Evaluation of a New Functional near Infrared Spectroscopy (fNIRS) Sensor, the fNIRS Explorer™, and Software to Assess Cognitive Workload during Ecologically Valid Tasks

Bethany K. Bracken[1], Colette Houssan[2], John Broach[2], Andrew Milsten[2], Calvin Leather[1], Sean Tobyne[1], Aaron Winder[1] and Mike Farry[1]

[1]*Charles River Analytics, 625 Mount Auburn St, Cambridge, MA, U.S.A.*

[2]*University of Massachusetts Medical School, 55 N Lake Ave, Worcester, MA, U.S.A.*

Keywords: Cognitive Workload, Functional Near Infrared Spectroscopy (fNIRS), Medical Simulation, Training, Ecologically Valid, Real World, Disaster Medicine Training.

Abstract: Medical personnel and first responders are often deployed to dangerous environments where their success at saving lives depends on their ability to act quickly and effectively. During training, non-invasive measurement of cognitive performance can provide trainers with insight into medical students' skill mastery. Functional Near-Infrared Spectroscopy (fNIRS) is a direct and quantitative method to measure ongoing changes in brain blood oxygenation (HbO) in response to a person's evolving cognitive state (i.e., cognitive workload or mental effort) that has only recently received significant attention for use in the real world. The work presented here includes data collection with a new, more portable, rugged design of an fNIRS sensor to test the functionality of this new sensor design and our ability to measure cognitive workload in a medical simulation training environment. To assess sensor and model accuracy, during breaks from the training, participants completed a gold-standard, laboratory task and during training in a medical simulation environment. Linear mixed model ANOVA showed that when we accounted for fixed effects of intercept and slope in our model, there was a significant difference in the HbR Ch1 model for n-back load (coef=0.009, p=0.034), intercept (coef=0.96, p=1.21e-07***), and load (slope) (coef=-0.09, p=0.03). Future work will present data collected across all disaster response medical trainings.

1 INTRODUCTION

Medical personnel are often deployed to a wide range of environments (e.g., sites of earthquakes, hurricanes, and other disasters) where their success at saving lives depends on their ability to act quickly and effectively. They are required to put to use the skills they have learned in the classroom and simulated trainings in some of the most stressful situations imaginable. To be truly effective, personnel must train to ensure skills transfer to environments that are chaotic and require performance over multiple days of sub-standard conditions (e.g., long working hours, sleep deprivation, abnormal food habits). In the field, personnel who experience cognitive overload due to inexperience or lack of skill may hesitate, make judgment errors, or fail to attend to critical situational details. Skills that are not mastered to the point of automatic response will not transfer optimally to these situations, putting patients at risk. Realistic training simulations ranging from classroom to simulated disaster scenarios provide medical teams with the opportunity to efficiently practice and hone medical skills; however, even the most rigorous training cannot ensure that personnel will perform effectively when faced with the aftermath of a disaster. Currently, trainers must infer trainees' competence through behavioural observation alone. This is a challenging task as even highly experienced trainers cannot always reliably determine which trainees have mastered a task to the desired point of automatic response, or whether task execution still requires significant individual cognitive resources that will be exhausted in operational environments.

Non-invasive measurement of cognitive performance can provide trainers with insight into

Bracken, B., Houssan, C., Broach, J., Milsten, A., Leather, C., Tobyne, S., Winder, A. and Farry, M.
Evaluation of a New Functional near Infrared Spectroscopy (fNIRS) Sensor, the fNIRS Explorer™, and Software to Assess Cognitive Workload during Ecologically Valid Tasks.
DOI: 10.5220/0008902701790186
In *Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2020) - Volume 4: BIOSIGNALS*, pages 179-186
ISBN: 978-989-758-398-8; ISSN: 2184-4305

trainee skill mastery without overloading trainers with additional tasks, and support assessment of cognitive measures such as attention and cognitive workload. Objective information on attention and cognitive workload can help trainers understand the skill level of trainees and can provide insight into how much cognitive effort is required for trainees to accomplish certain tasks. (i.e., whether applying a new junctional tourniquet correctly was done effortlessly or still required significant attentional resources). A comprehensive understanding of trainee knowledge acquisition and skill application will both improve educational assessment techniques and increase the cost-effectiveness of current training practices by enabling trainers to focus on areas where trainees and teams require the most improvement.

