Measuring workload using a combination of electroencephalography and near infrared spectroscopy

Emily B.J. Coffey[a,b], Anne-Marie Brouwer[b], Jan B.F. van Erp[b]

[a] *Cognitive Science Center of the Universiteit van Amsterdam, Sarphastistraat 104, Amsterdam, The Netherlands 1018GV*
[b] *TNO Human Factors, Kampweg 5, Soesterberg, The Netherlands 3769ZG*

Abstract

The ability to continuously monitor workload in a real-world environment would have important implications for the offline design of human machine interfaces as well as the real-time online improvement of interaction between humans and machines. The present study explored the usefulness of combining electroencephalography (EEG) with the newer technique of near infrared spectroscopy (NIRS), under data acquisition and processing conditions that could be applied to real-time usage, for example as an input to adaptive automation. Eight EEG channels (Cz, Pz, FCz, Fz, C3, C4, F3, and F4) and three NIRS channels over the left forehead were acquired simultaneously, during repetitions of blocks of three difficulty conditions of the N-back task. The resulting data were separated into five-second windows and binary classifications on condition were performed on bandpower-derived features for EEG, and average hemoglobin levels for NIRS. Each type of data was classified independently, and in combination. In general, EEG could be used to reliably classify workload condition for most subjects, whereas the NIRS signal was less helpful and did not contribute to classification accuracies when combined with EEG. Implications and future directions are discussed.

Key words: Workload, Working Memory, Brain-Machine Interface, Brain-Computer Interface, EEG, NIRS

1. Introduction

The limited capacity of the human information processing system may be expanded by redesigning the information flow between humans and computers. An interdisciplinary research approach known as ‘neuroergonomics’ seeks to integrate our understanding of the neural basis of cognition with the design and development of technology [1]. Subsets of this field known as ‘adaptive automation’ [2] and ‘augmented cognition’ involve using real-time measurements of user states to permit machine systems to detect and react appropriately to reduce errors and increase performance. Research to date suggests that performance in various tasks such as driving [3], monitoring technical and security systems [4], and even learning [5] can be enhanced, for example by alerting users to lapses in attention or by modifying the user’s tasks in real-time to prevent overload or underload. In the past, researchers have mainly used measures of behavior, performance data, or physiological parameters such as heart rate variability, occulomotor activity, pupilometry, and galvanic skin response as inputs to these systems [1]. While these measures have been somewhat successful, they are indirect indicators of cognitive processes and show limited predictive power. This has lead to a demand for more direct measures of cognitive phenomena such as task engagement, cognitive workload, surprise, satisfaction, or frustration [6, 7].

Research using brain-sensing technologies to determine user states for human-machine interaction applications is currently in its initial phases. One research focus in which progress has been made to distinguish differences in brain states in real-time is that of brain-machine interfaces (BMIs; also known as brain-computer interfaces). BMIs are devices which obtain information for their operation through the measurement of correlates of neural activity associated with mental processes [8]. They are usually designed for direct, intentional control of computer cursors, games, or typing programs for clinical populations with extremely limited motor control. Despite equipment and machine learning algorithms of increasing complexity, modern non-invasive BMIs have extremely low information transfer rates (meaning only a few commands can be recognized per minute) and have levels of accuracy that are highly variable between subjects [9], while they require most of the user’s attention. For healthy users, this is not a practical alternative to conventional methods of human-computer interaction. Instead of replacing traditional modes, BMI tools and technology can help to fill the direct-measure gap in neuroergonomics applications [7]. As use of these BMIs does not demand effort or attention, they may be practically employed to improve safety and efficiency in many industrial, educational, and everyday environments - both offline in the system design process and online during operations.

Before these predictions can be realized, a number of technical challenges must be overcome. The main obstacles to advancement lie in identifying stable, robust signals of the cognitive state, in ecologically valid settings, and using relatively inexpensive, portable equipment that does not interfere with the task [1]. If the device is to be used for online adaptive automation, these signals must also be obtained rapidly, therefore extracting information from short windows of data and without recourse to data processing techniques that require knowledge of the entire data set.

EEG has been well-studied in the context of brain computer interfaces, including several studies on workload, a state of primary interest for safety and performance in the workplace (e.g. [1, 6]). EEG uses scalp electrodes to record weak electrical signals generated by the post-synaptic potentials of large groups of cortical neurons firing simultaneously. The signal is attenuated and distorted by the skull and other tissue, and is easily contaminated by artifacts from muscle activity and ambient electrical noise. Temporal resolution is high (within the millisecond range), but deriving the location of the active brain region, particularly with few electrodes, is difficult or impossible. Nonetheless, even inexpensive EEG systems with few channels in an electrically un-insulated room can be used to distinguish between brain states induced by different tasks [6, 7].

Near infrared spectroscopy (NIRS) is a non-invasive optical technique which infers relative changes in the concentration of oxygenated and deoxygenated hemoglobin in the cerebral cortex from scattering and absorption properties of light projected through the skull (see [10] for a description of the working priciples of NIRS). The resulting signal is similar to that obtained in blood oxygen level-dependent functional magnetic resonance imaging (fMRI), though only cortical regions can be accessed [10]. In contrast to fMRI, NIRS is relatively inexpensive, portable, and allows for measurements during activities such as computer use. NIRS has been successfully used for several preliminary BMI studies, such as by Ogata and colleagues [11] to classify signals from the prefrontal cortex related to different cognitive tasks (see [8] for a description of current NIRS-based BMIs). NIRS can have reasonably good spatial resolution, particularly if high numbers of optodes are used. Temporal resolution is limited by the delayed nature of the hemodynamic response. NIRS is insensitive to electrical environmental noise, but may be disturbed by excessive user movement, and contains artifacts from breathing, heart rate, and a slow baroreceptor-related oscillation known as the Mayer wave (∼0.1Hz). The amount of light transmitted and absorbed is affected by the degree of pigmentation of hair and skin, and the thickness of the skull. Comparatively little research has been conducted using NIRS and its full potential for real-time brain state signal use is not yet known.

