of measuring mental workload, physiological measures offer promise because they can be more closely linked to brain function, which as mentioned previously, needs to be assessed prior to performance breakdown. Neural and psychophysiological...
Neurophysiological and psychophysiological variables are known to respond to cognitive demand in a relatively predictable manner (Fairclough et al., 2005). Direct measures of central nervous system function such as electroencephalography (EEG) and event-related brain potentials (ERPs) have been particularly strong candidates for accurate, objective measures of operator workload. Increasing task difficulty, for instance, is known to be associated with EEG changes such as increased power in the beta bandwidth, increased theta activity at frontal sites and the suppression of alpha activity (Brookings et al., 1996; Gevins et al., 1998; Klimesch, 1999). Although EEG has many excellent qualities for monitoring mental workload, including superior temporal resolution, it is limited in its capacity for spatial resolution. In addition, setup up time and susceptibility to motion artifact should be considered for minimally intrusive deployment. Optical imaging techniques offer a viable alternative for operator cognitive state monitoring.

Functional near infrared (fNIR) spectroscopy provides a potential portable system for measuring mental workload under work field conditions. fNIR is safe, highly portable, user-friendly and relatively inexpensive, with rapid application times and near-zero run-time costs (Bunce et al., 2006; Coyle et al., 2007; Izetoglu et al., 2004; Strangman et al., 2002; Villringer and Chance, 1997). The most commonly used form of fNIR uses light, introduced at the scalp, to measure changes in blood oxygenation as oxy-hemoglobin (HbO2) converts to deoxy-hemoglobin (HbR) during neural activity, i.e., the cerebral hemodynamic response. fNIR has been shown to compare favorably with other functional imaging methods (Huppert et al., 2006) and demonstrates solid test–retest reliability for task-specific brain activation (Plichta et al., 2007). fNIR provides good spatial localization compared to EEG, on the order of 1 cm2, and is amenable to integration with EEG/ERPs (Gratton and Fabiani, 2008; Strangman et al., 2002; Villringer et al., 1993).

fNIR’s capacity for spatial resolution has important benefits for use as a measure of mental workload in neuroergonomic studies as people develop expertise in a task. Such a capability is important for neuroergonomics given that human operators in work settings typically have developed considerable expertise in the tasks that they have to perform for their job, whereas laboratory studies with college students generally examine performance of untrained participants on artificial tasks. In contrast, neuroergonomics attempts to understand brain mechanisms underlying performance in experienced workers in operationally-relevant tasks (Parasuraman, 2003). One aspect of expertise is the development of automaticity. Current models of automaticity related to the development of expertise in certain tasks suggest that there are shifts in the functional neuroanatomy of task performance that support ongoing cognitive effort assessments. Operator skill and mental workload are generally inversely related (Gopher and Kimchi, 1989; Liu and Wickens, 1994). This inverse relationship between expertise and the cognitive demand of a given task impacts the accuracy and interpretation of psychophysiological variables as measures of mental workload (O’Donnell and Eggemeier, 1986; Wierville and Eggemeier, 1993). However, as automaticity develops in various tasks, shifts in the functional neuroanatomy of task performance free up attentional resources, largely associated with the prefrontal cortex, to allow for performance on the other tasks. The development of expertise, or automaticity, can be characterized as the freeing up of the limitations on those attentional resources (Schneider and Shiffrin, 1977; Shiffrin and Schneider, 1977).

There is an extensive literature regarding the effect of practice and expertise on the functional anatomy of task performance. The development of expertise has been studied across a wide range of motor, visuomotor, perceptual and cognitive tasks, and from disparate research perspectives (Ericsson, 2006). Four main patterns of practice-related activation change have been identified in this literature (Kelly and Garavan, 2005), including either 1) increased or 2) decreased activation in the brain areas involved in task performance, or 3) a functional redistribution of brain activity, in which some initial areas of activation increase while others decrease, or 4) a functional reorganization of brain activity, in which the pattern of increasing and decreasing activation occurs across distinct brain areas in addition to the initial areas.

