Using fNIRS for Realtime Cognitive Workload Assessment

Samuel W. Hincks[1], Daniel Afergan[2], Robert J.K. Jacob[1]

[1] Tufts University, MA 02145, USA
{shincks,jacob}@cs.tufts.edu
[2] Google Inc. Mountain View, CA 94043
{afergan}@google

Abstract. In this paper, we evaluate the possibility of detecting continuous changes in the user's cognitive workload using functional near-infrared spectroscopy (fNIRS). We dissect the source of meaning in a large collection of n-backs and argue that the problem of controlling the content of a participant's mind poses a major problem for calibrating an algorithm using black box machine learning. We therefore suggest that the field simplify its task, and begin to focus on building algorithms that work on specialized subjects, before adapting these to a wider audience.

Keywords: fNIRS · physiological interface ·implicit interface ·machine learning ·default mode network ·cognitive workload

1 Introduction

User interfaces typically deduce their user's intentions by measuring their physical gestures, within a tiny and agreed-upon space of device-dependent commands. Since the user produces each such input consciously, the mapping between physical gesture and digital effect should be transparent and immediate. Ongoing HCI research attempts to improve this bandwidth: to provide more information to the computer without posing an additional cognitive tax on the user. For example, whereas a conventional user interface (UI) is designed for a single prototypical user, a better UI recognizes differences among users, and can adapt its design appropriately.

But critical dimensions of the user vary from moment to moment. Humans, who like computers can be described with the metaphor of a continuously changing information processor, support multiple editions of themselves. Optimally, a user interface would characterize the user at each point in time. The goal of an *Implicit Interface* is to deduce the user's mood, intention, preference, and more general cognitive state from non-intentionally transmitted information. It then capitalizes on this information for the purpose of improved user experience in real time.

The design of an implicit or physiological interface can be divided into three parts:

- The *Measurement Component* dissects the user's cognitive processes into useful abstractions that can be inferred from behavioral or physiological data
- The *Data Mining Component* translates unorganized user data into accurate state predictions
- *The Design Component* crafts interfaces that adapt based on these predictions

Work in implicit interfaces generally covers all three parts. This paper is motivated by the problem of building and calibrating a brain-computer interface based on functional near-infrared spectroscopy (fNIRS). We combine knowledge from cognitive and neuroscience with observations from single trial analysis to arrive at a set of recommendations. We first analyze a concrete implementation of an implicit interface using fNIRS, focusing on the possible improvements to the data mining components. The paper makes the following contributions:

- First, we reanalyze the data collected from a previous experiment [1] using machine learning.
- Second, to gauge the portability of this algorithm in a real-time cognitive workload prediction context, we examine the character of individual fNIRS trials.
- Third, we suggest a possible alternative method for inducing mental states involving controlled self-report and adaptively filtered fNIRS data
- Finally, we consider how thinking about psychological states in terms of task-positive and resting-networks may provide a more useful framework for interpreting fNIRS data.

2 Dynamic Difficulty Adjustment

2.1 Design Component

In [1], we identified a design problem. With j Unmanned Aerial Vehicles (UAVs) and k human monitors of these UAVs, what is the optimal distribution of the j UAVs among the k human monitors? The naive solution is to distribute workload equitably among the monitors (each one receiving j/k UAVs). However, each monitor has different cognitive endowment. In addition, their cognitive energy and capacity might fluctuate unpredictably over the course of an hour or a day. Equal distribution thus appears to be sub-optimal. The UAV-monitor designer therefore poses a question to psychology: what cognitive abstractions can be characterized to elicit optimal user performance?

