Brain Measurement for Usability Testing and Adaptive Interfaces: An Example of Uncovering Syntactic Workload with Functional Near Infrared Spectroscopy

Leanne M. Hirshfield* **Erin Treacy Solovey*** **Audrey Girouard*** **James Kebinger*** **Robert J. K. Jacob*** **Angelo Sassaroli+** **Sergio Fantini+**

*Computer Science Department
Tufts University
Medford, MA 02155, USA
{leanne.hirshfield, erin.solovey, audrey.girouard, james.kebinger, robert.jacob}@tufts.edu

+Biomedical Engineering Department
Tufts University
Medford, MA 02155, USA
{angelo.sassaroli, sergio.fantini}@tufts.edu

ABSTRACT

A well designed user interface (UI) should be transparent, allowing users to focus their mental workload on the task at hand. We hypothesize that the overall mental workload required to perform a task using a computer system is composed of a portion attributable to the difficulty of the underlying task plus a portion attributable to the complexity of operating the user interface. In this regard, we follow Shneiderman's theory of syntactic and semantic components of a UI. We present an experiment protocol that can be used to measure the workload experienced by users in their various cognitive resources while working with a computer. We then describe an experiment where we used the protocol to quantify the syntactic workload of two user interfaces. We use functional near infrared spectroscopy, a new brain imaging technology that is beginning to be used in HCI. We also discuss extensions of our techniques to adaptive interfaces.

ACM Classification Keywords: H5.2 [Information interfaces and presentation]: User Interfaces. - Graphical user interfaces.

Author Keywords: evaluation, syntactic, BCI, workload

INTRODUCTION

A well designed computer interface should be nearly transparent, allowing the user to focus on the task at hand [20]. This is a common goal for experts in Human Computer Interaction (HCI) who conduct research on designing and evaluating user interfaces (UIs). It is well known that effectively designing and evaluating UIs enhances user performance, increases user satisfaction, and increases safety [23]. Therefore, determining the most effective techniques for evaluating UIs remains a popular area of research.

Measurements of accuracy and time to complete a task are common quantitative measures used in UI evaluation. However, measuring user states such as mental workload is done by qualitatively observing subjects or administering subjective surveys to subjects. These surveys are often given after a task has been completed, lacking insight into the user's changing experiences during the task. Our research addresses these evaluation challenges with respect to mental workload. We use a new, non-invasive brain sensing technique called functional near infrared spectroscopy (fNIRs) to make real time, objective measurements of users' mental workload while working with UIs. fNIRs was introduced in the 1990s [3, 7] to complement, and in some cases overcome, practical and functional limitations of EEG and other brain devices.

Brain measurement in HCI has typically been used to investigate overall system difficulty where the UI and task are viewed as one entity [10, 15, 22]. However, UI designers generally want to minimize the workload required to work with UI, allowing the user to focus more workload on the primary task. Therefore, **a high workload detected while a user works with a system is not necessarily a bad thing**; it could indicate that the user is immersed in the task. How can UI evaluators know if a high workload measurement is due to the UI or to the task?

Also, most research measuring mental workload in HCI involves the creation of controlled tasks where workload is manipulated as an independent variable throughout the experiment [10, 15, 22]. For example, researchers in the Augmented Cognition program developed the GUI based Warship Commander Task (WCT) [11]. In this mock command and control environment, subjects made actions based on planes flying through their airspace during a 75 second time period. Workload was manipulated by changing the number of planes that required attention in the airspace during that time [11]. Figure 1 shows two workload conditions in the WCT. Researchers looked at the brain activity of subjects to determine if the increased workload (number of planes) caused different brain activity, as measured by various brain imaging devices.

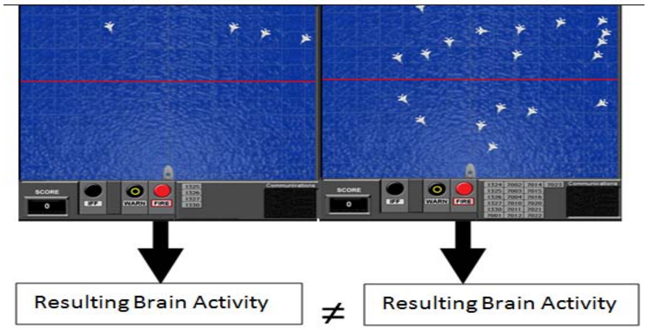

Figure 1: Two workload conditions from the WCT [11], with lower workload on the left. The UI and task are seen as one entity, and workload is determined as an independent variable

This setup is similar to many experiments on workload in HCI where the UI and task are seen as one entity and known levels of workload are used as independent variables in the experiment. How can we gain measures of users' workload while working with UIs in real world settings, when the workload associated with the UI is *not determined* beforehand? Also, UIs and tasks that have been created outside of lab settings will cause complex brain activity in the brain's cognitive resources. Thus, it will be difficult to gain measures of users' workload, and even more difficult to attribute workload to aspects of the UI.