Because EEG and NIRS measure different physiological correlates of neural activity and are susceptible to different noise sources, it has been suggested that using a combination of both techniques could improve our ability to differentiate between brain states [1, 8]. A handful of studies have been conducted using EEG and NIRS simultaneously for neuroimaging studies, including event-related designs (see [12, 13, 14]). To the best of our knowledge, a combination of EEG and NIRS for use in real time workload classification has not been attempted.

The focus of this research is to test the hypothesis that the combination of electrophysiological and hemodynamic information from the brain, obtained under conditions conducive to future real-time use, might help to classify a user's workload level. Prior research has identified several neurophysiological correlates of workload level change. For example, in EEG results, an increase in frontal theta and a decrease in alpha frequency band activity has been reported [6]. Hemodynamic studies generally show increases in oxygenated blood in the dorsolateral prefrontal cortex with increasing cognitive workload [10, 15].

2. Materials and Methods

2.1. Participants

Twelve participants were recruited from a pool of volunteers, consisting mostly of students and young professionals from the local area. The data from two subjects was not included due to a technical problem which precluded complete data analysis; all results are reported for the remaining ten subjects. All participants were fair-skinned (favorable for near infrared light transmission). All participants were unfamiliar with the N-back task. The mean age was 23.5 (SD = 3.0). Five participants were male, and three were left-handed. All participants used the first two fingers of their right hand to respond to the task, except for one who experienced discomfort in the right hand and switched to the left hand after several experimental cycles. The participants signed informed consent forms and were compensated for their time.

2.2. Experimental Design and Task

The experiment was approved by the TNO internal ethics committee. It was performed in an office environment, with the lights off to minimize extraneous light collected by the NIRS system, resulting in dim lighting from partially blocked daylight. The participant sat on an office chair at a computer desk in front of an LCD computer screen and used a chin rest (see Figure 1).

As outlined by Berka *et al.* [1], it is necessary to define a relatively pure task that consistently elicits the brain state of interest in order to identify robust signal correlates which are required to train a classifier. This classifier can later be validated in more realistic task circumstances. In human computer interaction research, high working memory load is recognized as predictive of errors and of slowed procedural skill acquisition, and it is considered a major component and reasonable approximation of workload [1, 6]. We therefore selected the 'N-back task' as our workload manipulation, an experimental paradigm which has been studied extensively in functional

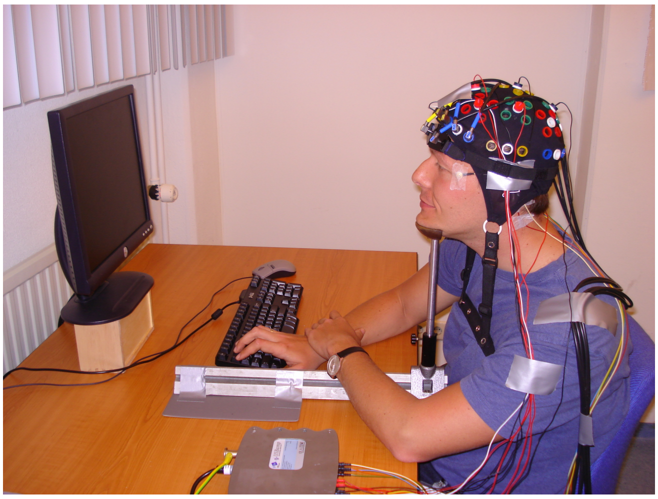


Figure 1: Experimental set-up (participant photo used with permission)

neuroimaging studies of working memory and cognitive load [15, 16].

The N-back task was presented using E-Prime (Psychology Software Tools, Pittsburgh). White letters (font style: Arial bold, approximately 3cm high) were presented for 500ms on a black background, followed by a 2000ms inter-stimulus interval during which a small white fixation cross was presented. Responses were collected using the '1' and '2' buttons on the computer's keyboard. Accuracy and reaction time data were recorded by E-prime. Performance data from the first three letters of each condition were not collected, due the pre-loading requirement of the 2-back condition. In the 0-back condition, the target letter was 'X'; in the 1-back condition the letter was a target if it was identical to the letter presented immediately before; and in the 2-back condition the letter was a target if it was identical to the letter presented two letters previously. A 3-back condition was not used, due to evidence that many subjects find it too difficult and tend to give up [10, 17]. In all difficulty conditions, 33% of letters were targets. Except for 'X' in the 0-back task, letters were randomly selected from English consonants. Vowels were excluded to reduce the likeliness of participants developing chunking strategies which reduce mental effort, as suggested in [6].

Prior to beginning the experiment, the participants performed one practice session on each of the 0 and 1-back tasks, and two practice sessions on the more difficult 2-back task such that the average performance accuracy in each condition was over 80%. Participants were asked to avoid movement as much as possible while performing the task. Eight blocks of each of the three levels of the N-back task were presented in pseudorandom order (Figure 2). Each block lasted for about 100 seconds with short breaks between them in which subjects were asked to sit still and relax for the first and last 10 seconds, and then were allowed to move, review the instructions for the following condition, rest if necessary. Most subjects continued with the next condition within about 20 seconds. This procedure introduced some jitter in the block presentations which reduced the possibility of confounding the block presentation with low-frequency oscillatory physiological noise such as the Mayer wave. After 4 blocks, participants had a mandatory rest of about ten to fifteen minutes in which the lighting and airflow was increased, a non-caffeinated beverage was provided, and participants were encouraged to stretch and engage in conversation with the experimenter. This was deemed necessary to reduce effects of boredom with the repetitive nature of the task and mild discomfort due to sitting still and wearing a cap reported by some pilot study participants.