2.2 Cognitive Workload.

We chose the cognitive workload or working memory state as cognitive abstraction. Working memory has fixed capacity and supports separate and somewhat parallelized buffers for different formats of data (spatial, verbal, episodic) [4]. It can be induced by the n-back task. In a visuospatial 1-back, the subject identifies whether or not a square on a grid matches what was seen in a previous iteration; in a visuospatial 2-back, the subject responds whether or not it matches the arrangement two iterations

ago. Engagement of working memory activates the dorsolateral prefrontal cortex (dlPFC) [17]. It is a prime candidate for detection via EEG [9] as well as through peripheral physiological correlates such as changes pupillary response (Iqbal, 2004) and heart rate variability [14]. A useful physiological sensor for HCI has a likely entry-point into consumer grade electronics as well as a very specific physiological trace that is unlikely to be tricked by a user in motion in an unpredictable real world setting. For this reason, we focus on functional Near Infrared Spectroscopy (fNIRS).

2.3 Functional Near Infrared Spectroscopy

Given its relatively exterior neurobiological housing in the dlPFC, cognitive workload invites convenient detection using non-invasive light-based neuroimaging. fNIRS is a neuroimaging technique especially well-suited as a supplementary input device in a Brain-Computer Interface [11,12,1,2,15,16]. A light source beams near infrared light that penetrates skin and bone; it is differentially absorbed by oxygenated and deoxygenated hemoglobin in the neural bloodstream, so that a measurement of the photons returning to a nearby sensor can indicate the relative proportion of either quantity. Since inter-neuronal communication (the basic computational process of any mental activity) demands a continuous stream of oxygen, measurements of oxygen provides a rough barometer of the activity of the probed region. The depth of effective probing is limited (the distance between light source and detector approximates the maximum depth), but, especially compared to EEG, fNIRS provides spatially well-resolved information, meaning the signal fluctuates mostly in response to changes in biology of the underlying region. It requires little calibration; light sources and detectors need only to be placed flush on bare skin adjacent to the cranium; and the basic components of fNIRS can likely be miniaturized and integrated cheaply and seamlessly into consumer grade electronics such as head-mounted wearable computers in the future [8]. In fact, several labs are in the process of building and refining such a low-cost fNIRS [6].

2.4 Data Mining Component.

Seeking to adaptively control task difficulty in a UAV-operation setting, we thus sought to continuously portray a user's cognitive workload using fNIRS. We used machine learning to translate successions of fNIRS data into discrete classifications of the user's state. We calibrated the machine learning algorithm on easy and hard versions of the n-back task. Specifically, subjects completed the easy 1-back task for 25 seconds, rested for 15 seconds, before completing the harder 3-back task. When complete, basic statistical features (mean and slope) were computed from the time series. These features, along with the associated class (1-back vs 3-back), were then fed into a support vector machine (SVM). Trained, the SVM could then theoretically estimate the probability that unseen 30-second time-series of fNIRS data pertained to a user experiencing either high workload or low cognitive workload.

2.5 Results

In the adaptive condition, the number of UAVs under the user's jurisdiction changed based on fNIRS-detected workload, and this significantly reduced the amount of user error compared to a non-adaptive condition with equivalent overall work.

3 Limitations with Machine Learning

Machine learning, which was also used previously [1,2,11,12], has several advantages for detecting cognitive state. It leverages optimized function fitting methods, allowing it to find patterns which might elude a human observer utilizing classical statistical techniques. It has a built-in system for confirming that a state can be reliably induced and measured in a subject by cross-fold validation. It generalizes nicely for calibrating new algorithms: one need only replace the calibration task and corresponding labels. It assumes, rightly, that subjects have different brains and that probe placements differ slightly from one experiment to another. Finally, it doesn't require the designer of the algorithm to possess advanced neuroscientific expertise as the burden of discovering neural correlates of state is left to automated pattern discovery.