To address these issues and move toward the goal of evaluating UIs with brain measurement, we propose to separate mental workload into multiple components. We hypothesize that the overall workload required to perform a task using a computer is composed of a portion attributable to the difficulty of the task itself plus a portion attributable to the complexity of operating the UI. In this regard, we follow Shneiderman's theory of syntactic and semantic components of a UI [20]. The semantic component involves the workload needed to complete the task. The syntactic component includes interpreting the UI's feedback and formulating and inputting commands to the UI. A goal in UI design is to reduce the mental effort devoted to the syntactic aspects so that more workload can be devoted to the underlying task, or semantic aspects.

We believe that brain measurement can be used as an additional metric in usability studies (and in adaptive UIs) to acquire real time, objective measurements that shed light on the syntactic workload associated with UIs. The brain is a complex structure, making it nearly impossible to completely separate resources devoted to processing the semantic (task) and syntactic (UI) elements of workload. However, we posit that brain measurement can be used to acquire valuable information about the syntactic workload of a UI. We focus on the interacting cognitive subsystems, or cognitive resources, that work together to process information (i.e., working memory, executive processing, visual search) while a user works with a UI and task.

As an initial step in this direction, we designed an experiment to measure the syntactic workload of two specially constructed UIs that involve users traversing through hyperspace while conducting an information retrieval task. These two UIs, described in detail later, are based on benchmark cognitive psychology tasks that place demands on users' spatial working memory (WM). We do not separate "semantic" and "syntactic" workload in the brain directly, but rather we constructed our UI and task so that the syntactic portion maps directly onto spatial WM, and the semantic portion maps onto verbal WM. We developed a novel experimental protocol that merges low level cognition experiments with high level usability evaluation. We used our protocol to acquire fNIRs brain measurements, and we created a set of data analysis algorithms that enable us to make inferences about the syntactic workload (i.e., spatial WM) of each interface.

Therefore, this work provides two primary contributions to the HCI realm. First, our novel experiment protocol and our data analysis algorithms can help usability experts, or designers of adaptive systems, to gain information about the workload experienced by computer users in the various cognitive resources in their brain while working with a computer system. As opposed to most brain measurement in HCI, our work demonstrates ways that workload can be measured in real working conditions, when the workload of operating a computer system is not known beforehand. Second, we designed two simplified UIs and a task, and we ran an experiment using our protocol where we acquired quantitative, real time measurements of the syntactic workload (i.e.,spatial WM) of our specially constructed UIs. The experiment ties Shneiderman's theory on syntactic and semantic workload to quantifiable brain measurements.

The rest of this paper proceeds as follows: First, we describe related work. Then we give an overview of our experimental protocol and we present our experiment designed to measure the syntactic workload of our UIs. We then describe the algorithms developed to analyze the brain data. After presenting the experiment results we discuss implications of our findings and future work in this area.

RELATED WORK

Our interdisciplinary research builds on work in biomedical engineering, cognitive psychology, HCI, and data mining.

Brain Imaging

Brain imaging techniques such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) have been widely used to learn about human brain activity. Although these techniques provide valuable insight into the brain, they require motionless subjects in constricted positions (fMRI), and they expose subjects to hazardous materials (PET) or loud noises (fMRI) [10]. These techniques are not suitable for measuring the brain in normal working conditions.

For this reason, the electroencephalograph (EEG) has been of interest to researchers looking to non-invasively measure users' brain activity [8, 13, 15, 17]. EEG is the most studied non-invasive brain imaging device due to its fine temporal resolution, ease of use, and low cost. While EEG is less invasive than other techniques, it is susceptible to noise since fluid, bone, and scalp separate the electrodes from brain activity. EEG also takes longer to set-up than fNIRs, has low spatial resolution, and is very sensitive to subject movement and electrical interference [15].

Functional Near Infrared Spectroscopy

We use fNIRs to add increased comfort and portability to subjects and to overcome some of the practical and functional limitations of EEG [10]. The tool, still a research modality, uses light sources in the near infrared wavelength range (650-850 nm) and optical detectors to probe brain activity. Light sources and detection points are defined by means of optical fibers which are held on the scalp with an optical probe (Figure 2). Deoxygenated and oxygenated hemoglobin are the main absorbers of near infrared light in tissues during hemodynamic and metabolic changes associated with neural activity in the brain [3]. We can detect these changes by measuring the diffusively reflected light that has probed the brain cortex [3, 10, 22]. Researchers have shown that by placing the probes on a subject's forehead, fNIRs provides an accurate measure of activity within the frontal lobe of the brain [10]. The frontal lobe has been found to play a part in memory and executive control. These results are promising when combined with the fact that fNIRs is safe, portable, less invasive than other imaging techniques, and has been implemented wirelessly, allowing for use in real world environments [10, 18].