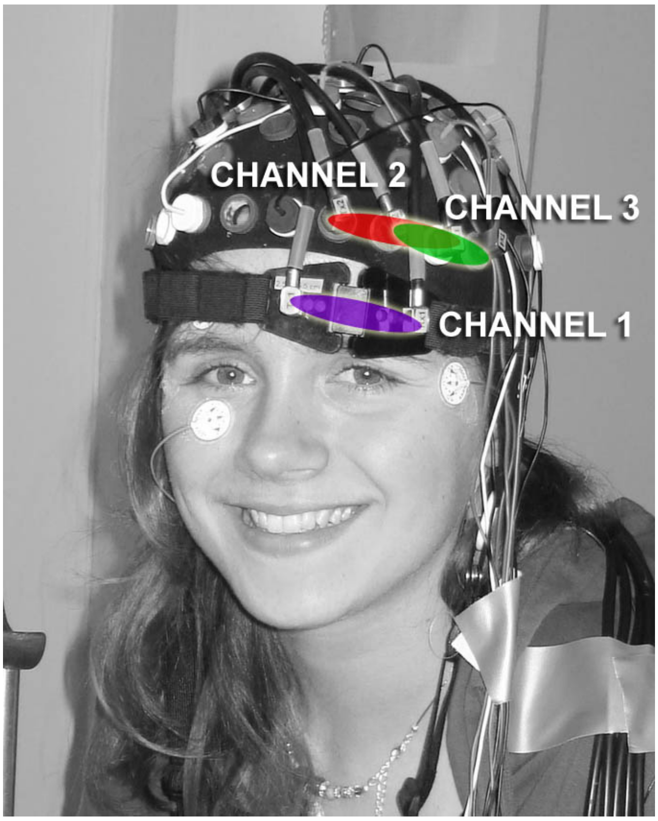


Figure 3: Locations of the three NIRS channels, located between three optode transmitter-receiver pairs. Note that channels 2 and 3 overlap. (Participant photo used with permission)

2.3. Equipment and Data Collection

The experiment used three interconnected computers. The participant's computer ran the N-back task in E-Prime, which sent event markers to the EEG and NIRS computers, allowing for accurate time calibration between the EEG and NIRS data. A desktop computer recorded EEG data using Simulink (The Mathworks, Natick, Massachusetts), and a laptop computer recorded NIRS and converted raw optical density scores into oxy, deoxy, and total haemoglobin concentrations (Artinis Medical Systems, Zetten, The Netherlands). Minor communication

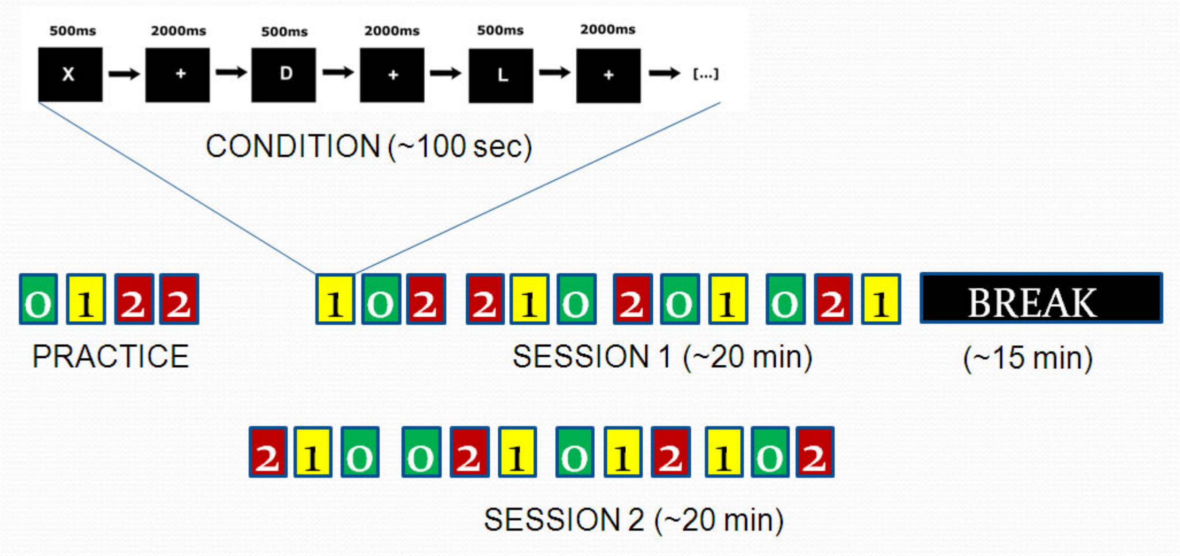


Figure 2: Graphic representation of the experimental design. Presentation order of N-back conditions is pseudorandomized

delays in the system were optically measured and accounted for during data processing.

A custom-modified electrode cap containing holes positioned for NIRS use (Guger Technologies OEG, Graz, Austria) was buttoned to a chest strap to reduce the risk of movement. A G-Tec USBamp amplifier (Guger Technologies OEG) was used with eight passive electrodes: Cz, Pz, FCz, Fz, C3, C4, F3, and F4, with a right-side forehead ground and linked mastoid references (frontal sites are strongly represented due to previous findings of changes in frontal theta activity with increases in workload [18]). The impedance of each electrode site was reduced to less than 5 kOhms prior to beginning the experiment. Data was sampled at 256Hz. A notch filter at 50Hz and a bandpass filter of 0.01 - 60Hz was applied during data collection. A continuous wave near infrared spectrometer (Oxymon MkIII, Artinis Medical Systems) was used to measure relative changes in oxyhemoglobin and deoxyhemoglobin concentrations on the left forehead, above the dorsolateral prefrontal cortex, a brain region strongly associated with working memory-dependent tasks in both NIRS and fMRI studies [17, 15]. Three independent channels were used, each with an inter-optode distance of 45mm. One was applied in a headband across the forehead, centred above the left eye (see Figure 3).