Non-invasive sensors can be used to supplement methods already used by trainers without significant extra effort and without further encumbering trainees (physically or cognitively). However, the majority of sensors commonly used to assess cognitive measures (e.g., electroencephalography (EEG)) are not designed for real world training environments, are sensitive to motion artifacts (Kerick, Oie, & McDowell, 2009), suffer from large variability across individuals (Mathan, Whitlow, Dorneich, Ververs, & Davis, 2007), and typically require post-hoc processing, preventing trainers from applying the resulting knowledge during training. Many sensors require significant training to learn how to correctly set up, use, and interpret and their measures are difficult to translate into a form that is easily understandable by the trainer (e.g., event-related potentials from EEG experiments). They also do not indicate to trainers which events (e.g., the entry of the 30th casualty, or the application of a new tourniquet) resulted in the highest cognitive workload. These sensor systems are therefore unsuitable for use during live training exercises.

Functional Near-Infrared Spectroscopy (fNIRS) is a quantitative method to measure ongoing changes in brain blood oxygenation (HbO) in response to a person's evolving cognitive state (i.e., cognitive workload or mental effort) (Boas, Elwell, Ferrari, & Taga, 2014; Ferrari & Quaresima, 2012) that has only recently received significant attention for use in the real world. When cognitive workload increases, there is a corresponding increase in prefrontal blood flow that correlates with increased task engagement. Once the task becomes too difficult, there is a decrease in blood flow that correlates with disengagement from the task and decreased performance (Ayaz et al., 2012; Bunce et al., 2011). Assessing cognitive workload with fNIRS when individuals are seated is well established. However, fNIRS sensor devices that can be used to assess cognitive workload during normal activities (e.g., combat medic training) are only recently emerging.

One analogous study used fNIRS during real world navigation where participants had to navigate the Drexel University campus using either Google Glass or a handheld smartphone (McKendrick et al., 2016). A secondary task was conducted concurrently to assess cognitive workload (an auditory version of the n-back). The n-back working memory task (Kirchner, 1958) is a gold-standard working memory task. The participants are presented with one stimulus at a time, and must respond "yes" if the current stimulus matches the one presented "n" items back. For the 1-back condition, this refers to the stimulus presented immediately before it. For the 2-back condition, this refers to the stimulus presented two items previously. Researchers found a decrease in hemodynamic response in right lateral prefrontal cortex (the location in which our fNIRS sensor is positioned) during correct responses.

Standard sensors are large (e.g., full-head), expensive (~$10K), and require heavy equipment (e.g., batteries, laptops). To address this gap, we have designed and developed two new sensors that are smaller and more cost-effective, and is designed to be used outside the laboratory. We previously validated these sensors against other, larger and more expensive systems that in other environments (Bracken, Festa, Sun, Leather, & Strangman, 2019; Bracken, Elkin-Frankston, Palmon, Farry, & Frederick, 2017; Bracken, Palmon, Elkin-Frankston, & Silva, 2018). This paper presents our work to validate one of our fNIRS sensors and our software system to process and model data into estimates of cognitive workload, with a focus on medical student trainers during high-tempo training simulations in order to assist trainers in optimizing learning. The work presented here includes data collection with our next generation of portable, rugged fNIRS sensor to test both the functionality of this new sensor design and our ability to measure cognitive workload in a real-world medical training environment.

2 METHODS

All methods were approved by both the University of Massachusetts Institutional Review Board (IRB) and the United States Department of Defense Human Research Protections Office (HRPO). All participants were fully informed of all elements of the study and completed informed consent forms. We used a rugged, portable, fNIRS sensor, the fNIRS Explorer™, shown in Figure 1. When compared to our previously-developed sensor, the fNIRS Pioneer™, the Explorer is smaller, consists of only one piece of hardware, is more comfortable due to including an adjustable headband, is more rugged with no charging ports or wired connects that could wear out or allow entry of sand, water, or dust.

Figure 1: fNIRS Explorer sensor, bottom view of sensor that gets placed against the person's forehead (left), and side view of the sensor with a pen shown for scale (right).