According to Petersen et al. (1998), a set of attention and control areas of the brain (the scaffolding) is used to support or cope with novel demands during unskilled, effortful performance. With practice processes or associations that can be more efficiently stored and accessed elsewhere are offloaded to those areas, after which the scaffolding network is pruned away. Decreased reliance on the ‘scaffolding’ is demonstrated by decreased activation in those areas during performance with a concomitant increase in activation brain areas underlying the task-specific processes. Therefore, activations seen earlier in practice involve generic attentional and control areas. Prefrontal cortex (PFC), anterior cingulate cortex (ACC) and posterior parietal cortex (PPC) are thought to be the predominant constituents of the scaffolding consistent with theories of PFC function and the involvement of these areas in the distributed working memory system. Increases associated with highly practiced performance are primarily seen in task-specific areas such as representational cortex — primary and secondary sensory or motor cortex, or in areas related to the storage of those representations, such as the parietal or temporal cortex. A majority of the studies examining task practice have found decreases in the extent or intensity of activations with ongoing practice, particularly in the attention and control areas (Kelly and Garavan, 2005). This finding is true whether the task is primarily motor (e.g. golf swing, (Milton et al., 2004)) or primarily cognitive in nature, as in the Tower of London task (Beauchamp et al., 2003). Decreases in activation are thought to represent a contraction of the neural representation of the stimulus (Poldrack, 2000) or a more precise functional circuit (Garavan et al., 2000). This finding provides an important overlap with the literature on expertise. Evidence suggests that overall, experts show lower brain activity relative to novices, particularly in prefrontal areas (Milton et al., 2004). Both practice and the development of expertise (the latter of which includes individual differences in ability) typically involve decreased activation across attentional and control areas, freeing these neural resources to attend to other incoming stimuli or task demands. As such, measuring activation in these attentional and control areas relative to task performance can provide an index of level of expertise. One way to conceptualize this approach is that a relative quantification of the attentional and control resources necessary to perform at a given level can serve as an index of the trainee’s neural “reserves,” a capacity that can be used to perform effectively under greater situational demands.

A neuronal measure of expertise must be defined in relationship to behavioral performance. However, at a given level of performance, a neuronal measure of expertise that monitors the attentional and control resources the individual must utilize to maintain that level of performance could be expected to differentiate between relatively lesser and greater expertise. That is, even at 98–100% performance levels, where performance measures cannot differentiate between trainee capacities, some individuals will be performing at close to their peak performance because their mental workload is close to the limit of their capacity, whereas others will be well below their workload capacity (Bunce et al., in press). An assessment of the cortical activity necessary to perform at a given level would indicate the cognitive resources available for more situational demands, consistent with greater expertise.

In this paper, we provide initial evidence from two studies that fNIR can be used in ecologically valid environments to assess: 1) mental workload levels of operators performing standardized (n-back) and
complex cognitive tasks (ATC, and 2) expertise development through learning a complex cognitive and visuomotor tasks (UAV). Previously, we have reported that the hemodynamic response over the dorsolateral and ventrolateral prefrontal cortex, assessed using fNIR, was responsive to mental workload in a realistic command and control task (Izzetoglu et al., 2004). In the current report, we examined the role of mental workload and expertise (relative levels of practice) on the hemodynamic response over dorsolateral and ventrolateral prefrontal cortex. The first study provides data on the hemodynamic response of certified air traffic controllers as they complete both a standard n-back task and realistic simulations of air traffic control situations with varying degrees of difficulty. The second study examined the impact of developing expertise on the dorsolateral and ventrolateral hemodynamic responses of novice college students as they trained in a simulated UAV operational environment.

The n-back task served as a baseline condition from which to establish the utility of fNIR to measure changes in mental workload. The n-back is a well-characterized paradigm with robust correlations between level of difficulty and cortical activation, including dorsolateral and ventrolateral prefrontal cortex (Owen et al., 2005). The n-back task requires that sequentially presented items (letters, spatial positions, or patterns) be evaluated for their identity to an element that was presented 0, 1, 2, or 3 items previously. As such, the task requires encoding, temporary maintenance and rehearsal, tracking of serial order, updating, and comparison and response processes, functions of working memory and attention, all of which are part of complex systems monitoring. Owen et al.’s meta-analysis of 24 n-back studies has shown that greater workload is consistently associated with greater cortical activation, providing a benchmark for evaluating the utility of fNIR for this purpose. A priori hypotheses regarding regions of interest for the n-back were derived from Owen et al.’s findings, particularly over left and right ventrolateral cortex and the edges of the dorsal poll.

Material and methods

Continuous wave fNIR system

Throughout all experiments, the prefrontal cortex of the participants were monitored using a continuous wave fNIR system first described by Chance et al. (1998), further developed at Drexel University (Philadelphia, PA), manufactured and supplied by fNIR Devices LLC (Potomac, MD; www.fnirdevices.com). The system was composed of three modules: a flexible headpiece (sensor pad), which holds light sources and detectors to enable a fast placement of all 16 optodes; a control box for hardware management; and a computer that runs the data acquisition (Fig. 1).