Optode sets for two channels were inserted into the modified G-Tec electrode cap, using custom-made black foam light baffles. The receptors and transmitters for these channels overlapped, each being 22.5 mm from one another and 45 mm from their counterpart (the Oxymon system allows for this due to a time-encoded pattern of laser firing). Laser wavelengths of 766 and 860 nm were used, and transmitter power and receiver gains were set to ensure sufficient light was collected. The signal was also checked visually during the training task, using the presence of the effect of the heart beat on the oxyhemoglobin concentration as an indicator that the light was being transmitted correctly through the head.

2.4. Data Analysis

The performance data was exported from E-prime and processed in MatLab to obtain average accuracies and reaction times for subject responses in each of the conditions. A one-way ANOVA was performed to assess the significance of differences between workload levels on the accuracy and reaction times, averaged over targets and non-targets for all subjects.

Custom-made MatLab (The Mathworks, Natick, Massachusetts) scripts were written to reconstruct the time course of the event markers and identify the beginning and end points of the experimental blocks within the EEG and NIRS data sets. Exploratory analyses were performed to visualize differences in the power spectra across conditions for each subject at each electrode, and to investigate patterns in the relative concentrations of oxy-, deoxy-, and total hemoglobin in each subject and each NIRS channel over the course of the experiment. As others [6] have found, we observed large differences in the patterns of the response of the power spectra between subjects. This was true also of the averages obtained from the NIRS data, where agreement with previous results (e.g. [17]) was only found in some subjects and some channels. Together, this reinforced the need for highly individually tailored classifiers.

In literature, longer window sizes tend to produce better classification results. For example, Grimes et al. [6] systematically varied the window size between 0 and 120 seconds for EEG data and found an improvement of about 15% with a 2-condition classifier between 2-second and 20-second window sizes. However, larger window sizes also increase the risk of overfitting of spectral details when using common spatial pattern filters, and reduce the captured trial-to-trial variation, which is undesirable (Christian Kothe, personal communication). Longer windows also lower the number of trials available within a data set, which is a problem when trying to estimate the accuracy of the classifier and the significance of the results from a limited amount of data [19]. A window length size of 5 seconds was selected with a view to making a real-time classifier that could make decisions for use in adaptive automation relatively rapidly. Thirty-six windows were obtained per experimental block, meaning that some overlap existed between windows. This is in keeping with the expected usage, in which classification results would be calculated on a moving window. Classification was carried out using a Beta version of PhyPA Team's MatLab based classification toolbox (TU Berlin), which is designed as a powerful classification platform for 2-class brain-computer interface data. Using two functions, it is possible to pre-process data for use by particular classifiers, and then to train and test the classifier. A cross-validation is performed to estimate the classifier's accuracy in which an entire block is kept out of the classifier training and is then used for testing. The standard deviation of a sequence of accuracy estimates is also calculated.

Due to the 2-class restriction of the software, the main analysis was done between the 0 and 2-back conditions, based on the expectation that this would produce the greatest workload difference. Classification was also performed between the 1 and 2-back conditions for comparison, due to concerns that the 0-back condition lacked the updating component of working memory, leading to a less pure measure. Due to non-stationary effects observed in the alpha frequency range of some subjects during exploratory analysis between the first and second sessions, three classifications were performed, one for each session (four blocks) separately and one for the entire experiment. In the case of one-session data, an 8-fold chronological/blockwise cross-validation was performed, and in the case of entire experiment data, a 16-fold cross-validation was performed. We report only the combined session results here, as it is a more stringent and valid measure of the classifier's ability to handle data collected over multiple sessions in which background signals related to other cognitive processes may change.

The EEG data was classified using five variations in classifier training paradigm recommended by signal processing experts from PhyPA Team for the specific experimental design. Each used either linear discriminant analysis (LDA), which uses a linear combination of features which best separate the two classes of data; or logistic regression, which bases the classification decision on fitting available data to a logistic curve. EEG features were all related to band-power within the 2-25Hz range. Some paradigms took the spatial location of the electrodes into account (using common spatial filtering or multiclass common spatial filtering). Others used different numbers spatial filters (6 or 8), or used a reduced frequency range centered around the alpha and theta peaks (3-6 and 9-12Hz).

The NIRS data were examined using an LDA classifier which took as input the average oxy-, deoxy-, and total hemoglobin concentration of the NIRS data in each channel after low-pass filtering at 0.14Hz. Only minor variations in classification accuracy were obtained by extending the upper boundary of the low-pass filter incrementally to 10Hz on a subset of data, so the cut-off value suggested most commonly in literature was retained.

A custom-written function by Christian Kothe of the PhyPA Team was then used to combine two separate training paradigms, one for the EEG and one for the NIRS. The single NIRS training paradigm was combined with each of the EEG training paradigms in turn to assess the benefits of the addition of NIRS data to EEG classifiers.

3. Results

3.1. Task performance

The results for the accuracy and reaction time of the subject per condition is displayed in Figure 4 (note that error bars display standard errors of the participant's respective averages). Although significance was not reached using a one-way ANOVA between conditions (accuracy F(2, 27) = 1.15, p = 0.33; reaction time F(2,27) = 1.83, p = 0.18), this is likely because of insufficient statistical power due to the low number of subjects and high variability between subjects in their accuracy and reaction time rather than a failure of the N-back task to adequately manipulate workload. Numerous studies have demonstrated the effectiveness of the task under comparable conditions. Using a similar N-back design, [6] only found an accuracy effect between 3-back and others, which was not used in this experiment. Reaction times were significantly different only for non-adjacent conditions. In this experiment, this effect seems to have been obscured by large variations between subjects'average response times in the 2-back condition. In comparison with Grimes' study, subjects tended to be much faster and somewhat less accurate in their responses, possibly due to differences in the wording of the task instructions.