When cognitive workload increases, there is a corresponding increase in prefrontal blood flow that correlates with increased task engagement. Once the task becomes too difficult, there is a decrease in blood flow that correlates with disengagement from the task and decreased performance (Ayaz et al., 2013, 2012; Bunce et al., 2011). However, sensor location matters. Because we are not using an EEG cap, and because our participants have different hair lines, we decided on the sensor placement in this picture with the optical density sensor (the square on the forehead-facing side of the sensor) positioned ~2 inches above the outside edge of the eyebrow and the two light emitting diodes (the two small circles) positioned medially. Figure 2 shows the preferred positioning of our fNIRS sensors. The Pioneer is shown on the left to demonstrate more clearly position on the forehead. The Explorer is shown on the right as it was positioned for this study.

Figure 2: fNIRS Pioneer sensor to demonstrate clear position on the forehead (left) and fNIRS Explorer as it was positioned for this study (right).

To assess cognitive workload, we are focusing on dorsolateral prefrontal cortex (dlPFC). We acquired data from participants taking part in a Basic Disaster Life Support (BDLS), Advanced Disaster Life Support (ADLS), and/or Disaster Pre-deployment trainings which occur across multiple training environments ranging from in-classroom trainings (lecture format and interactive table-tops sessions) to high-tempo, live-action role-playing simulated disaster events.

Each course allowed time for us to collect data from each student during three levels of the n-back working memory task (Kirchner, 1958) to allow us to optimize accuracy of our cognitive workload models. We added this into the protocol based on results of our data modeling during our related efforts that found that we could increase model accuracy by collecting ground truth data using a well-validated cognitive task in order to train models to account for individual differences (Bracken et al., 2019). Participants completed the n-back on a tablet within the simulation environment during breaks from training.

Although we only present data from the final training, we present the full set of trainings here so that the reader understands the training history of participants, and the full structure of the course. The first training included several classroom-based lectures. This is a full day and covered multiple topics. The second training included a lecture at the beginning of the day followed by splitting students into groups that move through four stations spread across multiple rooms including high-fidelity simulation labs. Each station lasted about an hour and topics included triage, basic life-saving skills (e.g., tourniquet application, airway), mass casualty triage (multiple simulated casualties in a room), donning and doffing personal protective equipment (PPE), and how to manage patient surge and patient flow (including logistics of handling a large surge in patients such as where there are open beds). The third, "pre-deployment", course is spread over several large rooms near the Emergency Department. For the pre-deployment course each station is 60-90 minutes. We planned to concentrate on two of the skills covered across all four types of training: triage and handling patient surges (similar to the mass casualty trainings), and to follow as many students across all three types of trainings as possible.

Here, we will present the results of the second training. Eighteen people attended the second course on May 21-22 2019, and participated in research (9 female; mean age 35, range 27-62). Participants were all recorded during three sessions – a series of lectures followed by interactive group table-top session. Recording sessions were randomized except for the nine people we followed for the triage and surge activities. All 18 completed a baseline assessment (1-, 2-, and 3-back versions of the n-back) on day one.

The enrollment protocol was built based on the expected number of participants and the number of available sessions to record subjects for trainings and n-back testing. We had five time points to record n-backs (registration, break1, lunch, break 2, and break 3) and 14 segments from the four sessions. We assigned four people to each training segment in order; each with a separate Explorer sensor. The Casualty Triage and Public Health & Population Health/Q&A were used for the 10 people who were followed across trainings since their lessons correlated to the lectures and interactive session presented in ADLS and the pre-deployment classes. The scenarios were pre-built into the MEDIC software so that the research assistants managing the training could press record when the subjects had their sensor placed on their heads. The scenarios for n-back were built as groups and before each set of subjects (3-4 at a time) took their n-back we added the subject number and which sensor was associated with each subject before we started collecting data. N-backs were completed on four laptops and the start of each n-back was recorded. As people checked in they completed their demographic survey, and a post-training session survey as soon as we collected the sensor from them.

3 RESULTS

To assess sensor and model accuracy, during breaks from the training, participants completed a gold-standard, laboratory task and during training in a medical simulation environment. In this paper, we present only the data from that gold-standard task, the n-back. In future papers, we will publish results of the training simulation scenarios.