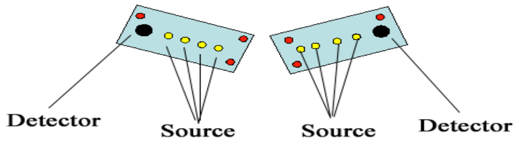

Figure 2: two probes and their sources and detectors.

Mental Workload

The term *workload* is used in literature from various fields and its definition varies widely[9, 21, 23]. In this section, we discuss workload in cognition, HCI, fNIRs research.

Workload in Cognitive Psychology Literature

Many studies in experimental psychology are based on Baddeley's model of WM [1, 21]. The original model posits that there are two separate storage spaces for short term memory, which provides volatile, short term maintenance of data. The visuo-spatial sketchpad holds visual and spatial information in short term memory and the phonological loop holds verbal storage (like remembering someone's phone number by rehearsing the number in one's mind). These storage spaces are often referred to as the slave systems for the central executive. While the phonological loop and the visuo-spatial sketchpad are the basis for much experimental research, the central executive has proved much more difficult to empirically validate [1, 21]. From this point on, we use the term spatial WM to refer to the visuo-spatial sketchpad and verbal WM to refer to the phonological loop.

Much research suggests that verbal and spatial WM tasks involve different signatures of brain activation [6, 16, 21], supporting the theory that these storage systems are separate in the brain. For example, Gevins [6] used EEG to differentiate between four levels of brain activation: low verbal WM, high verbal WM, low spatial WM, and high spatial WM. A variety of experimental tasks have been created to study the differentiation of the slave systems. These tasks often manipulate which slave system is used and the level of memory load placed on the slave system (varying the number of items to store in WM, or the number of updates made to WM in a given time). For a review of verbal and spatial WM see [21] and [1] .

Leung and her colleagues [16] provide another set of WM tasks which increase the number of updates made to spatial WM in a given set of time. These tasks will play an important role in the experiment presented next. As shown in the first row of Fig. 3, the subject views screens in order (from left to right). After viewing a grid for a set of time, a dot is introduced in a random location in the grid. Next, a set of twelve screens (abbreviated in Fig. 3 with 4 screens) direct the subject to either keep the spatial location of the dot stored in WM (denoted with a '-' on the screen), or to update the location of the dot in WM based on the direction of the arrow (←,↑,→, or ↓). At the end of the task, the subject views a screen with a dot in a particular location, and (s)he indicates if the dot is in the correct location. Using this experiment setup, Leung used MRI to show that activation in the brain increased linearly as the number of updates in spatial WM increased [16].

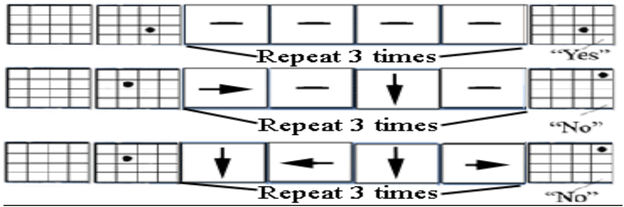

Figure 3: Leung's tasks, spatial WM is increased as we move from the top to the bottom task.

Workload in HCI Literature

HCI experts consider the various mental processes that make up workload when designing or evaluating an interface. For example, Boechler discussed the spatial cognition that is needed to navigate through web hierarchies. She urged web designers to understand the demands that spatial orientation places on web users, and she described the cognitive overload that occurs when a user "feels lost" while searching in hyperspace [2].

Larson and Czerwinski [14] looked at the connections between WM and the structure of multiple hyperlinks on web pages for information retrieval tasks. They tied their results into the memory and visual scanning ability of each subject, as assessed by tasks from the Kit of Factor Referenced Cognitive Tests [5]. In a later book chapter, Czerwinski and Larson [4] discussed the need to combine knowledge of "sensation, perception, attention, memory, and decision making[4]" from the cognition literature with experiments evaluating UIs. They noted that this is not an easy task, as most cognitive research focuses on low level tasks, with a small cognitive load. The jump from low level cognitive tasks to tasks in usability evaluation is large. They discuss bridging the gap between cognition experiments and UI evaluation [4]. The research presented in this paper is a step toward Czerwinski and Larson's goal.

Workload and fNIRs

Leung (Fig. 3) used fMRI to conclude that brain activity increased linearly as the number of updates to spatial WM increased [16]. This is in line with Parasuraman and Caggiano's [18] discussion on the neural activation of mental workload. They reviewed brain imaging research on workload, and found that WM and executive functioning tasks activate areas in the prefrontal cortex, and the amount of activation increases as a function of the number of items held in WM. The presence of workload activation and the relative load (of holding *n* items in WM or of making *n* updates to WM) can be quantified using PET, fMRI, fNIRs, and transcranial Doppler ultrasonography (TCD) [18]. This is promising, as fNIRs rates well when compared to the other brain imaging devices based on its functionality and practicality [11, 18].