But it also has several disadvantages. The state of cognitive workload induced by the benchmark task may not match the state of cognitive workload in the ordinary course of the experiment. It might be that that the n-back induces only a very specific cognitive workload state with a neural profile that does not reflect the more general state which the system endeavors to measure. Similarly, the n-back may not enlist processing in the dorsolateral regions under measurement for a particular subject, in which case the more general state of workload may be detectable even though calibration task failed to enlist it. Most problematically, the real time user experience is not organized into neat serial trials with preceding baselines periods and clear starting points. A machine learning algorithm is calibrated on 30-second time segments only to be required to make predictions of a real task in continuous time. In the best case, where a user would transition from a baseline resting state into a thirty second experience of high cognitive workload, the time series would only align with the trials of the calibration period at that one point: 30 seconds after rest. In theory, every other prediction is doomed because of fundamental misalignment with the structure assumed during calibration.

The lack of reasonable ground truth for a state outside of the format of a well-controlled induction task makes it impossible to evaluate how well a machine learning based algorithm is actually performing. All that is known is that, in some percentage of experiments, some percentage of dependent variables are significantly superior in an adaptive condition that hinges on accurate real-time prediction.

4 Machine Learning Reanalysis

The approach of only examining average data and delegating large portions of the problem to black box machine learning may work, but it stymies progress towards a better algorithm, and needs to be supplemented with a calculated break down into exquisitely well-controlled individual trials. In this paper, we try to dissect the fNIRS workload signal. We begin by looking at the average case, before breaking the problem down to individual subjects, and ultimately individual trials.

In an exploratory re-analysis of the 27 subjects in [1], leave-one-trial-out analysis of each subject resulted in an average case performance of 66.6%. In other words, when a machine learning algorithm was fed all but one of the fifteen 1-back trials and fifteen 3-back trials for any given subject, it correctly identified the excluded testing trial as belonging to either 1-back or 3-back with 66.6% accuracy. Although several features, filters, and machine learning algorithms were tested, optimal performance involved Weka's SMO support vector machine [18] processing the mean, standard deviation, and slope-of-best-fit on the whole and second half of every channel. This combination has proven most useful in this and other experiments [15]. Curiously, optimal performance omitted data preprocessing techniques and filtering, possibly because respiratory signals such as heart rate variability, captured by the standard deviation feature, provide information to the calculation of workload. The fact that offline analysis advises against preprocessing and filtering further substantiates the problem of building an online algorithm from evaluation in an offline setting.

Classification accuracies ranged from 34.4% to 89.2%, with a standard deviation of 13.9%. Five subjects had classification accuracies above 80% (subjects 27, 25, 10, 8, 12). These subjects warrant further investigation.

In order to dissect the source of meaning in the data and get a better sense of the difficulty of the problem, we have selected the five best subjects (as measured through machine learning), and applied the following steps of preprocessing to their data. First, we have applied the Modified Beer Lambert Law to extract Hb and HbO. For simplification, we then merged values from neighboring sources on the forehead into single channels, before z-scoring the data, and applying a moving average algorithm, which sets each value to be the average of 16 readings (the device samples at 11.9 hz) as well as a low pass filter at a cutoff of 0.5hz. Finally, we have anchored each trial at 0, setting each point in the 25 second trial to reflect the difference between it and the starting point

In figure 1, we have added all trials into a common dataset and then merged these into one graph, where the dotted line reflects the average value and the thickness of the area chart reflects one standard deviation. Selecting among four possible graphs: either left or right prefrontal cortex and either Hb or HbO, we selected the most visually distinguishable graph. This graph agrees with the average case reported in the literature; that higher workload tends to signify an increase in oxygenation, especially in the left PFC [10]. It also shows how the change in oxygenation happens slowly, beginning some time after the trial begins.