We started by visually inspecting the data from the n-back task and the trainings (see Figure 7), and running data through our standard processing pipeline developed to handle data collected in non-laboratory conditions (e.g., when participants are not instructed to remain still, and data are collected on mobile devices (e.g., tablets) or moving around their environment taking part in realistic activities). This processing procedure applies advanced motion correction algorithms including wavelet filtering, movement artifact removal algorithm (MARA), and acceleration-based movement artifact reduction algorithm (AMARA) (Metz, Wolf, Achermann, & Scholkmann, 2015; Molavi & Dumont, 2012; Scholkmann, Spichtig, Muehlemann, & Wolf, 2010).

analysing n-back data, pooling data across all participants. Figure 3 shows n-back response accuracy (correct responses) versus n-back level (load). There was a statistically significant decrease as determined by a one-way ANOVA ($F(2,24)=4.27$, $p=0.03$). A Tukey posthoc test with corrections for multiple comparisons revealed that accuracy significantly decreased between the 1-back and 3-back condition ($t=-2.913$; $p=0.02$). We did not find a relationship between response time with load (one-way ANOVA; $F(2,24)=0.10$, $p=0.90$).

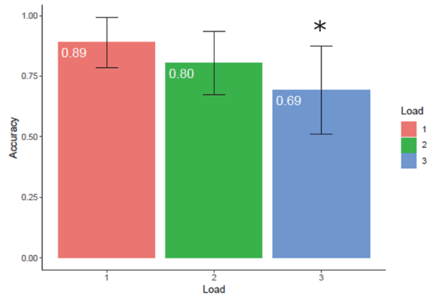

Figure 3: N-back performance—accuracy versus n-back load. There was a decrease with load (a one-way ANOVA ($F(2,24)=4.27$, $p=0.03$)). Tukey posthoc test revealed that accuracy decreased between 1-back and 3-back; no difference between 1-back and 2-back. Error bars are standard deviation.

We next compared blood oxygenation changes across n-back load conditions. The Explorer sensor collects data at two locations, so there are four variables to consider: oxygenated blood signal (HbO) and deoxygenated blood signal (HbR) from channel 1 (the location of assessment that is more lateral) and HbO and HbR from channel 2 (the location of assessment that is more medial). Figure 4 shows blood oxygenation versus n-back load. Oxygenated blood signal from Explorer channel 1 is shown in salmon (HbO Ch1); oxygenated blood signal from Explorer channel 2 is shown in green (HbO Ch2); deoxygenated blood signal from Explorer channel 1 is shown in blue (HbR Ch1); and deoxygenated blood signal from Explorer channel 2 is shown in purple (HbR Ch2). We saw no significant difference when we pooled data across participants (four separate one-way ANOVAs; all $p>0.41$).

Based on our previous results showing large inter-individual differences in both performance and blood oxygenation on both the standard working memory task, the n-back (Kirchner, 1958), and on a more complex task, the multi-attribute task battery (MATB; (Bracken et al., 2019; Comstock & Arnegard, 1992; Santiago-Espada, Myer, Latorella, & Comstock Jr, 2011)), a multi-task battery designed by NASA. We next broke out the data to examine each individual subject. This is shown in Figure 5. These data are indicative of the high degree of variability in performance we have previously noted in working memory tasks. We have found that incorporating this variability is beneficial to modelling efforts.

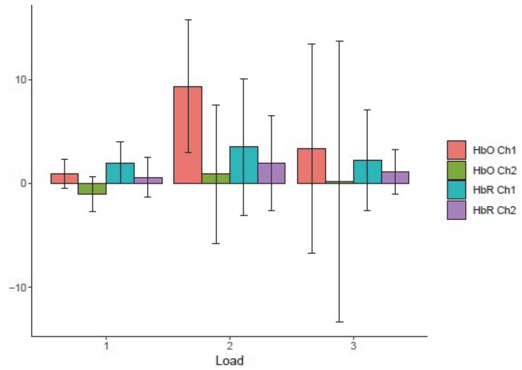

Figure 4: Blood oxygenation versus n-back load: no significant difference on pooled data.