The trend indicates that accuracy decreases and reaction time increases with increasing workload, although there may be little difference in difficulty between the 0 and 1-back conditions, a result confirmed by the subjective reports of most participants and apparently found by Grimes *et al.* [6], although only main effects across all conditions are reported. The accuracy results for non-target letter responses are likely higher than target letters due to the difference in the frequency of their appearance. The 'non-target' response is likely to be the default option. It should also be considered that subjects may partially compensate for increasing task difficulty with a greater arousal state and increased effort, in which case we could expect bigger effects between conditions of the brain signals as compared with the performance results.

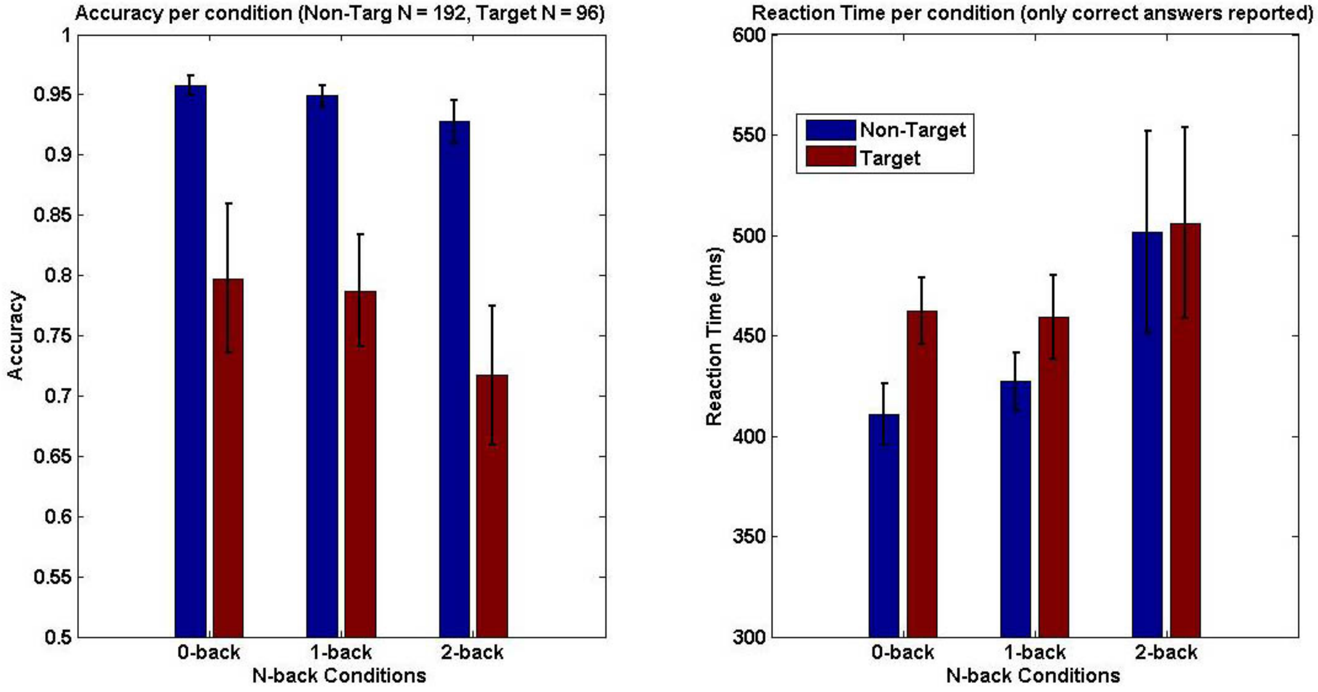


Figure 4: Accuracy and reaction time averages for participatns (N=10) over N-back conditions

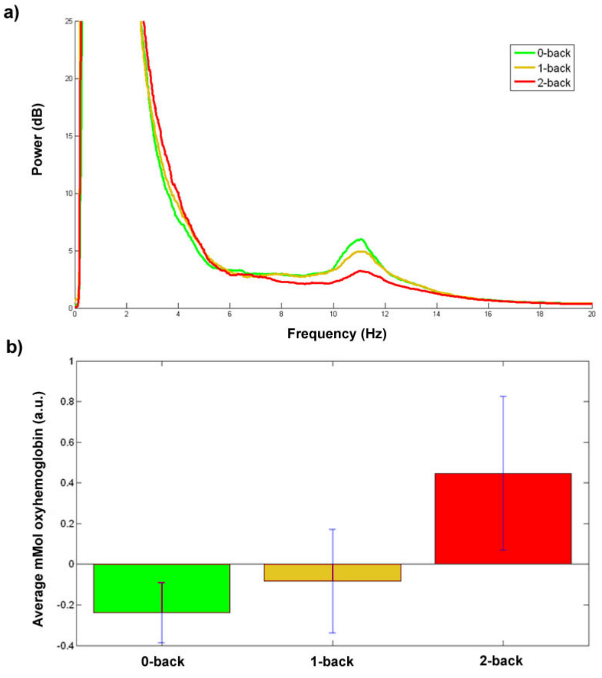


Figure 5: Sample data visualization from the subject with the greatest classification accuracy (subject 2). a) A smoothed periodogram averaged across all blocks from electrode position Fz shows an increase in theta-band activity and a decrease in alpha-band activity with increasing workload levels. b) Relative oxyhemoglobin concentrations across all blocks in a NIRS channel shows an increase with increasing workload levels. Error bars represent the standard error of the averages of the raw data within each block

3.2. Qualitative correlates of workload

In order to ascertain whether there are major and consistent differences between conditions in the data prior to classification, the data were visualized in various ways including periodograms for EEG and averages over experimental blocks for NIRS. Figure 5 shows an example of 'good' results. The power spectra averaged over all 8 experimental blocks for each condition are clearly distinguishable from one another, with theta activity increasing and alpha activity decreasing with increased workload. Similar EEG patterns were found at all electrode locations for this subject, with theta differences more strongly represented in the frontal channels. The averages of the oxyhemoglobin concentrations within each condition are presented, showing the expected increase with higher workload. Quantitative analysis at this stage is less relevant for our research question, which focuses on using short time windows and real-time applicable filters rather than grand averages over entire blocks after data cleaning, as was used in [17].