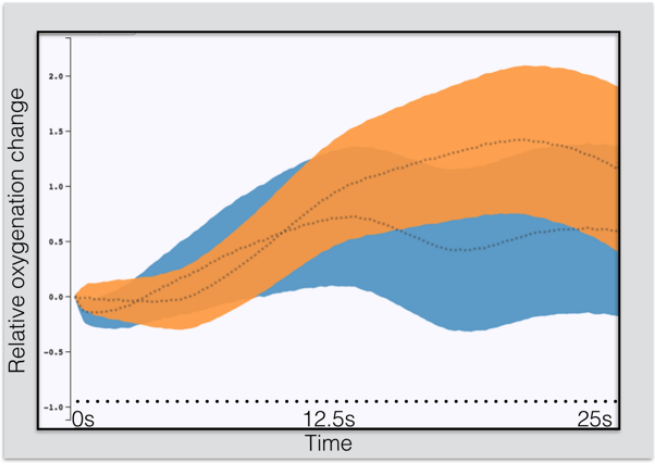

Fig. 1. Average left prefrontal cortex HbO activation for five subjects

In figure 2, we have applied the same selection procedure but for each individual subject. We show each individual trial to get a sense of the underlying variation. We note that for four out of five subjects the most visually distinguishable was on the left prefrontal cortex.

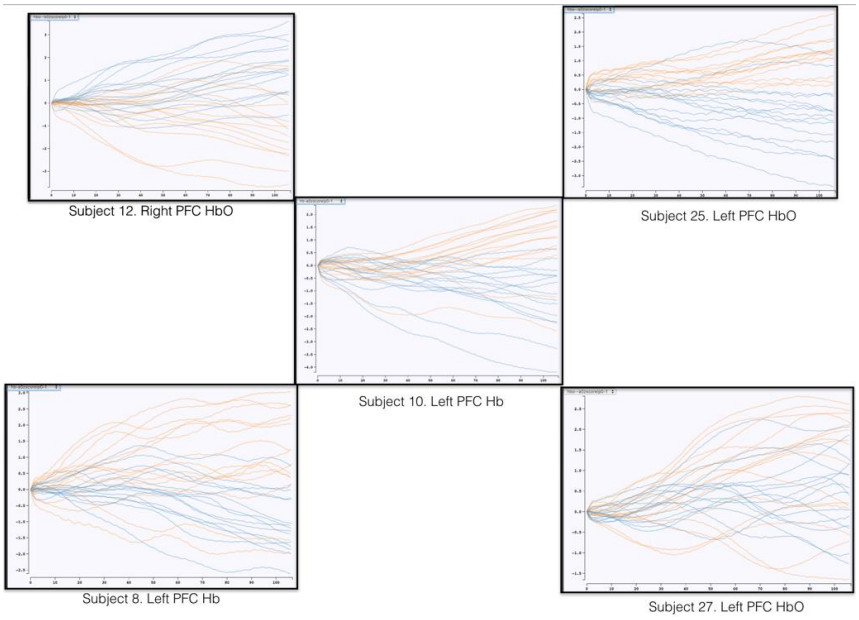

Fig. 2. Changes in Hb for five subjects

Each subject exhibits a visually apparent pattern with individual exceptions to that pattern. The key to designing an accurate algorithm is to figure out why certain trials break the trend. There are two plausible reasons:

- A spontaneous and unrelated respiratory trend might be overwhelming a consistent neurological response.
- The user might have lost focus or otherwise solved the problem under a different cognitive profile, thereby generating a different neurological response.

5 Towards a Better Algorithm

5.1 Adaptive Filtering

To mitigate the effect of spontaneous respiratory signals interfering with the true neurological response, it is possible to apply adaptive filtering (Zhang, 2009; Aqil, 2012). An adaptive filter requires there to be one source-detector pair which measures oxygenation changes in only the skin; this occurs if the source is less than a centimeter away from the detector. With one signal measuring both brain and skin response and one signal only measuring skin response, a channel including nothing but brain response can be obtained by subtracting the frequencies the two have in common. The adaptive filter also fulfills the purposes of the bandpass filters of eliminating the heart rate and respiratory components. It may also eliminate spontaneous low frequency oscillations that can drown out an otherwise consistent neurological response.