The sensor has a temporal resolution of 500 milliseconds per scan with 2.5 cm source-detector separation allowing for approximately 1.25 cm penetration depth. The light emitting diodes (LED) were activated in turn one light source at a time and the four surrounding photodetectors around the active source were sampled. The positioning of light source and detectors on the sensor pad yielded a total of 16 active optodes (channels) and was designed to monitor dorsal and inferior frontal cortical areas underlying the forehead (Ayaz, 2010; Bunce et al., 2006; Izzetoglu et al., 2005). COBI Studio software (Drexel University) was used for data acquisition and visualization (Ayaz et al., in press). During the n-back and UAV tasks, a serial cable between the fNIR data acquisition computer and stimulus presentation computer was used to transfer time synchronization signals for marking the onset of sessions and stimuli.

Changes in light absorption, as measured by fNIR at each of the two wavelengths, can be used to calculate relative changes of HbO2 and HbR dependent measures versus time by using the modified Beer–Lambert Law. Jobsis (1977) first reported that near infrared light could diffuse through the intact scalp and skull and can be used for tracing hemoglobin concentration changes within the brain. Cope et al. (1988) proposed a modified Beer Lambert Law to quantify changes in chromophore concentration. The method relies on a simple basic principle that near infrared light is emitted from the light source and travels through the tissue, and undergoes multiple scattering and partial absorption (Cope et al., 1988; Cope and Delpy, 1988; Obrig et al., 2000). Within the near infrared range of light, the biological tissue is transparent enough that the spectroscopic analysis allows detection of the two main chromophores, HbO2 and HbR.

The optical density (OD) for a specific input wavelength (\( \lambda \)) is the logarithmic ratio of input light intensity (\( I_{in} \)) and detected (\( I_{out} \)) light intensity. OD is also related to the concentration (c) and absorption coefficient (\( \varepsilon \)) of chromophores, the distance (d) between the light source and detector, a differential pathlength factor (DPF) for the increase in path length due to high scattering, plus a constant attenuation factor (G). G is a factor that accounts for the measurement geometry and scattering. Scattering is high within the tissue but is generally assumed to be constant throughout the measurement (Obrig and Villringer, 2003). This simplification allows constant G and DPF factors.

\[
OD_{\lambda} = \log \left( \frac{I_{in}}{I_{out}} \right) \approx \varepsilon_{\lambda} \cdot c \cdot d \cdot DPF_{\lambda} + G
\]

Having the same \( I_{in} \) at two different time instances and detected light intensity during baseline (\( I_{test} \)) and during performance of the task (\( I_{test} \)), the difference in OD is

\[
\Delta OD_{\lambda} = \log \left( \frac{I_{test}}{I_{test}} \right) = \varepsilon_{\lambda}^{HbR} \Delta \cdot \varepsilon_{\lambda}^{HbR} \cdot d \cdot DPF_{\lambda} + \varepsilon_{\lambda}^{HbO_2} \cdot \Delta \cdot \varepsilon_{\lambda}^{HbO_2} \cdot d \cdot \Delta \cdot DPF_{\lambda}
\]

Measuring the OD at two different wavelengths gives

\[
\begin{bmatrix}
\Delta OD_{\lambda 1} \\
\Delta OD_{\lambda 2}
\end{bmatrix} =
\begin{bmatrix}
\varepsilon_{\lambda 1}^{HbR} \varepsilon_{\lambda 1}^{HbO_2} \\
\varepsilon_{\lambda 2}^{HbR} \varepsilon_{\lambda 2}^{HbO_2}
\end{bmatrix}
\begin{bmatrix}
\Delta \varepsilon_{\lambda 1}^{HbR} \Delta \varepsilon_{\lambda 1}^{HbO_2} \\
\Delta \varepsilon_{\lambda 2}^{HbR} \Delta \varepsilon_{\lambda 2}^{HbO_2}
\end{bmatrix}
\begin{bmatrix}
\Delta \cdot \varepsilon_{\lambda 1}^{HbR} \Delta \cdot \varepsilon_{\lambda 1}^{HbO_2} \\
\Delta \cdot \varepsilon_{\lambda 2}^{HbR} \Delta \cdot \varepsilon_{\lambda 2}^{HbO_2}
\end{bmatrix}
\]

This equation set can be solved for concentrations if the 2×2 matrix is non-singular. Typically, the two wavelengths are chosen i) within 700–900 nm where the absorption of HbO2 and HbR are dominant as compared to other tissue chromophores, and ii) below and above the isosbestic point (~800 nm) where absorption spectra of HbO2 and HbR cross. The fNIR sensor used in this study had LED light sources with peak wavelengths at 730 nm and 850 nm.

Study 1 — mental workload assessment

Participants

Twenty-four certified professional controllers (CPCs) between the ages of 24 to 55 volunteered for this study. All participants were non-supervisory CPCs with a current medical certificate and had actively controlled traffic in an Air Route Traffic Control Center for a duration of 3 to 30 years. Prior to the study, all participants provided a written informed consent.