NOVEL EXPERIMENTAL PROTOCOL

We designed a protocol to shed light on the workload experienced by users' various cognitive resources while working with a computer. We designed the protocol to aid usability experts to measure workload as a dependant variable while a user works with a UI and/or task. First we present the protocol in its most general form. Then we present our experiment and show how we used the protocol to measure the syntactic workload of our two simple UIs.

The general protocol is as follows: Given a UI to evaluate and an underlying task, we conduct a task analysis on the UI and task. For each subtask, we determine the cognitive subsystems that one would use while conducting the subtasks (i.e., spatial WM, visual search, etc.). Next we gather benchmark exercises from cognitive psychology designed to elicit high and low levels of workload on the target cognitive resource(s) associated with our UI.

Next, we run an experiment where users complete the benchmark cognition exercises, giving us a measure of their brain activity while they experience high and low workload levels in their various cognitive subsystems. Users also work with the UI that we are attempting to evaluate. Lastly, we use fNIRs data analysis tools to find similarities and differences in the users' brain activity while working with the UI to be evaluated and while completing the cognitive psychology exercises. While the protocol, in its most general form, will not yield exact measures of syntactic workload for any given UI, usability experts can incorporate the protocol into their studies and use the knowledge gained as an added usability metric. For example, web page designers using this protocol in a usability study might find that their users were visually overloaded while searching for items on a web page. They could determine this by finding that the users' brain activity while working on that particular web page was similar to the users' brain activity while conducting a cognition exercise designed to cause high visual search workload. In this case, the designers could re-design the page to place less demand on users' visual search resources.

In the future, one could imagine a *training period*, where users work with a set of benchmark cognitive psychology exercises designed to target particular cognitive resources (i.e., verbal WM, spatial WM, visual scanning, auditory processing). After determining the patterns of brain activity induced by the various benchmark exercises, users could work with a computer system and usability experts could search for similarities between the users' brain activity while working with the computer system, and the brain activity already established during the *training period*.

EXPERIMENT: UNCOVERING SYNTACTIC WORKLOAD

We designed an experiment to shed light on the syntactic (interface) components of workload. We created two interfaces to evaluate which were based on Leung's cognitive psychology tasks. Our interfaces allow users to traverse through hyperspace, which has been shown to involve spatial WM [2]. We chose a simple information retrieval (IR) underlying task that primarily uses verbal WM. We kept the underlying task difficulty level constant throughout all experimental conditions. We designed our UIs to map directly to users' spatial WM and our task to map directly to users' verbal WM. In this specially constructed scenario, if we could measure different levels of users' spatial WM demands while working with the UIs, we could acquire information about the syntactic workload of each UI. When a user worked with these UIs and simplified task, we say (s)he completed *interface exercises*.

We used our novel protocol and we conducted a task analysis on each hyperspace UI and IR task. We chose two exercises from cognitive psychology experiments that involve both spatial WM and verbal WM. Both of these exercises had low verbal WM demands (mirroring the IR task). However, one of these exercises had high spatial WM, and the other involved low spatial WM. When a user worked with our low level cognitive psychology tasks we say (s)he completed *cognition exercises*. We used fNIRs to record brain activity while subjects worked with our interface exercises and with our cognition exercises. By using benchmark exercises from the cognition literature with more realistic UIs and tasks, we attempted to bridge the gap between cognition experiments and UI evaluation.

Experiment Research Questions

We ran a pilot study and measured two subjects' brain activity while they completed four benchmark cognitive psychology tasks that induced only low spatial, only high spatial, only low verbal, and only high verbal WM demands. We were able to distinguish between these four conditions, and results were in line with cognition literature; that verbal WM and spatial WM use different resources in the brain, and we can measure high and low levels of workload in these different resources with fNIRs.

Pilot results indicated that we could, indeed, measure the different brain activity associated with spatial and verbal WM. Next, we aimed to tackle our primary research question: to determine whether or not we could use our experiment protocol and analysis algorithms to measure the syntactic workload of our two UIs. However, acquiring a reliable signal and measuring workload with fNIRs remains challenging. Therefore, we had two preliminary questions to address in order to reach our primary goal of measuring the syntactic workload of our UIs.

1) *Preliminary Question 1*: Can we use fNIRs to differentiate brain activity associated with each of our experimental conditions from brain activity at rest?
2) *Preliminary Question 2:* Can we use fNIRs to distinguish between no, low, and high demands on spatial WM?
3) *Primary Question:* Can we use well established exercises from the cognitive psychology literature on spatial WM to shed light on the syntactic workload involved in our higher level user interface exercises?

We hypothesize that we can answer 'yes' to all three research questions, which build on one another. In order to shed light on the syntactic workload involved in our high level interface exercises (question 3), we must distinguish between no, low, and high demands on spatial WM resources with fNIRs (question 2), and to do this, we must ensure that our fNIRs device can detect a brain signal induced by our conditions (question 1).