While the alpha and theta band patterns occurred frequently amongst the subjects, it is not a generalizable finding. Other idiosyncratic frequency band differences were observed, such as a shift of the peak alpha towards the low frequencies with increasing workload, or even patterns of alpha and theta band powers inverse to those expected. The results reported in literature of increased oxyhemoglobin for the NIRS data were only apparent in some subjects' data, with others showing inverse or simply inconsistent results. This could not be improved by biasing each block's average to the resting period immediately prior, taking a subset of the block where a peak hemodynamic response would be expected as a time window, or various smoothing techniques. In general, a high degree of variability made significant differences between conditions rare, with differences more likely to be found between the 0 and 2-back conditions for most subjects.

3.3. Classification

The classification paradigms are designed to optimize the weighting of specific features in order to produce the highest classification accuracy possible. This implies a high level of customization which makes group statistics on the effectiveness of a particular classifier less meaningful. No consistency between subjects was observed in which EEG classification paradigm gave the best result. Most EEG classification results for the same data set were within the range of about 5% from the average classification accuracy. We therefore extend this subject-wise customization to selecting the most effective of the classification paradigms for each subject in order to assess the effectiveness of the addition of NIRS data. Mueller-Putz and colleagues [19] emphasize the need to adequately determine if a classifier performs above chance by using confidence limits considering the number of classes, the number of trials per class, and the desired alpha value. Given our two-condition classifications, 36 windows per testing block, and an alpha level of 0.05, we can set the upper confidence limit for single tests at approximately 60% accuracy. Since the same data are used for two classifications (EEG or NIRS independently as well as a combined result), a Bonferroni correction is appropriate (alpha = 0.025), which according to experimental results [19] means a confidence interval cut-off of approximately 63% accuracy.

The classification results are presented in Figure 6. In the 0 vs. 2-back classification, 8 out of 10 subjects' EEG data and only 5 out of 10 subjects' NIRS data was classifiable above chance level. For those who had both EEG and NIRS above chance level, three showed a small increase in classification accuracy with the addition of NIRS, in the order of 2%. The relatively large standard deviations of classification accuracy during the cross-validation process (in the order of 15-20%) imply that this is not a substantial increase. The results for the 1 vs. 2-back classification are similar, with an even lower number of subjects' NIRS classification above chance levels.

4. Discussion

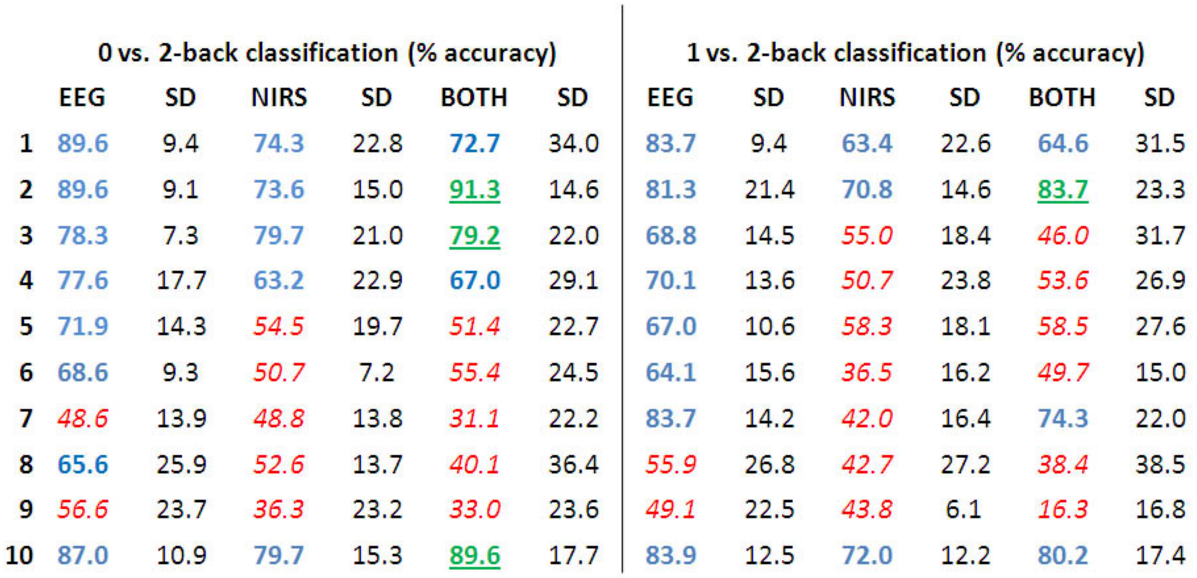
The results of this experiment do not promote the use of NIRS for the real-time recognition of correlates of workload. However, the results suggest several areas in which further investigation would help to clarify whether and under what circumstances NIRS signals could be helpful.

The classifiers showed a high standard deviation, meaning that the classifier performance varies considerably during iterations of cross-validation testing. This could result from a combination of several factors: the classifier may be overfitting the data, the data may not allow for the formation of a good/stable model, or the data set may be insufficiently large (Christian Kothe, personal communication). It is difficult to interpret the standard deviations in the context of published BCI results, as most publications to date have not reported this important information [19].