Experiment protocol

Participants were asked to complete two tasks: a visual identity n-back and two types of ATC tasks. The n-back task was a standardized working memory and attention task with four incremental levels of difficulty (Smith and Jonides, 1997). Participants were asked to monitor stimuli (single letters) presented on a screen serially and click a button when a target stimulus arrived. Four conditions were used to incrementally vary working memory load from zero to three items. In the 0-back condition, participants responded to a single target letter (e.g., “X”) with their dominant hand (pressing a button to identify the stimulus). In the 1-back condition, the target was defined...
as any letter identical to the one immediately preceding it (i.e., one trial back). In the 2-back and 3-back conditions, the targets were defined as any letter that was identical to the one presented two or three trials back, respectively. The total test included seven sessions of each of the four n-back conditions (hence, a total of 28 n-back blocks) presented in a pseudo-random order. The task was designed and presented in E-prime (Psychology Software Tools).

For the ATC task, each CPC controlled traffic on workstations with a high-resolution (2048 × 2048), 29 in. radarscope, keyboard, trackball, and Direct Access Keypad for 10 min (see Fig. 2). Simulated air traffic was presented using the DESIREE ATC simulator and the TGF systems developed by software engineers at the FAA William J. Hughes Technical Center (Willems et al., 2010).

Two types of communications between the CPCs and pilots, either voice (VoiceComm) or data (DataComm) communications were used in ATC simulations in a pseudo-random order (Willems et al., 2006, 2010). For each communication type, task difficulty was varied by the number of aircraft in each sector, containing 6, 12 or 18 aircraft. Three simulation pilots supported each sector within voice-based scenarios and entered data at their workstations to maneuver aircraft, all based on controller clearances.

Data analysis

For each participant, raw fNIR data (16 optodes × 2 wavelengths) were low-pass filtered with a finite impulse response, linear phase filter with order 20 and cut-off frequency of 0.1 Hz to attenuate the high frequency noise, respiration and cardiac cycle effects (Ayaz et al., 2010; Izzetoglu et al., 2005). Saturated channels (if any), in which light intensity at the detector was higher than the analog-to-digital converter limit were excluded.

Using time synchronization markers, fNIR data segments for rest periods and task periods (28 sessions per participant for n-back task, 6 sessions per participant for ATC task,) were extracted. Oxygenation changes for each 16 optodes were calculated separately using the Modified Beer Lambert Law (MBLL) for task periods with respect to rest periods at the beginning of each task (Ayaz et al., 2010). As feature
extraction, the average oxygenation change (HbO2 – HbR) for each session was used as the dependent measure (Izzetoglu et al., 2004). For statistical channel-wise analysis of the ATC task, 2 (Communication: Data-based, Voice-based) × 3 (Task Difficulty: 6, 12, 18 aircraft) ANOVA with repeated measures on both factors for average self-reported ratings and oxygenation changes were applied for each optode. For N-back task, the main effect of condition was tested using one-way repeated measures ANOVA. Geisser–Greenhouse correction was used for sphericity and Tukey’s post hoc tests to determine the locus of main effects, with the significance criterion set at α = 0.05. For multiple comparison correction, False Discovery Rate (FDR) approach was used (Benjamini and Hochberg, 1995). This FDR based procedure has been reported to provide better balance between specificity and power than other available methods for multi-channel near-infrared spectroscopy functional neuroimaging data (Singh and Dan, 2006).

Normalization by linear modeling. A personalized model was developed based on the fNIR responses for the standardized task to quantify the responses measured during ATC tasks. A separate first order polynomial regression model was established for each participant and the model parameters were estimated by using minimum (min) and maximum (max) of the respective participant’s fNIR n-back data. This individualized model, mapped oxygenation (of ATC task) through an affine transformation that preserved the collinearity, i.e., it applied the same scaling and translation for all conditions for that subject. This model transforms ATC oxygenation values on a standardized n-back conditions’ (0 to 3) axis by using the minimum and maximum n-back oxygenation values as a scale. The general model:

\[ Y_i = \beta_0 + \beta_1 X_i \]

where X is the oxygenation value and Y is the normalized output response. \( \beta_0 \) and \( \beta_1 \) are scalar model parameters (specific to a participant) estimated by using two (n-back-oxygenation, n-back-condition) coordinates, where the n-back-condition is either 0, 1, 2 or 3. Using minimum and maximum oxygenation (and respective condition), parameters can be solved from the following equation set:

\[
\begin{align*}
\beta_0 & = \frac{n_{\text{min}} \cdot \text{oxy}_{\text{max}} - n_{\text{max}} \cdot \text{oxy}_{\text{min}}}{n_{\text{max}} - n_{\text{min}}} \\
\beta_1 & = \frac{n_{\text{max}} \cdot \text{oxy}_{\text{min}} - n_{\text{min}} \cdot \text{oxy}_{\text{max}}}{n_{\text{max}} - n_{\text{min}}} 
\end{align*}
\]

Finally, using this model, ATC oxygenation values for all 6 (2 Communication × 3 Task Difficulty) conditions of the participants were transformed.