Experiment Description

We used a randomized block design and we randomly presented each of our conditions in nine trials throughout the experiment. The experiment had five conditions:

1) *Cognition exercises* that have been shown to cause *low spatial* WM load and low verbal WM load
2) *Cognition exercises* that have been shown to cause *high spatial* WM load and low verbal WM load
3) *Interface exercises* that show users their location in hyperspace while they search for verbal content. We hypothesize that this design will cause *low spatial* WM load
4) *Interface exercises* that do not show users their location in hyperspace while they search for verbal content. We hypothesize that this design will cause *high spatial* WM load
5) *Controlled rest exercises*

The first two conditions are depicted in Figure 4.

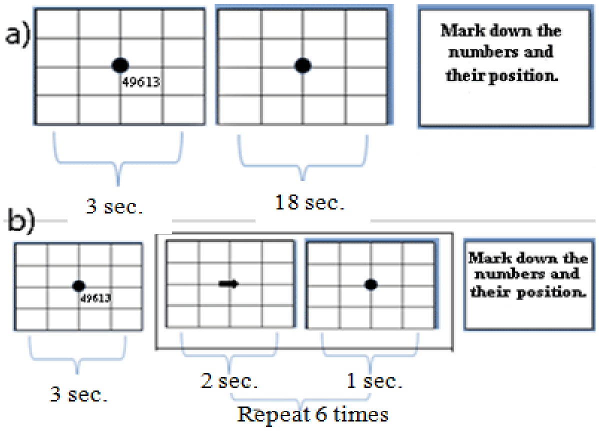
a)
49613
Mark down the numbers and their position.
3 sec.
18 sec.
b)
49613
Mark down the numbers and their position.
3 sec.
2 sec.
1 sec.
Repeat 6 times

Figure 4: a) *Low spatial WM cognitive psychology exercise*: recall digits and location b) *High spatial WM cognitive psychology exercise*: use arrows to update location of digits

The cognitive psychology tasks described previously by Leung [16] provide the basis for these exercises. During the first condition (Fig. 4a), subjects viewed a screen for 3 seconds which had a five digit number displayed in a spatial position in the grid. The grid had a circular fixation point in the center, on which the subjects were to keep their eyes focused. Fixation points are used to minimize eye movement and to maintain the onscreen stimuli position in memory. Subjects were instructed to recall the number and position of the cell that the number was located in. They kept the number (verbal) and position (spatial) in WM for 18 seconds, until prompted by the third screen to write down the numbers and position on a blank answer sheet.

In the second condition, participants saw a screen with a fixation point (Fig 4b). As in the first condition, subjects viewed a screen for 3 seconds with a 5 digit number located at a random spatial location in a grid. Next, they saw a screen for 2 seconds depicting an arrow (←,↑,→, or ↓) followed by a screen with just a fixation point for one second and they had to update the spatial location of the numbers based on the direction of the arrow. During this time they kept the number in their verbal WM and the current spatial position in spatial WM.

Variants of these two slides (an arrow followed by a fixation point) were repeated 5 more times, as indicated in the figure. For each of these exercises, there were six arrows (requiring WM updates) throughout the 21 second exercise span. As Leung showed previously, the second condition (Fig. 4b) requires more spatial WM updates than condition 1, and is associated with a higher level of spatial WM load. We refer to these conditions as the *low spatial* and the *high spatial* conditions.

The third and fourth conditions (Fig. 5) were placed into a more realistic, UI setting. These two conditions represent our interface exercises. We will attempt to measure the syntactic workload of each UI variation. A web hierarchy consisting of 36 web pages was created for each condition. Each web page had an image at the top of the page with a picture of the hyperspace. There was a five digit number on the bottom of each web page.

In condition 3 (Fig. 5a), referred to as the *display location* condition, subjects were randomly directed to a page in the hyperspace, and their position in the space was displayed for them. They were instructed to conduct a simple IR task; to remember a five-digit zip code (verbal WM) while searching through hyperspace for a matching zip code. Subjects used the arrows at the bottom of the web page to navigate. To ensure that subjects used similar WM processes in a similar span of time as the first two conditions, the hyperlink arrows were not active for the first three seconds after each web page was displayed.

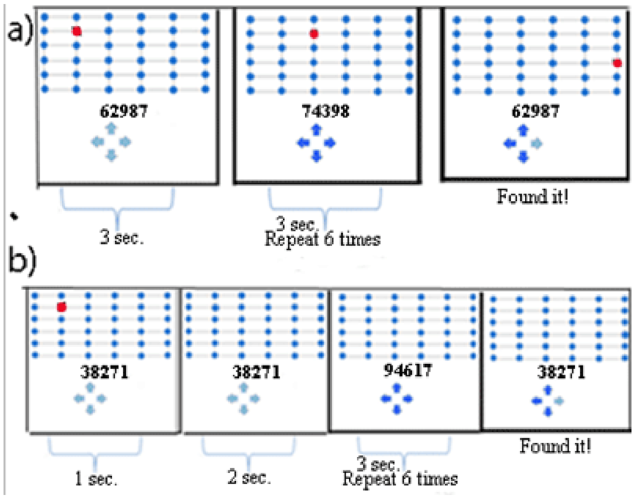

Figure 5: In the *interface exercises*, participants had to traverse hyperspace to find matching zip code, and remember their current location in hyperspace at all times. There were two conditions: a) Display location UI b) No location UI.

