Article

Effects of Performance and Task Duration on Mental Workload during Working Memory Task

Kosar Khaksari, Emma Condy, John B. Millerhagen©, Afrouz A. Anderson, Hadis Dashtestani and Amir H. Gandjbakhche *©

National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD 20892, USA
* Correspondence: Gandjbaa@mail.nih.gov

Received: 2 July 2019; Accepted: 23 August 2019; Published: 28 August 2019

Abstract: N-back is a working memory (WM) task to study mental workload on the prefrontal cortex (PFC). We assume that the subject’s performance and changes in mental workload over time depends on the length of the experiment. The performance of the participant can change positively due to the participant’s learning process or negatively because of objective mental fatigue and/or sleepiness. In this pilot study, we examined the PFC activation of 23 healthy subjects while they performed an N-back task with two different levels of task difficulty (2-, and 3-back). The hemodynamic responses were analyzed along with the behavioral data (correct answers). A comparison was done between the hemodynamic activation and behavioral data between the two different task levels and between the beginning and end of the 3-back task. Our results show that there is a significant difference between the two task levels, which is due to the difference in task complication. In addition, a significant difference was seen between the beginning and end of the 3-back task in both behavioral data and hemodynamics due to the subject’s learning process throughout the experiment.

Keywords: fNIRS; working memory; N-back; prefrontal cortex; workload

1. Introduction

Working memory (WM) refers to a system that holds sensory information for processing and integration, allowing for cognitive understanding [1]. The manipulation of items held within this short-term memory storage is done during everyday cognitive functions, such as decision making and problem solving, where neural correlates of WM are a topic of investigation [1,2]. The ability to access WM is critical for daily life, and WM dysfunction has been correlated with deficits in neurodegenerative disorders such as Alzheimer’s or Parkinson’s [1,3,4], along with neurodevelopmental disorders such as attention-deficit/hyperactivity disorder (ADHD) [5] and autism spectrum disorder (ASD) [6]. An imaging approach to objectively measure WM would provide broader understanding of its neural correlates along with aiding in the detection of impairments affected across a range of psychopathologies.

Functional magnetic resonance imaging (fMRI) has been the traditional imaging modality used to study the associations between functional brain activation patterns and brain regions involved in working memory [7–9]. However, fMRI is relatively expensive [10,11] and requires that the subject remain still to obtain usable data. In some instances, functional near-infrared spectroscopy (fNIRS) should be the modality of choice for observing brain activation when the region of interest is within the cortex [12,13]. fNIRS, which is non-invasive and relatively inexpensive, can be used for screening in more patient-friendly settings [13–15]. fNIRS measurement is based on the diffusion of photons from the near-infrared region of the light spectrum. The difference in light absorbance of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) allows for the detection of hemodynamic fluctuations in the cortex [15,16]. Studies using fNIRS have replicated fMRI findings for tasks evoking...
prefrontal activity [17,18]. Compared to fMRI, fNIRS has advantages and disadvantages. While fMRI is the gold standard for in vivo imaging of the human brain, fNIRS has exceptional portability and robustness to noise, along with a higher temporal resolution [13,19–23].

Previous imaging studies have shown that working memory exertion is localized to prefrontal cortex (PFC) activity [1,24] and that specific regions of the PFC may be responsible for different aspects of working memory [10]. The dorsolateral prefrontal cortex (DLPFC) is important in monitoring working memory and manipulating retained information, as opposed to the storage of working memory information [9,25]. It is also involved in increasing task performance, potentially assisting in lessening overall mental exertion [26]. The medial prefrontal cortex (MPFC) is involved in decision making [27,28], along with short-term maintenance of emotion [29], and may also be involved in working memory maintenance [27]. A study using fNIRS reported higher activation of DLPFC regions than MPFC regions during a WM maintenance task, further suggesting the role DLPFC regions play in WM monitoring and cognitive control [7]. It is worth investigating the functionality of the DLPFC and MPFC in behavioral tasks to understand the role these regions play in monitoring and maintenance.