Study II – assessment of expertise development

Participants
Seven college students between the ages 21 to 28 with no prior flight simulator experience volunteered for this study. One participant was disqualified due to insufficient ability to perform basic flight tasks and one had to travel abroad before completing all 9 days. Prior to the study, all participants signed informed consent forms.

Experimental protocol
Participants practiced approach and landing scenarios while piloting a virtual unmanned aerial vehicle (UAV) in a flight simulator. The scenarios were designed to expose novice subjects to realistic and critical tasks for a UAV ground operator directly piloting an aircraft. The first scenario was a turn-to-approach task, in which the pilot flies through several waypoints on an approach to land at an airfield. The second scenario was a landing task, in which the pilot performs the actual touchdown. In both scenarios, subjects were told to fly as smoothly as possible, learn the optimal paths, cope with crosswinds, and operate within certain speed and bank angle constraints (see Fig. 3).

The flight simulator system was designed in house and consisted of a custom desktop computer with high performance graphics capabilities, semi-immersive angled triple-monitor display, Thrustmaster HOTAS Cougar joystick-and-throttle system and a CH Pro Pedals rudder pedal.

Fig. 3. Unmanned aerial vehicle simulators on triple panel display running turn-to-approach task (top row) and task landing (bottom row) screenshots.
system. The scenarios were rendered on Microsoft Flight Simulator X and FS Recorder, an add-on for Flight Simulator X, was integrated to record behavioral data during the simulated flights. The experiment protocol involved a total of nine sessions per subject, one session per day. The first session on day 1 was to allow subjects to become acquainted with the flight simulator; by the end of this session, they needed to demonstrate basic understanding of flight simulator controls. Study data were collected during the following eight practice sessions.

Practice sessions consisted of ten repetitions each of two scenarios (one approach, one landing), a total of twenty flights per session, for a total of 160 trials per participant over the 8 days. Participants provided subjective mental effort and performance evaluation using the NASA Task Load Index (TLX) questionnaire (Hart and Staveland, 1988). Each session lasted 2 to 3 h, with no more than one session per day.

Data analysis

UAV fNIR data was pre-processed using the low-pass filters and saturated channel attenuation criteria as described in Study 1 above. fNIR data epochs for the rest and task periods were extracted from the continuous data using time synchronization markers. Blood oxygenation and volume changes within for each 16 optodes were calculated using the MBLL for task periods with respect to rest periods at beginning of each task (Ayaz et al., 2010; Izzetoglu et al., 2004).

The main effect for practice level was tested using one-way repeated measures ANOVA, with Subject and Practice Level designated as fixed effects factors. Geisser–Greenhouse (G–G) correction was used when violations of sphericity occurred in the omnibus tests. Tukey’s post hoc tests were used to determine the locus of the main effects with a 0.05 significance criterion. Three practice levels were defined for each participant: the beginner phase included days 2 through 4, intermediate phase included days 5 through 7 and the advanced phase included days 8 and 9. Number Cruncher Statistical Software (NCSS) 2007 (www.ncss.com) was used for the statistical tests. As in Study 1, False Discovery Rate (FDR) approach was used for multiple comparison correction (Benjamini and Hochberg, 1995; Singh and Dan, 2006).

Results

n-back — standardized working memory and attention task

Behavioral measures

The behavioral data for the n-back were submitted to a one-way repeated measures ANOVA. The results indicated a significant main effects of task difficulty (0-, 1-, 2- and 3-back conditions) for accuracy ($d'$) ($F_{3,69} = 40.68$, $p < 0.001$, $\eta^2_p = 0.639$) and reaction time ($F_{3,69} = 42.76$, $p < 0.001$, $\eta^2_p = 0.65$). Accuracy and reaction time for all participants are presented in Fig. 4. Tukey post hoc tests showed that for accuracy ($q_{0.05/2, 69} = 3.72$, $p < 0.05$), the 3-back was significantly lower than all other conditions and the 2-back differed from 0-back. Results for reaction time ($q_{0.05/2, 69} = 3.72$, $p < 0.05$) showed that the 3-back and 2-back conditions were significantly slower than the 1-back and 0-back conditions.

fNIR measures

A repeated measures ANOVA performed on the fNIR data revealed that reliable changes in oxygenation as a function of n-back condition occurred only at optode #2 ($F_{3,69} = 4.37$, $p < 0.05$, $\eta^2_p = 0.16$) (see Fig. 4).