Arrows were initially light blue, depicting inactive links, and after three seconds passed, they became dark blue, indicating their active nature. As subjects traversed through the web space, the pictorial representation of the hyperspace displayed their current location at all times. While the zip code numbers on each page were randomly chosen, each exercise was set up so that a subject would not find a match until he or she reached an exercise length equal to the length of the cognition exercises. When they found a zip code match, subjects wrote down the zip code and its spatial location on an answer sheet.

Figure 5b depicts the fourth condition, which we refer to as the *no location* condition. The setup was the same as the third condition, except for one crucial change. When subjects were randomly directed to their starting webpage, they were shown their spatial location within the web hierarchy for one second (as shown in the first screen shot), and then their spatial location disappeared (second screen shot). The subjects had to navigate the web space to find a zip code match, and the arrow hyperlinks behaved the same as they did in the third condition, but the picture of the hyperspace on each page gave no information about the current page location. As subjects searched the hyperspace, they had to update their location in spatial WM while reciting the target zip code in verbal WM. When they found a zip code match, subjects wrote down the zip code and the location that they found the match at in the hyperspace.

In the fifth condition, subjects rested for 18 seconds.

Experiment Setup

Ten subjects completed the experiment. Six subjects were women and nine were college students ranging in age from 19 to 23 years old. One subject was a lecturer, aged 42 years. Nine of the subjects were right handed. There were nine trials in the experiment. Each trial consisted of each of the five conditions presented in random order. After writing down their answers, subjects rested for an additional 20 seconds to allow their brains to return to baseline. Subjects were asked to keep movement to a minimum and to keep their hands on the mouse during all conditions.

When describing our experiment data from this point on we use the term *task* to refer to an 18 second period of time that a subject was working with one of the five conditions described previously. We refer to a *trial* as one block of five tasks, where each condition was randomly presented once to the subject. Therefore, for each subject, there were five conditions tested and nine trials, resulting in 45 tasks.

fNIRs Equipment and Data Analysis

The fNIRs device is an ISS OxyplexTS frequency-domain tissue spectrometer with two probes. Each probe has a detector and four light sources. Each light source produces near infrared light at two wavelengths (690nm and 830nm) which are pulsed intermittently in time. This results in 2 probes x 4 light sources x 2 wavelengths = 16 light readings at each time point.

To analyze the data, we implemented several analysis algorithms that helped us make connections between the conditions in our experiment. With these algorithms, we found similarities and differences between our interface exercises and the cognition exercises that place high or low workload demands on subjects' cognitive resources.

Data Preprocessing

Each experiment lasted about 45 minutes, with data recorded every .16 seconds, resulting in approximately 16,875 data readings recorded throughout the experiment. Since we record 16 channel readings at each timepoint, our raw data is approximately 16,875 rows x 16 columns, and each column represents the readings of one source detector pair at one wavelength, which we refer to as one *channel*.

As brain activity differs widely on a person to person basis, we run all analyses separately for each subject. We first normalize the intensity data in each channel by their own baseline values. We then apply a moving average band pass filter to each channel and we use the modified Beer-Lambert Law [3, 10] to convert our light intensity data to measures of the change in oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (Hb) in the brain. Therefore, we have a recording of HbO at four depths on the left side (labeled L1, L2, L3, L4) and four depths on the right side of the brain (R1, R2, R3, R4). We have the same recordings of Hb data. We discard the data in the L1 and R1 channels. These channels pick up the shallowest level of activity, which consists mostly of physiological noise from the forehead (sweat, movement, etc.) [25]. We also shift each of our tasks so that the ΔHbO and ΔHb data begins at 0. We now have six time series depicting the ΔHbO and six time series depicting the ΔHb in the brain activity throughout the experiment. Figure 6 shows ΔHbO on the left and right side of the brain during one task of 18 seconds for subject 4. In this task, the subject was completing the *display location* interface exercise. Next we discard the rest time between tasks (including time when subjects recorded answers). Lastly, we cut off two seconds of data from the start of each task, as blood in the brain takes a few seconds to reach its area of activation. Since some of the interface exercises lasted longer than other conditions, we truncate each task to the length of the shortest task. This results in each task lasting ~18 seconds.

Folding Average Analysis: ANOVA

We use Analysis of Variance (ANOVA) to determine whether or not our experimental conditions cause different patterns of brain activation. We follow the same process used by Izzetoglu [10] to compare various levels of mental workload encountered while users worked with a mock command and control center application by looking only at HbO data [11]. For each like condition we conduct folding averages across all trials to remove noise from the data and to acquire a template of the average HbO recorded for each subject during that condition. This results in one prototype of each condition for our six HbO channels. We then average together the HbO channels on the left side of the brain by averaging together L2, L3, and L4 at each of their time *t =1, t=2… t=n*, where *n* equals the last time point in the ~18 second long condition. We do the same for the right side of the brain. For each condition, we have two time series representing the activity on the left and on the right side of the brain. We use ANOVA with a confidence level of 95%, to determine whether or not each condition elicits different brain activity than the other conditions.