Behavioral assessments of WM involve mental workload exertion and assess performance based on successfully recruiting WM while employing varying degrees of mental workload. Mental workload describes the exertion of cognitive systems; tasks with higher cognitive demands elicit a higher mental workload [7,30]. States of increased mental workload occur with working memory tasks of a higher difficulty (e.g., greater memory spans) and result in increased hemodynamic activity in fNIRS studies [7,31]. Some studies have suggested WM to be trainable, where participants can acquire learning throughout a task and thus increase performance [32,33]. Additionally, there is support for a “dual-processing theory”, where increased task performance eventually induces neural efficiency through a lower mental workload and less hemodynamic activity [33–35]. Moreover, other cortical areas may be involved in WM training. Studies suggest WM training may result from increased connectivity between frontal and parietal networks or increased DLPFC activity [32–36]. Thus, there is a need for more neuroimaging research on WM training [32], and no fNIRS studies to date have investigated the neurological correlates of training on WM tasks.

The N-back task is a working memory assessment that was first introduced by Kirchner (1958). The N-back uses varying levels of mental load, where performance is indicative of WM activation [8]. Imaging studies have found N-back load level to correlate with greater PFC activation [7,10,15]. N-back performance is also potentially trainable, such that repeated exposure to the task increases performance [37]. Using the N-back task and fNIRS, we intend to further investigate the neurological correlates of WM in prefrontal regions involved in WM exertion. Additionally, we aim to investigate WM training through behavioral and hemodynamic data.

In this work, we examined how the subject’s performance changes, hemodynamically and behaviorally, between two different tasks with different mental workloads (2-back vs. 3-back) and over the course of the task with a higher mental workload (3-back). Thus, we focused on the three different conditions of an N-back experiment consisting of 2-back and 3-back tasks. These three conditions are: (1) the last four blocks of the 2-back task (2-back(e)), (2) the first four blocks of the 3-back task (3-back(b)), and (3) the last four blocks of the 3-back task (3-back(e)). We analyzed the total hemoglobin (HbT) concentration in conjunction with performance on the task for each of these conditions. Statistical analysis was done in order to compare the results. We also investigated the spatial variation of hemodynamics on the forehead between four different regions of the PFC. We were interested in assessing hemodynamic activity differences across prefrontal regions, along with hemodynamic exertion and behavior. We hypothesized that training occurs if participants have a lower hemodynamic response over time with higher behavioral performance, indicating less cognitive effort exerted for increased performance.
2. Materials and Methods

This study was a part of an ongoing protocol approved by the National Institute of Health Institutional Review Board (IRB #10-CH-0198).

2.1. Subjects

Fifteen female volunteers with a mean age of 35 (+/−15 years) and 8 male volunteers with a mean age of 39 (+/−16) were recruited for this study. In order to ensure that there were no pre-training influences, only those participants who had never participated in a prior N-back study were selected as subjects. Every participant provided consent prior to their participation.

2.2. Task and Experimental Design

Subjects completed 2-back and 3-back conditions for visual N-back tasks. This experiment was modeled after the N-back study described by Herff and colleagues [38]. The task was performed inside a quiet room. Prior to start of the task, the system was checked to ensure the participants were able to see the letters clearly on the screen. Participants’ responses and reaction times (milliseconds) during the task were recorded using the E-Prime 2.0 software package (Psychology Software Tools, Inc., Sharpsburg, PA, USA). Each task contained of 10 trials in which subjects would observe the letters on the screen. Every trial included 3 ± 1 correct answers (targets). The inter-stimulus duration was 1.5 s, and each stimulus was presented for 500 milliseconds. Each block presented 22 letters for a total trial time of 44 s. There was a 15 s rest period in between trials. Figure 1 shows a set of 3 trial tasks with their corresponding rest periods in between. General instructions regarding N-back tasks were given to the subjects prior to the start of the experiment and placement of the NIRS probes. Specific instructions for each set were provided before the onset of the set. Subjects were asked to click when they saw the target stimuli, based on the instructions. For 2-back conditions, the subjects clicked when the target stimulus letter was the same as the letter from 2 steps prior. For 3-back conditions, they clicked when the target stimulus letter was the same as the letter from 3 steps before. A cross was displayed for 15 s between trials, and the subjects were instructed to relax. This relaxation period was meant to provide time for hemoglobin levels as monitored by NIRS to return to baseline values. These resting periods were not included in our analysis. The task blocks were administered in order of workload: 2-back and then 3-back. We recorded each subject for a total of 14 min (14 trials of 44 s, 15 baseline pauses of 15 s, and a set of instructions of 5 s) [39,40].