Fig. 4. N-back task: average response time for each condition (left, top) and correct response ratio for each condition (right, top). Average oxygenation changes (bottom) of all subjects (24 participants, and 28 trials for each participant) with increasing task difficulty. Error bars are standard error of the mean (SEM).
This site, close to AF7 in the International 10–20 System, is located within the left PFC (inferior frontal gyrus). In correspondence with previous studies, increase in activation level was found with increasing task difficulty during task performance. Moreover, the significant region within left PFC in this study was implicated in many previous studies of n-back task with Positron Emission Tomography (PET) (Reuter-Lorenz et al., 2000; Smith et al., 1996), fMRI (Cohen et al., 1997; Owen et al., 2005) and fNIR (Schreppel et al., 2008). Post hoc analyses showed that the 3-back task elicited greater oxygenation increases than either the 0- or the 1-back conditions ($q_{0.05/2, 69} = 3.72, p < 0.05$).

Simulated air traffic control tasks

Self-reported ratings

Subjective workload ratings during the air traffic control tasks were submitted to a 2 (Communication: Data-based, Voice-based) × 3 (Task Difficulty: 6, 12, 18 aircraft) repeated measures ANOVA. The interaction between aircraft number and communication type was significant ($F_{2,46} = 4.66, p < 0.05, \eta^2_p = 0.167$) and is depicted in Fig. 5. Both main effects were also significant, Task Difficulty, denoted by number of aircraft ($F_{2,46} = 6.79, p < 0.05, \eta^2_p = 0.223$) and Communication ($F_{1,23} = 4.53, p < 0.05, \eta^2_p = 0.165$). Tukey post hoc tests for Task Difficulty ($q_{0.05/2, 42} = 3.44, p < 0.05$) showed that the 18 aircraft condition had significantly higher workload ratings than the 6 and 12 aircraft conditions.

fNIR measures

Two subjects were excluded from analyses of fNIR measures of hemodynamic response during the air traffic control tasks due to high motion artifact and low signal-to-noise ratios. A 2 (Communication: Data-based, Voice-based) × 3 (Task Difficulty: 6, 12, 18 aircraft) repeated measures ANOVA was applied on average oxygenation change for the remaining participants.

The significant measurement location was optode #8 within the medial PFC/frontopolar cortex, and there were two significant main effects, Task Difficulty ($F_{2,42} = 4.52, p < 0.05, \eta^2_p = 0.177$) and Communication ($F_{1,21} = 5.03, p < 0.05, \eta^2_p = 0.193$) which are depicted in Fig. 5. Tukey post hoc tests for Task Difficulty ($q_{0.05/2, 42} = 3.44, p < 0.05$) showed that the 18 aircraft condition had greater oxygenation increase relative to the 6 aircraft condition. Practice effect was investigated by grouping tasks in the order of completion. Each of the six conditions was presented to subjects in a pseudo-random order so that each condition’s order can be from 1 to 6. As expected, repeated measures ANOVA on task order indicated no significant effect with oxygenation ($F_{5,105} = 0.64, p > 0.5$) and it is depicted in Fig. 5. The individualized linear model trained by n-back data was used to normalize oxygenation changes (see Fig. 5). Fitted ATC responses indicated that this normalization improved the spatial localization of activity pattern in dorsolateral PFC and represent a monotonic increase with increasing task difficulty within anterior medial PFC ($F_{2,21} = 11.26, p < 0.05, \eta^2_p = 0.518$).

Fig. 5. ATC task: average subjective workload ratings (top, left) and average oxygenation changes (top, right) for data-based and voice-based for 6, 12 and 18 aircraft conditions. Average oxygenation for all tasks grouped by task order (bottom, left) indicates no practice effect. Average oxygenation changes normalized by the individualized linear model (bottom, right). (N = 22). Error bars are SEM.
Piloting unmanned aerial vehicle tasks

Self-reported ratings

The NASA TLX index results for the UAV tasks were analyzed using one-way repeated measures ANOVA. The results indicated a significant main effect of practice level (beginner/intermediate/advanced conditions) for mental demand ($F_{2,8} = 17.87, p < 0.01, \eta^2_p = 0.817$), effort ($F_{2,8} = 16.32, p < 0.01, \eta^2_p = 0.803$) and frustration ($F_{2,8} = 8.60, p < 0.01, \eta^2_p = 0.682$). Both mental demand and perceived effort followed a monotonic decrease from beginner to advanced phase (see Fig. 6). Tukey post hoc tests revealed that mental demand ($q_{0.05/2, 8} = 4.04, p < 0.05$) for the beginner phase was significantly higher than the other conditions, whereas results for effort ($q_{0.05/2, 8} = 4.04, p < 0.05$) showed the beginner phase required higher effort than the other conditions. Similar to mental demand and effort, the frustration demand ($q_{0.05/2, 8} = 4.04, p < 0.05$) was higher in the beginner phase than in the intermediate or advanced phases.