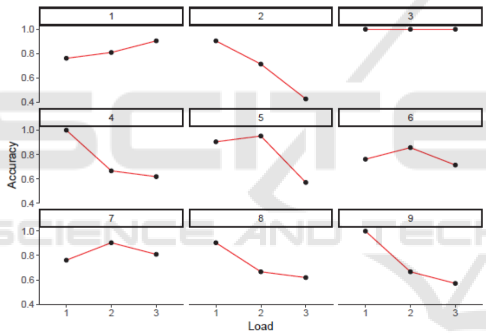

Figure 5: N-back accuracy versus load broken out by individual subject.

Figure 6 shows changes in blood oxygenation versus load broken out by individual subject. As previously noted, the individual subject variability in performance and hemodynamics means any modelling efforts must incorporate this complexity. To this end, we turned to linear mixed effect models for our efforts to predict accuracy as a function of brain oxygenation and workload by incorporating a fixed effect of slope and intercept by subject. There was a significant difference in the HbR Ch1 model for n-back load (coef=0.009, p=0.034), intercept (coef=0.96, p=1.21e-07***), and load (slope) (coef=-0.09, p=0.03). Model variants incorporating different blood oxygenation variables (HbO Ch1, HbO Ch2, HbR Ch2) also included significant intercept and load coefficients but not relationship between blood oxygenation and accuracy (i.e., cognitive workload).

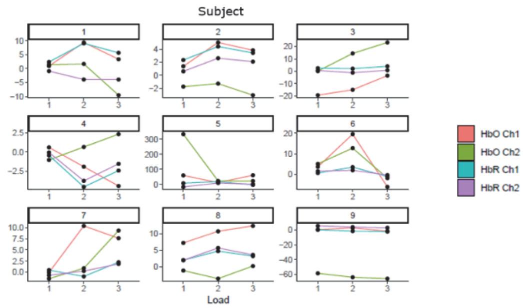

Figure 6: Changes in blood oxygenation versus load broken out by individual subject. Linear mixed model ANOVA showed that when we accounted for fixed effects of intercept and slope in our model, there was a significant difference in the HbR Ch1 model for n-back load (coef=0.009, p=0.034), intercept (coef=0.96, p=1.21e-07***), and load (slope) (coef=-0.09, p=0.03).

We next analyzed the medical curriculum training data. Each separate training focus (e.g., triage) was saved as a separate data file. However, the UMass experimenters only indicated the beginning of each training session, and did not annotate the data as to when the session ended. So we first visualized the data to decide if we should exclude some of the data collected during the session (e.g., if the last 25% of the data was a large outlier in terms of blood oxygenation, accelerometry, or any of our quality control (QC) variables that it seemed likely that it was collected after the termination of the course curriculum). In fact, this QC process was designed specifically for exploring issues of data quality. Figure 7 shows the corresponding data for HbR Ch1. This shows that data could not reliably be excluded from any particular quartile as there was not a characteristic difference in data split with this method (e.g., the first quartile is reliably different due to a difference in the experience of the participant such as donning of removing the sensor during this period).

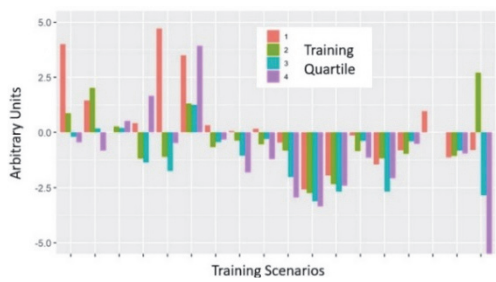

Figure 7: HbR Ch1 data collected during training scenarios binned by time within scenario (first 25% in pink, second 25% in green, third 25% in blue, last 25% in purple).

We will next visualize accelerometry data and QC variables in the same manner to determine which chunk of the data we should exclude, if any. We began by asking whether the first or last quartile of each dataset contained significantly more artefacts than the middle portions of the dataset, and whether a heuristic could be used to remove portions of the data that consisted primarily of noise (e.g., by removing the last quartile of data for all subjects). We first plotted each individual subjects' accelerometer data. We discovered that signal artefacts were primarily at the end of the data time series, rather than at the beginning of the time series, likely corresponding to the removal of the sensor from the subject's head before the sensor was shut off or recording was terminated (see Figure 8).