5.2 Controlling for Spontaneous Mental Activity

With the software tools and hardware in place to extract the exclusively neurological component of the signal, the next challenge is to find a way to assess whether or not the user is maintaining a consistent cognitive profile across trials. This requirement demands a degree of meta-awareness that would be rare in the average participant. We therefore combined our research that aims to build an algorithm that works moderately on many people with a parallel investigation of an algorithm that works extremely well on one person. It would be informative to study just one mentally disciplined subject who generated reasonably consistent fNIRS-detected neurological responses and could assist explain the character of their own mental activity. It is possible that none of the results effectively generalize to others. But the more likely scenario is that this format invites rapid iteration for testing of new algorithms, and can quickly suggest what is and what is not possible.

To prototype the possibility of this approach, one of the authors of this paper (SH) regularly examined his own fNIRS activity. The following section is therefore written in the first person to show the advantages of having the same person as interpreter and supplier of the data.

5.3 4-Back

In this experiment, I completed a total of fourteen 30 second aural 4-backs (in orange) and aural 0-backs (in blue), interspersed with 10 second resting periods. (In an aural n-back, participants repeat the number heard n iterations ago). The data (which have been adaptively filtered and anchored to start at zero) are shown in figure 3. To solve a 4-back, I subvocally rehearse a 4-item mental buffer. When I hear a new number, I say the first number of the string recently rehearsed, then immediately repeat the string but with the last string's second element in the first position, and the new element in the last position. As long as I keep escalating n, it is virtually impossible to solve the problem without granting the task the exclusive province of my attention. During a 0-back, I do not exert the same level of control over the activity of my mind. I complete the task, but occasionally my conscious mind is occupied by other thoughts or feelings.

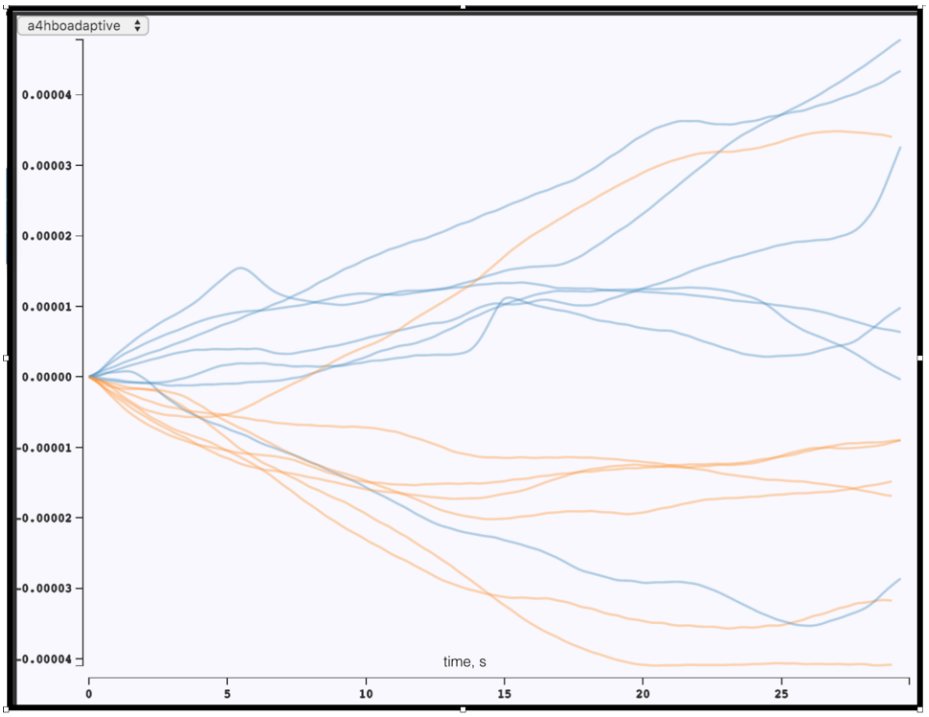

Fig. 3. Adaptively filtered change in HbO

What is especially interesting with fNIRS self-analysis is the possibility to dissect the trials which break an otherwise coherent pattern. The perfect self-analytical experiment is when you have a clear pattern, with exactly one exception. In this experiment, each condition has one exception. For the easy trials (when I was merely repeating numbers), one trial has the character of the 4-back trials. And for the 4-back trials,