Behavioral measures

Behavioral performance was calculated as root-mean-square (RMS) deviation from the 4th-order polynomial fits to the path for latitude and longitude. For the approach task, there was a significant monotonic decrease in deviation from the optimal path when comparing practice level (beginner through advanced) for latitude ($F_{2,8} = 18.30, p < 0.05, \eta^2_p = 0.793$), longitude ($F_{2,8} = 17.58, p < 0.05, G-G, \eta^2_p = 0.815$) bank ($F_{2,8} = 15.80, p < 0.05, G-G, \eta^2_p = 0.798$) and heading ($F_{2,8} = 16.41, p < 0.05, \eta^2_p = 0.804$), see Fig. 7. Post hoc analyses confirmed that an error at the beginner phase was significantly higher than the other two phases ($q_{0.05/2, 8} = 4.04, p < 0.05$). Similarly, for the landing task, errors for the following parameters were significant: latitude ($F_{2,8} = 6.65, p < 0.05, \eta^2_p = 0.624$), longitude($F_{2,8} = 8.52, p < 0.05, \eta^2_p = 0.68$), bank ($F_{2,8} = 8.30, p < 0.05, \eta^2_p = 0.675$) altitude ($F_{2,8} = 6.76, p < 0.05, \eta^2_p = 0.628$) and heading ($F_{2,8} = 8.31, p < 0.05, \eta^2_p = 0.675$), see Fig. 8. Post hoc analyses confirmed that an error at the beginner phase was significantly greater than the error at the other two phases ($q_{0.05/2, 8} = 4.04, p < 0.05$).

fNIR measures

For each type of UAV task (approach and landing), the average total hemoglobin (HbT) concentration changes throughout the practice levels (beginner/intermediate/advanced) were analyzed using one-way repeated measures ANOVA. Only significant measurement location was optode #2, that is close to AF7 in the International 10–20 System, located within the left PFC (inferior frontal gyrus). The response was significant for both the approach task ($F_{2,8} = 6.08, p < 0.05, \eta^2_p = 0.603$) and the landing task ($F_{2,8} = 7.70, p < 0.05, \eta^2_p = 0.658$), see Fig. 9. Post hoc analyses confirmed that activation for the beginner phase was significantly higher than the other two phases ($q_{0.05/2, 8} = 4.04, p < 0.05$).

Discussion

The n-back task served as a baseline condition from which to establish the utility of fNIR to measure changes in mental workload. As task difficulty increased, the results show a monotonic decrease in accuracy and an increase in response time. The fNIR results were also sensitive to task difficulty specifically at left inferior frontal gyrus. These results are in line with earlier PET and fMRI studies that have used the n-back task (Badre and Wagner, 2007; Cohen et al., 1997; Osaka et al., 2007; Owen et al., 2005; Smith et al., 1996; Smith and Jonides, 1997) and also fNIR studies (Izzetoglu et al., 2004; Schreppel et al., 2008). Moreover, the findings help establish the validity of fNIR for measuring cortical changes in oxygenation associated with

---

Fig. 6. NASA TLX indices for UAV tasks (5 participants, 8 days per participant): mental demand (left, top), effort (right, top) and frustration (bottom). Error bars are SEM.
cognitive workload in this subject population in a simple laboratory task that has been widely used in the cognitive neuroscience research literature.

The primary goal of the first study was to test if the same methodology could be used for assessing mental workload in a complex cognitive task that very closely simulated the activities of air traffic controllers. Experienced controllers were used as the subject population. The average oxygenation level, as measured by fNIR in the anterior medial PFC/frontopolar cortex (imaged at optode #8), increased monotonically with increased task difficulty (number of aircraft that the controllers had to monitor and control) as illustrated in Fig. 5. The results provide strong evidence that activation in this brain region provides a valid measure of mental workload in this realistic air traffic control task. Furthermore, the distinction of focal optodes for the n-back (working memory) and air traffic control task (planning/decision making) is also in parallel with the fMRI findings of functional dissociation of lateral and medial PFC (Bechara et al., 1998; Koechlin et al., 2000; Simons et al., 2005). The fNIR results from the main effect of communication type, corroborates with subjective effort ratings and implicates higher brain activation for VoiceComm with a small to moderate effect size (Cohen’s d = 0.28). Moreover, the VoiceComm condition resulted in higher oxygenation than DataComm for each aircraft condition suggesting that, given the same cognitive workload (identical scenarios), DataComm required fewer cognitive resources. This is also in line with the self-reported difficulty ratings and previous studies (Hah et al., 2006; Willems et al., 2006).