![Diagram of task flow](image)

**Figure 1.** The N-back task flow for this study. The task consists of three periods: a 5 s instruction, a 44 s N-back trial, and a 15 s resting period. During the N-back trials, different letters were shown on the monitor in front of the participant, who was asked to click on the target letters as quickly and accurately as possible.
2.3. NIRS Data Analysis

NIRS data was collected using a 16-channel NIRS probe (fNIR Devices, LLC, Potomac, MD, USA) placed on the subject’s forehead. The sampling rate of the machine was 2 Hz. CobiStudio software was used to record the raw NIRS signals at the two wavelengths of 730 and 850 nm [41]. The NIRS headband with embedded sensors consisting of 4 sources and 8 detectors (16 channels) was positioned on the subject’s forehead, with the center of the band placed approximately on the Fpz location (international 10–20 system) [42]. This position has been used several times in prior NIRS studies. The source-detector separation distance was 2.5 cm for all channels. Figure 2 shows the probe design. The NIRS data processing was done in MATLAB R2017a.

![Figure 2. Probe diagram on the prefrontal cortex. Red circles are sources, and blue squares are detectors. All source-detector distances are 2.5 cm.](image)

Biological and technical artifacts affect the signals recorded by fNIRS. Based on suggested techniques for artifact removal, we used a low pass filter with a cutoff frequency of 0.1 Hz to remove higher frequency noise such as heartbeat and respiration signals [43]. Subsequently, the linear and nonlinear trends in the signals were removed by fitting a low order (an order of 3) polynomial to the fNIRS signals and subtracting it from the original signal [44]. This enabled us to focus our analysis on the hemodynamic response to the given task.

Overall, each trial was 44 s long with a resting period of 15 s between trials. For the analysis, we considered an extra 0.5 s after each N-back trial, which added one data point (sampling frequency of 2 Hz) for each trial.

The recorded raw intensity NIRS signals were converted to oxyhemoglobin, deoxyhemoglobin, and total hemoglobin using the modified Beer–Lambert law (mBLL) [45]. Two differential pathlength factors (DPFs) were used for the two wavelengths in the analysis (6.51 at 730 nm and 5.92 at 850 nm) based on values found from previous NIRS studies on the human forehead [46]. We examined the NIRS signal with respect to the baseline for each subject in order to minimize the effects of extra-cerebral layer contamination. The source-detector distance (2.5 cm) was enough to ensure that the cerebral cortex was sampled [47]. Additionally, in previous NIRS studies, it has been shown that the task-related effects on extra-cerebral layer hemoglobin concentration is negligible [13,18,47–49].

In this study, only changes in HbT were considered in our analysis because these signals are more robust than HbO and HbR when examining task-related oscillations. The HbT signal contains both HbR and HbO signals, and in studies such Gagnon et al., HbT is less sensitive to pial vein contamination than HbO or HbR and provides better spatial specificity of the cerebral activation [23].

In this study, we studied different mental workloads using the spatial characteristics of the hemodynamic responses during an N-back trial (region of interest (ROI)). We also examined how the length of the experiment affected the subject’s performance. In order to do so, we considered the last 4 blocks of the 2-back (2-back(e)), the first 4 blocks of the 3-back (3-back(b)), and the last 4 blocks of the 3-back task (3-back(e)) (3 conditions). The concentration of total hemoglobin was determined for each channel and each subject. Total hemoglobin concentrations were averaged over each condition, excluding the resting periods.
We also analyzed the behavioral data in conjunction with the hemodynamic data in order to determine if there is a relationship (correlation) between the fNIRS signals and the behavioral data.