The classification results also show a decrease in accuracy after the addition of NIRS data, which demonstrates that the classifiers are not performing well, since data which does not contribute to the classification should ideally be ignored. When the degrees of freedom are too high relative to the class information available in the data, many classifiers (including those used here) deal with the situation by enforcing simplicity, which also has the effect of simplifying the parts of the data which are informative, leading to lower classification accuracies. This could be improved by recording much more data per subject. Alternately, more advanced machine learning techniques could be used such as relevance vector machines can theoretically handle this nature of problem very well (Christian Kothe, personal communication). A systematic examination of the how the features selected for classification vary with related yet distinct aspects of the user's cognitive state (e.g. boredom, distraction, drowsiness, and stress levels) would also be highly informative for guiding feature selection.

A related problem is the paucity of tools and methods, including classifier training paradigms, available to handle the classification of NIRS data. For example, in keeping with current practice of using EEG for BCIs, the PhyPA toolbox's classifiers have been designed and built primarily around EEG data, focusing on techniques for extracting bandpower and event-related potential features. Only one of the included functions was appropriate for NIRS data, and the analysis was based on a simple low-pass filter and averaged hemoglobin levels. It is likely that dedicated efforts by the machine learning and signal processing research community will develop better ways to handle the high degree of variability found in the NIRS data, allowing for better identification of robust features. Related efforts are already underway, as demonstrated by the recent release of a statistical parametric mapping tool for NIRS data [20]. Luu *et al.* have also experimented with using a large number of NIRS channels at multiple depths (different inter-optode distances) to mathematically reduce the effects of superficial hemodynamic changes from the deep sources, thus obtaining a cleaner signal from the brain area of interest [21].

It is also possible that a more extensive search through parameters such as window size, bandwidth filters, and others, would increase the accuracy of the NIRS classification results. Unfortunately, the classification as described for a single subject took approximately five hours of processing time on a high-end desktop computer, for which it was not possible to systematically vary all available classification parameters within the scope of the project.



	0 vs. 2-back classification (% accuracy)						1 vs. 2-back classification (% accuracy)					
	EEG	SD	NIRS	SD	BOTH	SD	EEG	SD	NIRS	SD	BOTH	SD
1	**89.6**	9.4	**74.3**	22.8	**72.7**	34.0	**83.7**	9.4	**63.4**	22.6	**64.6**	31.5
2	**89.6**	9.1	**73.6**	15.0	**91.3**	14.6	**81.3**	21.4	**70.8**	14.6	**83.7**	23.3
3	**78.3**	7.3	**79.7**	21.0	**79.2**	22.0	**68.8**	14.5	*55.0*	18.4	*46.0*	31.7
4	**77.6**	17.7	**63.2**	22.9	**67.0**	29.1	**70.1**	13.6	*50.7*	23.8	*53.6*	26.9
5	**71.9**	14.3	*54.5*	19.7	*51.4*	22.7	**67.0**	10.6	*58.3*	18.1	*58.5*	27.6
6	**68.6**	9.3	*50.7*	7.2	*55.4*	24.5	**64.1**	15.6	*36.5*	16.2	*49.7*	15.0
7	*48.6*	13.9	*48.8*	13.8	*31.1*	22.2	**83.7**	14.2	*42.0*	16.4	**74.3**	22.0
8	**65.6**	25.9	*52.6*	13.7	*40.1*	36.4	*55.9*	26.8	*42.7*	27.2	*38.4*	38.5
9	*56.6*	23.7	*36.3*	23.2	*33.0*	23.6	*49.1*	22.5	*43.8*	6.1	*16.3*	16.8
10	**87.0**	10.9	**79.7**	15.3	**89.6**	17.7	**83.9**	12.5	**72.0**	12.2	**80.2**	17.4

Figure 6: Classification accuracies and standard deviations of repeated accuracy measurements using cross-validation. Results from two binary classifications are reported. Results higher than the 63% confidence interval cut off are highlighted in bold font (blue). Results supporting the hypotheses are underlined (green)

Only three NIRS channels were available, leading to an increased risk of missing a region of workload-related activity. This problem could be resolved either by using a more extensive network of NIRS channels (e.g. [17] used 16 channels over the forehead region), which would also allow classifiers to take spatial relationships into consideration when selecting features. The area of interest could first be functionally localized using a short session in an fMRI scanner in which a working memory task is performed in a block design. The resulting fMRI activation map could then be used with a stereotactic guidance system to pinpoint the appropriate scalp region over which to measure NIRS. Other signals such as the deactivation of the brain's resting state network, which is found to be negatively correlated with frontal theta activity [22], may yield more robust workload-correlated hemodynamic signals (Dr. Sander Daselaar, personal communication). Hemodynamic signals originating in more posterior regions such as components of the default state network in the parietal cortex may be less susceptible to contamination from superficial blood flow changes, and classification features made up of combinations of positively and negatively correlated regions might further improve the classifier performance.

A last consideration for future work is that EEG measurements may be more practically combined with the physiological parameters mentioned in the introduction than with NIRS. For example, Iqbal *et al.* [23] have found a relationship between task difficulty and pupillary response. If further testing with NIRS systems fail to produce sufficiently powerful classifiers, workload measurement may be improved by combining this or other approaches with EEG measurement.

5. Acknowledgments

The authors gratefully acknowledge the support of the BrainGain Smart Mix Programme of the Netherlands Ministry of Economic Affairs and the Netherlands Ministry of Education, Culture and Science; Dr. Thorsten Zander and Christian Kothe for use of and assistance with the PhyPA MatLab toolbox; and Antoon Wennemers and Giovanni Beltrame for technical assistance.