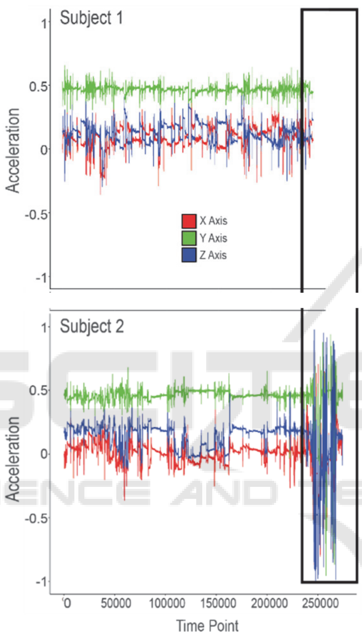

Figure 8: Accelerometry data from two subjects showing large motion artefact at the end of the session, likely corresponding to removal of the sensor. The x-axis is shown in red, y-axis is shown in green, and z-axis shown in blue.

We also discovered that we could not reliably remove a percentage of the data (e.g., the last quartile) as the length of the time series differed between subjects within the same training due to differences in how the subject progressed through the training or how quickly the research assistants were able to attend to the sensors once the training was completed. In order to perform the necessary scrubbing of non-training data from the end of the time series, we used the accelerometer data as an indicator of end-of-session sensory removal (see section indicated by black box for Subject 2 in Figure 8 for example) and manually clipped these times from each individual subject's dataset. In order to facilitate comparison of individual datasets with slightly different end-of-session motion profiles we clipped all subject's datasets within an individual training to the same time point (compare panels A and B in Figure 8). For example, both subjects in Figure 8 participated in the same training ("Triage for Disaster and Public Health Emergencies") and displayed similar levels of motion across the training. Subject 2's time series is longer than that of Subject 1, and the extra time contains a significant amount of motion. This likely corresponds to the removal of the sensor for Subject 2 without immediately turning the sensor off or stopping recording on the tablet. The accelerometer then continues to register movement which, due to the lack of annotations in the data, cannot be distinguished without manually inspecting the dataset. For the two subjects in question in Figure 8, no data was clipped from Subject 1 due to signal artefacts, while several thousand data points were clipped from Subject 2 to account for the erroneous data captured while the sensor was not placed on the subject.

This gave us confidence to trim the data using this protocol. We visualized accelerometer data from all training scenarios together and evaluated each subject's time series for motion. If motion was present, we determined an approximate point at which the aberrant motion started and then removed that data from the time series. When possible, subject time series' from the same training were clipped at the same point to facilitate a fair comparison across subjects. Figure 9 displays the raw (left) and AMARA-filtered (right) signal for a single subject before (top) and after (bottom) the clipping procedure. It is clear that the clipping procedure does not alter the characteristics of the signal in any way as it is performed after all online pre-processing is performed. In addition to verifying that the data are not altered in some way by clipping out the aberrant signal at the end of the time series, we also noted that the initial transient at the beginning of the time series, likely due to the initial online sensor calibration, is still present in the data. We might not have been aware of this if we were not evaluating each dataset individually. We will add an automated procedure to our existing processing pipeline to find and remove large transient in the first few sample of the time series. We will also pursue automated measures of detecting aberrant signal at the end of the time series based on the accelerometer data.

Analysis of the training data began with determination of difficulty levels for the different trainings. No performance measures are available for these trainings, as is often the case for real-world, ecologically valid tasks, so we relied on the subjects self-report level of difficulty experienced during the training as reported via an after-action survey administered through REDCap (https://www.project-redcap.org/software/). Figure 9 displays stacked bars of the difficulty ratings per training, with the median difficulty marked by a black point. The range of difficulty spanned only 1 ("Not challenging at all") to 3 ("Somewhat challenging"), with the median overall perceived difficulty across trainings equal to 2. Overall, this meant that subjects may not have experienced the level of difficulty seen in the 2-back version of the n-back, which is typically reported as a very challenging test of attention and working memory.