2.4. Behavioral Data Analysis

Performance on the working memory tasks was scored first by calculating the z-transformed hit rate (HR) and z-transformed false alarm rate (FAR) in service of calculating $d'$, a sensitivity metric that accounts for both HR and FAR to measure participants’ performance accuracy on a task [50]. A higher $d'$ score indicates that a participant performed better on the task. Performance on the 2-back(e), 3-back(b), and 3-back(e) was calculated by deriving $d'$ for each block and then averaging the $d'$ value within each condition. Since performance ($d'$) on the 2-back and 3-back tasks was not normally distributed in our sample, a traditional parametric ANOVA could not be used to analyze these data. For this reason, a non-parametric one-way repeated measures ANOVA (i.e., Friedman test) was used to evaluate differences in performance on the three tasks in R.

3. Results

First, we analyzed HbT concentration over four different regions on the forehead during different tasks to see whether regions of interest (ROIs) of the PFC were differentially activated across the three task conditions (Figure 3). A repeated measure analysis of variance (ANOVA) using R software was performed to examine whether HbT concentration varied by task condition or ROI. Based on the ANOVA output, there was no significant interaction between ROI and task condition ($F(6,23) = 0.087, p = 0.9975, d = 0.5$), nor was there a main effect of ROI ($F(3,23) = 0.22, p = 0.89, d = 0.32$). However, a significant difference was seen between the three task conditions ($F(2,23) = 8.18, p = 0.00036, d = 0.45$). Since there were no differences in ROI activity across the conditions, we averaged the HbT concentration across all four ROIs on the forehead and analyzed differences in global prefrontal activation across the tasks and in relation to behavioral performance on these tasks.

![Figure 3](image-url)  
**Figure 3.** Difference in total hemoglobin (HbT) concentration between regions of interest (ROIs) and tasks. ROIs include the left dorsolateral prefrontal cortex (LDLPFC), the left medial prefrontal cortex (LMPFC), the right dorsolateral prefrontal cortex (RDLPC) and the right medial prefrontal cortex (RMPFC). Bars are +/- 1.5*IQR.

To compare differences in global prefrontal activation across the three tasks, a series of post-hoc paired sample Wilcoxon signed-rank tests were used. Global prefrontal HbT concentration was significantly lower during the 2-back(e) ($M = 0.55, SD = 0.72$) than the 3-back(e) ($M = 1.64, SD = 1.44$) ($V = 9, p = 0.0000079$) and the 3-back(b) ($M = 2.04, SD = 1.80$) ($V = 5, p = 0.000024$). Global prefrontal HbT concentration was also lower during the 3-back(e) than the 3-back(b) ($V = 207, p = 0.035$) (Figure 4b).
Wilcoxon signed rank paired samples t-tests revealed that participants scored significantly higher on the performance. Thus, hemodynamic measures (HbT) and behavioral performance changed from the beginning to the end of a working memory task. We also examined whether spatial patterns of PFC activation captured through fNIRS differed as a function of workload; however, no effect was found between the conditions across the four regions of the PFC (left DLPFC, left MPFC, right MPFC, and right DLPFC).

While prior studies have investigated working memory-load-dependent activation in different regions of the PFC with fMRI [51], we did not see regional differences in our analysis with fNIRS. Thus, we averaged the signals over the four different regions of the PFC and focused on global PFC activation for further analysis. The results of our analysis demonstrate that we are able to differentiate between hemodynamic activity produced by performing tasks of varying workloads in the case of 2- vs. 3-back and for the 3-back end vs. beginning. This replicates previous findings that fNIRS is capable of determining differences in mental workload exertion induced through varying N-back levels [17–19,22,39,52–59]. Additionally, we found decreased HbT concentrations in the 3-back(e) compared to the 3-back(b), along with increased behavioral responses on the 3-back(e) compared to the 3-back(b). This learning effect suggests a trend of neural efficiency, where behavioral learning is correlated with decreased hemodynamic activity [60,61]. While previous literature has found increased hemodynamic activity in regard to N-back learning, these studies focused on long-term differences between N-back loads, ROI connectivity, or resting state hemodynamics and did not look at learning...
across a specific task block [24,28]. Our results thus show that short-term learning may result in decreased hemodynamic activity across a block of the N-back, supplementing previous research by providing evidence on short-term learning through an N-back task.