one trial has the character of the 0-back trials. The 4-back exception is especially interesting because it was the last trial in the series and I know exactly how it mentally unfolded. It was the last trial and I was getting tired. For the first five seconds, the trial unfolded like every other 4-back trial. I heard 0, then subvocally said zero. I heard 3; then subvocally said zero-three. I heard 5; then subvocally said zero-three-five. But then I heard six, and instead of adding it to a four-item buffer, I said aloud the first item in my buffer (zero), immediately noticed the mistake and then fumbled through the remainder trial unable to recover. Coincidentally, five seconds into the trial (approximately when I heard 5), there is a spike in the oxygenation levels in my left ventromedial prefrontal cortex.

There are three possible explanations for this. First, it could be entirely coincidental. I think this would be the wrong interpretation. The data is otherwise consistent, and the shift in oxygenation is a slope-value greater than anything observed in the dataset up until that point. Second, it is possible that I made the mistake, noticed the mistake, and the data reflects my frustration. But a subtle clue in the data dismisses this possible. The break in activation occurs two seconds *before* my noticing that I have failed the trial. That suggests the third possibility is true.

Five seconds into the trial, either a rhythmic biological force or a string of computational association, cemented computation in one of the nodes antithetical to proper focus. For 6/7 of the 4-back trials, a combination of task difficulty and mental preparedness enabled me to block out the otherwise near-continuous presence of my default mode network. For the 7th and last trial, I got ready, doing my best to sustain focus, but five seconds in, neurological circumstance concentrated computation in a network of mind that interferes with task-related rumination. This explanation aligns exactly with the observed data, my private experience of it, as well as the cognitive science literature.

6 Task Positive vs Default Mode Network

Neuroscience has changed since the advent of BCI [13].Early BCI applications focused on the distinction high vs low cognitive workload in part because it seemed the most neurologically rich and inducible state. Since then, neuroscience has grown increasingly interested in studying the brain at rest, discovering an active and energy-consuming set of regions known as the default mode network (DMN), which together compute the tendency of the mind to wander and ruminate when given the opportunity, becoming relatively more silent when the host engages a consuming task. It is an open question whether fNIRS PFC oxygenation values correlate better with (a) the current strain of working memory or (b) the enlistment of general mental resources towards task performance. Note that it is possible to have an engaged working memory without utilizing this memory in the service of a task and also possible to be dedicating attention wholeheartedly to a task without utilizing working memory. But either cognitive abstraction explains the differentiability of an n-back signal from fNIRS data. Furthermore, the primary sensitivity of the fNIRS to (b) not (a) better fits the between subject variability of the signal.

With this more general portrayal of cognitive state, the failure to coerce a consistent cognitive state can easily be explained. In some cases, a 3-back requires the full enlistment of task-positive attention and the subject is both cognitively able and motivated to grant it. In other cases, because the task is too easy, too hard, or the subject is generally distracted, the subject is unable to enter a task-positive network. Similarly for the low workload induction: sometimes these trials are met with relative tranquility and other times a busily exploring default mode network.

Between participants, fMRI studies have shown that extroverts have a greater change in the fMRI signal in the dlPFC between three-backs and rest compared to introverts even though, behaviorally, they perform the same [7]. One explanation for this discrepancy is that introverts, being consistent self-monitors and relatively less capable to surrender attention to the external environment, fail to dislodge mild concurrent self-monitoring scripts as they complete n-backs. So while the cognitive processes in charge of their working memory remain their same, perhaps they fail to fully enter the task-positive network like the extroverts.

Task-related attention and cognitive workload overlap, but knowing them invites a different set of design considerations. The basic goal of any user interface is to facilitate task performance. Good user interfaces therefore mitigate distractibility and narrow their user's attention. Knowing the direction of the user's attention, internal or external, provides the basic measure governing the current effectiveness of a user interface: a broader and more useful piece of information than knowing the weight of their cognitive workload.