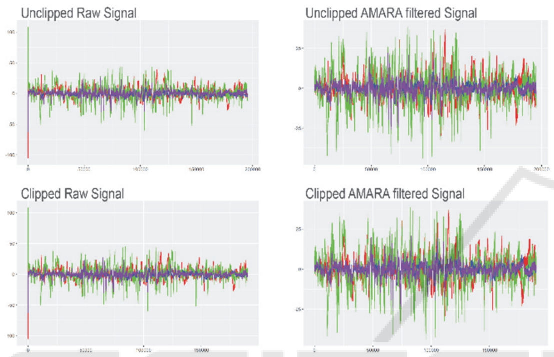

Figure 9: Raw (left) and AMARA-filtered (right) signal for a single subject before (top) and after (bottom) the clipping procedure.

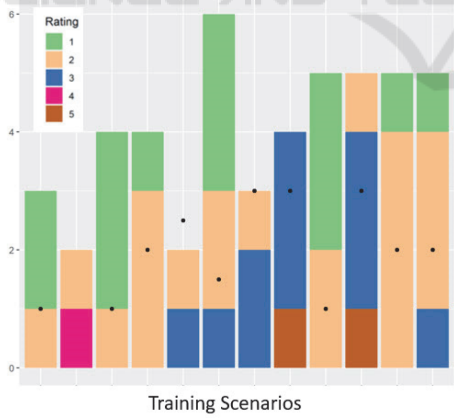

Figure 10: Self-reported difficulty ratings for each training. Median difficulty for each training is indicated by a black point.

In addition to limited variability in the self-reported difficulty rating, analysis of the training data was further complicated by the mismatch between n-back data and training data. Not all subjects who completed the medical trainings possessed n-back dataset required for adapting modelling procedures to each subject. This limited the number of subjects available for a full analysis. As was done with the n-back modelling, we attempted to model the training data according to level of experienced difficulty (i.e., performance), here indicated by the self-reported difficulty measure, using mixed effects models. We used standard mixed effects analysis with a random intercept. Unlike our n-back analysis, we could not include a random effect of slope as all subjects did not participate in trainings of all difficulties. As we have shown in the past, without the ability to model the subject specific baseline, we were unable to predict subject difficulty reported in the REDCap survey with any fNIRS-derived blood oxygenation metrics. In addition to mixed effects modelling, we also attempted to predict self-report difficulty using multinomial logistic regression and ordered logistic regression and found no relationship between fNIRS and difficulty.

4 DISCUSSION

The results of our n-back data analysis have validated that this new form factor of the sensor collects reliable data and that we are able to quantify cognitive workload. Our blood oxygenation data changed as expected with effort level on the task. However, we did not see significant changes until we had accounted for individual differences in performance on the task, which fits with our previously-published results.

Unfortunately, the problems with dropped signals, unannotated data files, and device malfunction during acquisition of the medical training simulation data meant that we did not have adequate parity between n-back data and training data on individual subjects. Therefore, we are unable to add individual performance on the n-back into the model.

Here we primarily present pooled data, whereas all of our previous results have shown that large individual differences are likely present (e.g., an increase in HbR with increase mental effort for some subjects and a decrease in HbR with increased mental effort for others) (Bracken et al., 2019).

Our future work is focusing on additional adding an individualization parameter to our model by adding information on each individual's change in blood oxygenation during the n-back task to assess changes in cognitive workload during the medical simulation training data. Our prior work shows that this level of individualization of the model is required for adequate characterization of cognitive workload. We plan to automate the individualization procedure in future data acquisition and modelling efforts.

An objective, accurate, real-time capability to inform trainers of the level of cognitive workload experienced during training would enabe trainers effectively tailor trainings to maximize impact and decrease cost associated with over-training particular skills or trainees. Additional studies must be conducted to further validate our sensors and data analysis and modelling software to prove validity of such a system.

ACKNOWLEDGEMENTS

This work was supported by United States Army Medical Research and Materiel Command under Contract Nos. W81XWH-14-C-0018 and W81XWH-17-C-0205. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Army Medical Research and Materiel Command. In the conduct of research where humans are the participants, the investigators adhered to the policies regarding the protection of human participants as prescribed by Code of Federal Regulations (CFR) Title 45, Volume 1, Part 46; Title 32, Chapter 1, Part 219; and Title 21, Chapter 1, Part 50 (Protection of Human Participants).