Folding Average Analysis: Clustering

We also implemented a hierarchical clustering algorithm to draw similarities between our various conditions. In particular, we use this algorithm to find similarities between the known cognition exercises and our interface exercises in order to gain information about the syntactic workload of each UI. When clustering fNIRs data, we must be mindful that irrelevant data can be detrimental to the performance of the clustering algorithm [24]. If, for example, a condition induces no activation on the left side of the brain, time series from that side of the brain are irrelevant. We also note differences in the use of HbO and Hb data in existing fNIRs research. Some use HbO data only [10] while others use both Hb and HbO data [18, 22] for analysis. There are also cases when the Hb is the most relevant [19]. Therefore, we select the most relevant channels before clustering. We randomly choose one trial to remove from the nine available trials. We run an ANOVA comparing the statistical difference between these five conditions on HbO channels L2, L3, and L4 and R2,R3,R4. We do the same for the Hb data. For each subject, we choose the two HbO or Hb channels that result in the best F-statistic differentiating between the conditions.

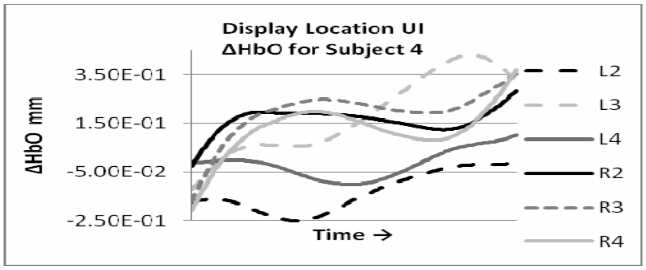

Figure 6: ΔHbO in subject 4's brain while completing one 18 second long *display location* UI exercise. L2, corresponds to the left side of the head, channel 2 on the fNIRs device.

Once we've determined the most relevant channels we conduct a folding average of each condition over the eight remaining trials for our two channels. This gives us a prototype of each of our five conditions for our two best channels. We concatenate the time series for these two channels together, and we run hierarchical clustering, with an unweighted average Euclidian distance similarity metric.

RESULTS AND ANALYSIS

We discuss our results within the context of each of our research questions. In this section we use abbreviations for our *low spatial* (LS), *high spatial* (HS), *display location* (DL), *no location* (NL) and *no workload* (0WL) conditions.

1) *Preliminary Question 1:* Can we use fNIRs to differentiate brain activity associated with each of our experimental conditions from brain activity at rest?

To answer the first question, we look for a difference in subjects' brain activity while resting and while completing each of our experiment conditions. We run ANOVAs comparing the brain activity of each of our cognitive psychology and UI conditions with our *no workload* condition. Results are in Table 1a. The Table shows the ANOVA comparison between each condition, for each subject on the left and right side of the brain. It is possible that each condition caused activity on both or on only one side of the brain per subject. The presence of a '√' indicates that ANOVA showed the conditions were significantly different, with confidence of 95%. Out of 40 pair-wise comparisons in Table 1a, only one comparison had no significant differences (both the right *and* left side of the brain showed insignificant results). These results indicate that we successfully addressed our first question. Next we tackle our second question.

a)	NL v. 0WL	DL v. 0WL	LS v 0WL	HS v 0WL
	L R	L R	L R	L R
s1	√√	√√	√√	√√
s2	√√	√√	√√	√√
s3	√√	√√	√√	√√
s4	√√	√√	√√	√√
s5	√√	√√	√√	√√
s6	√√	√√	√√	√√
s7	√√	√√	√√	√√
s8	√		√√	√
s9	√√	√√	√	√√
s10	√√	√√	√√	√

b)	LS v 0WL	HS v 0WL	LS v HS
	L R	L R	L R
s1	√√	√√	√√
s2	√√	√√	√√
s3	√√	√√	√√
s4	√√	√√	√
s5	√√	√√	√√
s6	√√	√√	√
s7	√√	√√	√√
s8	√√	√	√√
s9	√	√√	√√
s10	√√	√	√√

Table 1: ANOVA results for the left and right side of the brain. A '√' denotes that the two conditions were different with confidence of 95%. s1 = subject 1, etc. a) and b) address the first and second research questions, respectively.

2) *Preliminary Question 2:* Can we use fNIRs to distinguish between no, low, and high demands on spatial WM?

To answer our second question, we use our ANOVA results (Table 1b) to do pair-wise comparisons of the *LS, HS*, and *no workload* exercises. As Table 1b shows, we can distinguish between these conditions on one side (often both) for all subjects. Therefore, we can distinguish between no, low, and high WM demands in users' brains. Now we tackle our primary question.