7 Conclusion

We suggest a return to simplicity in the design fNIRS-based cognitive state predictors. First, we recommend steering away from machine learning until the meaning of the data can be understood on a single trial basis. Second, we recommend adaptive filtering so that channels with exclusive brain activity can be examined. Third, we recommend perfecting an algorithm on only one or a handful of trained and disciplined subjects, before deciding how to train, calibrate, and deploy it on a random audience. Finally, we encourage future BCI studies, like contemporary neuroscience, begin to consider the relevance and usefulness of calibrating algorithms mind-wandering resting states.

References

1. Afergan, Daniel, et al. "Dynamic difficulty using brain metrics of workload."*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2014
2. Afergan, Daniel, et al. "Brain-based target expansion." *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 2014
3. Aqil, Muhammad, et al. "Cortical brain imaging by adaptive filtering of NIRS signals." *Neuroscience letters* 514.1 (2012): 35-41
4. Baddeley, Alan D., and Graham Hitch. "Working memory." *The psychology of learning and motivation* 8 (1974): 47-89
5. Iqbal, Shamsi T., Xianjun Sam Zheng, and Brian P. Bailey. "Task-evoked pupillary response to mental workload in human-computer interaction."*CHI'04 extended abstracts on Human factors in computing systems*. ACM, 2004
6. Juanning Si, Ruirui Zhao, Yujin Zhang, Nianming Zuo, Xin Zhang, and Tianzi Jiang. 2015. A portable fNIRS system with eight channels. In *SPIE BiOS*. International Society for Optics and Photonics, 93051B–93051B
7. Kumari, Veena, Steven CR Williams, and Jeffrey A. Gray. "Personality predicts brain responses to cognitive demands." *The Journal of neuroscience*24.47 (2004): 10636-10641
8. Ferrari, Marco, and Valentina Quaresima. "A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application." *Neuroimage* 63.2 (2012): 921-935
9. Gevins, Alan, and Michael E. Smith. "Neurophysiological measures of cognitive workload during human-computer interaction." *Theoretical Issues in Ergonomics Science* 4.1-2 (2003): 113-131
10. Herff, Christian, et al. "Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS." *Frontiers in human neuroscience* 7.1 (2013): 935-940
11. Peck, Evan M., Daniel Afergan, and Robert JK Jacob. "Investigation of fNIRS brain sensing as input to information filtering systems." *Proceedings of the 4th Augmented Human International Conference*. ACM, 2013
12. Peck, Evan M., et al. "Using fNIRS to measure mental workload in the real world." *Advances in physiological computing*. Springer London, 2014. 117-139
13. Randy L Buckner, Jessica R Andrews-Hanna, and Daniel L Schacter. 2008. The brain's default network. *Annals of the New York Academy of Sciences* 1124, 1 (2008), 1–38
14. Rowe, Dennis W., John Sibert, and Don Irwin. "Heart rate variability: Indicator of user state as an aid to human-computer interaction."*Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., 1998
15. Treacy Solovey, Erin, et al. "Designing implicit interfaces for physiological computing: Guidelines and lessons learned using fNIRS." *ACM Transactions on Computer-Human Interaction (TOCHI)* 21.6 (2015): 35
16. Solovey, Erin Treacy, et al. "Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines." *Proceedings of the 22nd annual ACM symposium on User interface software and technology*. ACM, 2009
17. Wager, Tor D., and Edward E. Smith. "Neuroimaging studies of working memory." *Cognitive, Affective, & Behavioral Neuroscience* 3.4 (2003): 255-274

18. Witten, Ian H., et al. "Weka: Practical machine learning tools and techniques with Java implementations." (1999): 81-17
19. Zhang, Quan, Gary E. Strangman, and Giorgio Ganis. "Adaptive filtering to reduce global interference in non-invasive NIRS measures of brain activation: how well and when does it work?." *Neuroimage* 45.3 (2009): 788-794