References

[1] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, P. L. Craven, EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks, Aviation, Space, and Environmental Medicine 78 (5 Suppl) (2007) B231–244, PMID: 17547324.

[2] M. W. Scerbo, Adaptive automation, in: R. Parasuraman, M. Rizzo (Eds.), Neuroergonomics, Oxford University Press US, 2007, pp. 239–252.

[3] J. Kohlmorgen, G. Dornhege, M. L. Braun, B. Blankertz, K. Müller, G. Curio, K. Hagemann, A. Bruns, M. Schrauf, W. E. Kincses, Improving human performance in a real operating environment through Real-Time mental workload detection, in: G. Dornhege, J. del R. Millán, T. Hinterberger, D. J. McFarland, K. Müller (Eds.), Toward Brain-Computer Interfacing, The MIT Press, Cambridge, Massachusetts, 2007, pp. 409–422.

[4] T. de Greef, H. Arciszewski, J. Lindenberg, J. van Delft, Adaptive automation evaluated, TNO report TNO-DV 2007 A610, TNO Defence, Security and Safety, The Netherlands (2007).

[5] E. Palmer, D. Kobus, The future of augmented cognition systems in education and training, in: Foundations of Augmented Cognition, 2007, pp. 373–379.

[6] D. Grimes, D. S. Tan, S. E. Hudson, P. Shenoy, R. P. Rao, Feasibility and pragmatics of classifying working memory load with an electroencephalograph, in: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ACM, Florence, Italy, 2008, pp. 835–844.

[7] J. C. Lee, D. S. Tan, Using a low-cost electroencephalograph for task classification in HCI research, in: Proceedings of the 19th annual ACM symposium on User interface software and technology, ACM, Montreux, Switzerland, 2006, pp. 81–90.

[8] F. Matthews, B. A. Pearlmutter, T. E. Ward, C. Soraghan, C. Markham, Hemodynamics for brain-computer interfaces: optical correlates of control signals (2008).

[9] K. Müller, M. Tangermann, G. Dornhege, M. Krauledat, G. Curio, B. Blankertz, Machine learning for real-time single-trial EEG-analysis: from brain-computer interfacing to mental state monitoring, Journal of Neuroscience Methods 167 (1) (2008) 82–90, PMID: 18031824.

[10] M. Izzetoglu, S. Bunce, K. Izzetoglu, B. Onaral, , Pourrezaei, Functional brain imaging using near-infrared technology, IEEE Engineering in Medicine and Biology Magazine 26 (4) (2007) 38–46.

[11] H. Ogata, T. Mukai, T. Yagi, A study on the frontal cortex in cognitive tasks using near-infrared spectroscopy, Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2007 (2007) 4731–4734, PMID: 18003062.

[12] R. P. Kennan, S. G. Horovitz, A. Maki, Y. Yamashita, H. Koizumi, J. C. Gore, Simultaneous recording of Event-Related auditory oddball response using transcranial near infrared optical topography and surface EEG, NeuroImage 16 (3, Part 1) (2002) 587–592.

[13] J. C. Gore, S. G. Horovitz, C. J. Cannistraci, P. Skudlarski, Integration of fMRI, NIROT and ERP for studies of human brain function, Magnetic Resonance Imaging 24 (4) (2006) 507–513, PMID: 16677957.

[14] Y. Tong, E. J. Rooney, P. R. Bergethon, J. M. Martin, A. Sassaroli, B. L. Ehrenberg, V. V. Toi, P. Aggarwal, N. Ambady, S. Fantini, B. Chance, R. R. Alfano, B. J. Tromberg, M. Tamura, E. M. Sevick-Muraca, Studying brain function with near-infrared spectroscopy concurrently with electroencephalography, in: Optical Tomography and Spectroscopy of Tissue VI, Vol. 5693, SPIE, San Jose, CA, USA, 2005, pp. 444–449.

[15] A. M. Owen, K. M. McMillan, A. R. Laird, E. Bullmore, N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies, Human Brain Mapping 25 (1) (2005) 46–59, PMID: 15846822.

[16] T. D. Wager, E. E. Smith, Neuroimaging studies of working memory: a meta-analysis, Cognitive, Affective & Behavioral Neuroscience 3 (4) (2003) 255–274, PMID: 15040547.

[17] H. Ayaz, M. Izzetoglu, S. Bunce, T. Heiman-Patterson, B. Onaral, Detecting cognitive activity related hemodynamic signal for brain computer interface using functional near infrared spectroscopy, in: 2007 3rd International IEEE/EMBS Conference on Neural Engineering, Kohala Coast, HI, USA, 2007, pp. 342–345.

[18] O. Jensen, C. D. Tesche, Frontal theta activity in humans increases with memory load in a working memory task, European Journal of Neuroscience 15 (8) (2002) 1395–1399.

[19] G. Müller-Putz, R. Scherer, C. Brunner, R. Leeb, G. Pfurtscheller, Better than random: a closer look on BCI results, Rome, 2007, pp. 95–96.

[20] J. C. Ye, S. Tak, K. E. Jang, J. Jung, J. Jang, NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy, NeuroImage 44 (2) (2009) 428–447.

[21] S. Luu, T. Chau, Decoding subjective preference from single-trial near-infrared spectroscopy signals, Journal of Neural Engineering 6 (1) (2009) 016003.

[22] R. Scheeringa, M. C. Bastiaansen, K. M. Petersson, R. Oostenveld, D. G. Norris, P. Hagoort, Frontal theta EEG activity correlates negatively with the default mode network in resting state, International Journal of Psychophysiology 67 (3) (2008) 242–251.

[23] S. T. Iqbal, X. S. Zheng, B. P. Bailey, Task-evoked pupillary response to mental workload in human-computer interaction, in: CHI '04 extended abstracts on Human factors in computing systems, ACM, Vienna, Austria, 2004, pp. 1477–1480.