Working memory dysfunction is a noted symptom of many disorders, including ADHD [5], ASD [6], Parkinson’s disease, and Alzheimer’s disease [26]. In addition, WM capacity has been shown to be a reliable predictor of higher functions including problem solving, reasoning, and reading ability. Interestingly, Olsen et al. [50] and Buschkuehl, et al. [23] have shown that it is possible to observe and locate training-related changes in frontal and parietal activation. The ability to objectively measure WM by workload can aid in the detection of impairments in the functional anatomy that lie at the root of higher order conditions. fNIRS in conjunction with N-back as a workload-dependent WM task may be utilized as an objective early-onset diagnostic tool for disorders that are correlated with WM dysfunction.

There are limitations to this study. First, we did not counterbalance our administration of the N-back load levels. This would have helped answer more questions on learning and training, along with allowing us to assess mental fatigue, as it was possible that participants were simply becoming more tired toward the end of the task. Based on our results, the decrease in hemodynamic activity is unlikely to be due to mental fatigue since task performance did not decrease with task progression. Moreover, significant increases in performance and decreases in hemodynamic activity only occurred at the group level. It is important to note that the results of this study are based on the overall group analysis of all subjects. Although the observed trend was true for this study group, it does not imply that the trend was observed per individual subject. Additionally, multiple administrations of the same paradigm across different time points would have better accounted for learning effects. For instance, if a decrease in hemodynamic activity was always correlated with increases in behavioral performance across one week or month, this would further provide evidence for neural efficiency. Finally, our results report no significant differences in DLPFC and MPFC activation, conflicting with prior research. This emphasizes the need for further investigation into the function of different PFC regions in WM.

In the future, we plan on applying this method to study populations affected by working memory problems. Our ultimate goal is to use the fNIRS and N-back paradigm to develop a robust system that can be used to aid in the clinical diagnosis of developmental delays and Parkinson’s disease. Since this was a pilot study, the task was designed based on the order of workload. In future studies, a counterbalanced task design would be useful to further clarify the physiological or behavioral effect. We designed the study in this way to look at the learning effect for each of the workloads separately. Although we applied appropriate filtering to remove physiological noise, it would be useful to include physiological measures in the future to help further improve signal quality. In order to clarify possible effects of superficial tissue layers, future studies should also include short-channel distances. In this study, we only focused on the learning process since the length of the experiment was short. For future work that may have longer duration experiments starting from 0-back and ending with 3-back, we will have to consider the effects of objective mental fatigue, sleepiness, as well as the subject’s willingness to participate [60,61].

5. Conclusions

Using fNIRS and behavioral analysis, we explored the effects of task duration on mental workload in an N-back experiment. In this work, the results suggest an important role of task duration in mental workload due to the learning process. In the case of longer experiments, we will consider other terms such as sleepiness, mental fatigue, and willingness, as other parameters may influence changes in hemodynamics and behaviors throughout the experiment. The results of this study highlight the importance of task duration in the performance of functional tasks such as working memory. These findings can have implications for understanding the N-back behavioral findings during one task, which can help us design the experiments and analyze the data better.

Funding: This research received no external funding.

Acknowledgments: We thank our participants who devoted their time to this research.

Conflicts of Interest: The authors declare no conflict of interest.

References
27. Euston, D.R.; Gruber, A.J.; McNaughton, B.L. The Role of Medial Prefrontal Cortex in Memory and Decision Making. *Neuron* 2012, 76, 1057–1070. [CrossRef]
53. Di Lorenzo, R.; Pirazzoli, L.; Blasi, A.; Bulgarelli, C.; Hakuno, Y.; Minagawa, Y.; Brigadoi, S. Recommendations for motion correction of infant fNIRS data applicable to data sets acquired with a variety of experimental designs and acquisition systems. *NeuroImage* 2019, 200, 511–527. [CrossRef]