REFERENCES

Ayaz, H., Onaral, B., Izzetoglu, K., Shewokis, P. A., McKendrick, R., & Parasuraman, R. (2013). Continuous monitoring of brain dynamics with functional near infrared spectroscopy as a tool for neuroergonomic research: Empirical examples and a technological development. *Frontiers in Human Neuroscience*, *7*, 871. https://doi.org/10.3389/fnhum.2013.00871

Ayaz, H., Shewokis, P. A., Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2012). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage*, *59*(1), 36–47. https://doi.org/10.1016/j.neuroimage.2011.06.023

Boas, D. A., Elwell, C. E., Ferrari, M., & Taga, G. (2014). Twenty years of functional near-infrared spectroscopy: Introduction for the special issue. *NeuroImage*, *85*, 1–5. https://doi.org/10.1016/j.neuroimage.2013.11.033

Bracken, B., Festa, E., Sun, H.-M., Leather, C., & Strangman, G. (2019). Validation of the fNIRS Pioneer™, a Portable, Durable, Rugged Functional Near-Infrared Spectroscopy (fNIRS) Device. In *Biomedical Engineering Systems and Technologies: 12th International Joint Conference, BIOSTEC 2019, Prague, Czech Republic, February 22–24, 2019*. Springer.

Bracken, B. K., Elkin-Frankston, S., Palmon, N., Farry, M., & de B Frederick, B. (2017). *A System to Monitor Cognitive Workload in Naturalistic High-Motion Environments*.

Bracken, B. K., Palmon, N., Elkin-Frankston, S., & Silva, F. (2018). *Portable, Durable, Rugged, Functional Near-Infrared Spectroscpy (fNIRS) Sensor*.

Bunce, S. C., Izzetoglu, K., Ayaz, H., Shewokis, P., Izzetoglu, M., Pourrezaei, K., & Onaral, B. (2011). Implementation of fNIRS for Monitoring Levels of Expertise and Mental Workload. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems* (Vol. 6780, pp. 13–22). https://doi.org/10.1007/978-3-642-21852-1_2

Comstock, J. R., & Arnegard, R. J. (1992). MAT: Multi-Attribute Task Battery for Human Operator Workload and Strategic Behavior Research. *NASA Technical Memorandum*, (January).

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*(2), 921–935. https://doi.org/10.1016/j.neuroimage.2012.03.049

Kerick, S. E., Oie, K. S., & McDowell, K. (2009). *Assessment of EEG signal quality in motion environments*. Army Research Lab Aberdeen Proving Ground MD Human Research and Engineering Directorate.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352.

Mathan, S., Whitlow, S., Dorneich, M., Ververs, P., & Davis, G. (2007). Neurophysiological estimation of interruptibility: Demonstrating feasibility in a field context. *In Proceedings of the 4th International Conference of the Augmented Cognition Society*, 51–58.

McKendrick, R., Parasuraman, R., Murtza, R., Formwalt, A., Baccus, W., Paczynski, M., & Ayaz, H. (2016). Into the wild: Neuroergonomic differentiation of hand-held and augmented reality wearable displays during outdoor navigation with functional near infrared spectroscopy. *Frontiers in Human Neuroscience*, *10*, 216.

Metz, A. J., Wolf, M., Achermann, P., & Scholkmann, F. (2015). A new approach for automatic removal of movement artifacts in near-infrared spectroscopy time series by means of acceleration data. *Algorithms*, *8*(4), 1052–1075.

Molavi, B., & Dumont, G. A. (2012). Wavelet-based motion artifact removal for functional near-infrared spectroscopy. *Physiological Measurement*, *33*(2), 259.

Santiago-Espada, Y., Myer, R. R., Latorella, K. A., & Comstock Jr, J. R. (2011). *The multi-attribute task battery ii (matb-ii) software for human performance and workload research: A user's guide*.

Scholkmann, F., Spichtig, S., Muehlemann, T., & Wolf, M. (2010). How to detect and reduce movement artifacts in near-infrared imaging using moving standard deviation and spline interpolation. *Physiological Measurement*, *31*(5), 649–662. https://doi.org/10.1088/0967-3334/31/5/004