These results must be interpreted with caution, however, as not all areas of the brain that are involved in air traffic control could be measured with fNIR technology. It is possible that mental workload was shifted to other areas of the brain such as parietal cortex that were not being monitored under the DataComm condition, rather than a reduction in working memory resources. This hypothesis requires further evaluation. However, combined with performance evaluations, the fNIR data provides preliminary evidence that DataComm may require fewer cognitive resources than VoiceComm by lowering the demand on working memory and attentional resources.

For both UAV tasks there was a reduction in the fNIR measures, shown in Fig. 9, which were significantly different across practice levels and matches the same trends reported in behavioral performance (Figs. 7 and 8) and self-reported measures (Fig. 6). A valid hypothesis can be derived from the evidence that expertise tends to be associated with overall lower brain activity relative to novices, particularly in prefrontal areas (Milton et al., 2004). Both practice and the development of expertise typically involve decreased activation across attentional and control areas, freeing these neural resources to attend to other incoming stimuli or task demands. As such, measuring activation in these attentional and control areas relative to task performance can provide an index of level of expertise and illustrate how task-specific practice influences the learning of tasks. The differences in activation of the attentional and control regions of the prefrontal cortex may also indicate neural plasticity as a function of task-specific practice (Kelly and Garavan, 2005). However, the level of neural activation must be interpreted in light of the performance and the “objective” difficulty of the task, as decreases in sustained mental effort during a difficult task result in a rapid drop in prefrontal oxygenation (Izzetoglu et al., 2004).

In summary, fNIR is a portable, safe, affordable and negligibly intrusive optical brain monitoring technology that can be used to measure hemodynamic changes in the prefrontal cortex. Changes in blood oxygenation in the dorsolateral and ventrolateral PFC, as measured by fNIR, were shown to be associated with increasing...
cognitive workload. The results further indicate that text-based communications required less brain activation of the operator than legacy voice based communication systems. These fNIR results are in agreement with the subjective assessments of operators and earlier studies (Izzetoglu et al., 2004; Owen et al., 2005; Willems et al., 2006). In parallel, the second study results illustrate that focused practice on the UAV tasks reduces neural activation of the left dorsolateral PFC and provides support for neural changes that were previously reported as a function of practice (Kelly and Garavan, 2005). The overall sensitivity of the hemodynamic response to changes in level of task difficulty, however, is still an open question. Whereas large changes in task difficulty resulted in statistically different levels of oxygenation change, smaller differences in task difficulty were not reliably differentiated with these methods. The overall capacity for workload sensitivity using the hemodynamic response has yet to be fully explored, and more sophisticated techniques, or indeed, an integration with EEG techniques, may yield more refined results in the future.

**Conclusions**

This paper provides important albeit initial information about fNIR measures of dorsolateral PFC hemodynamic response and its

Fig. 8. UAV landing task behavioral performance results: deviation from optimal for latitude (top, left), longitude (top, right), bank angle (bottom, left) and heading (bottom, right). (5 participants, 10 trials per day, 8 days). Error bars are SEM.

Fig. 9. Total hemoglobin concentration changes for the UAV approach (left) and landing (right) tasks. (5 participants, 10 trials per day, 8 days). Error bars are SEM.
relationship to mental workload, expertise, and performance, in a complex multitasking environment. Level of expertise does appear to influence the hemodynamic response in the left dorsolateral PFC, at least for some complex tasks. Since fNIR technology allows the development of mobile, non-intrusive and miniaturized devices, it has the potential to be deployed in future learning and training environments to personalize the training regimen and/or to assess the effort of human operators expend in critical multitasking environments.

Disclosure

fNIR Devices, LLC manufactures the optical brain imaging instrument and licensed IP and know-how from Drexel University. H. Ayaz, S. Bunce, K. Izzetoglu and B. Onaral were involved in the technology development and thus offered a minor share in the new startup firm fNIR Devices, LLC.

Acknowledgments

Authors gratefully acknowledge Prof. Raja Parasuraman for comments on the manuscript, Sehchang Hah, Atul Deshmukh for ATC data acquisition, Justin Mura, Merut Perit Cakir and Adrian Curtin for UAV data acquisition. This work was supported in part by the U.S. Federal Aviation Administration through BAE Systems Technology Solutions Services Inc. under Primary Contract, DTF-A01-00-C-00068 and Subcontract Number, 31–5029862. This investigation was in part funded under a U.S. Army Medical Research Acquisition Activity; Cooperative Agreement W81XWH-08-2-0573. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the funding agencies.

References


This work was supported in part by the U.S. Federal