3) *Primary Question:* Can we use well established exercises from the cognitive psychology literature on spatial WM to shed light on the syntactic workload involved in our higher level user interface exercises?

Clustering the data provides insights into our primary research question. Clustering results for each subject are in Figure 7, and Table 2 gives an overview of these results. Based on our success addressing our first two research questions and our use of established exercises from cognitive psychology, we assume that our *no workload*, *LS*, and *HS* exercises provide us with benchmark levels of spatial WM, where:

- The spatial WM demands of the *no workload* exercises are less than the demands of the *LS* exercises, which are less than the demands of the *HS* exercises (spatial WM load of 0WL < LS < HS).

We look at the similarities between our interface exercises and our benchmark spatial WM exercises to draw conclusions about the WM demands of the *NL* and *DL* UIs. We expect the *NL* interface exercises to cause higher spatial WM load than the *DL* interface exercises. Thus, when comparing the brain activity of the *NL* interface exercises with the *DL* interface exercises, we expect the clusters to indicate that the *NL* interface exercises are grouped closer to cognition exercises of known higher spatial WM than the *DL* interface exercises (where spatial WM of 0WL < LS < HS).

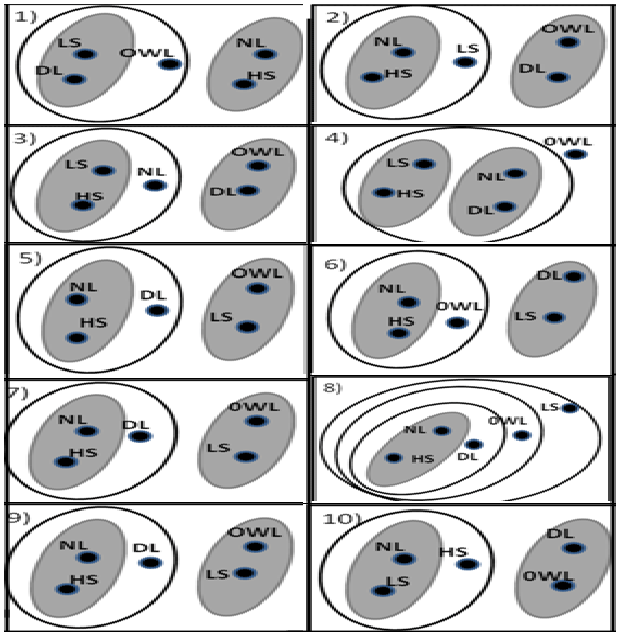

Figure 7: Clustering for each subject. (1 = subject 1, etc.)

As the first row of Table 3 indicates, this was the case for 90% of our subjects. In fact, the *NL* interface exercises were the *most* similar to the known *HS* cognition exercises for 70% of subjects (row 2, Table 2). Cluster results show that we successfully addressed our primary research question for 90% of our subjects. The spatial WM involved in the *NL* UI was higher than the spatial WM needed to work with the *DL* UI, indicating that the *NL* condition had higher syntactic workload (in this case, spatial WM) than the *DL* condition. This makes sense, as UI designers know the benefits of keeping users oriented in hyperspace.

Findings From Clustering	% of subjects	subject ID #'s
The *NL* UI exercises are grouped more closely with exercises of known higher spatial WM than the *DL* UI exercises (all cluster levels).	90%	1, 2, 3, 5, 6, 7, 8, 9, 10
The *NL* UI exercises are clustered together with known *HS* cognition exercises (1st level cluster).	70%	1, 2, 5, 6, 7, 8, 9

Table 2: Analysis of cluster results (3rd research question)

Extension to Adaptive interfaces

Using fNIRs data as an additional metric for usability testing allows us to conduct folding averages across trials to remove noise and make generalizations about the patterns of activation of our experiment conditions. We also foresee brain measurement in the future as an input for adaptive systems. To create adaptive fNIRs based systems, we need to classify brain patterns on a single trial basis, allowing the system to adapt in real time. With this goal in mind, we implemented a classifier to test our ability to distinguish between our conditions on a single trial basis.

We implemented a weighted k-nearest-neighbor classifier with a Dynamic Time Warping distance metric [12].

For each of the 45 experiment tasks, we averaged together the channel readings on the left side (L2,L3,L4) and on the right side (R2,R3,R4) of the brain. The result was four time series representing HbO on the left side of the brain (HbO_L), Hb on the left side of the brain (Hb_L), HbO on the right side of the brain (HbO_R), and Hb on the right side of the brain (Hb_R). We removed one of our nine trials of data in order to select the most relevant of the HbO_R, HbO_L, Hb_R, and Hb_L time series for classification of the eight remaining trials. Using the data from our removed trial, we ran ANOVA on each of the HbO_L, HbO_R, Hb_L, and Hb_R and we chose the time series that showed statistically significant differences between the two tasks we wanted to classify. We used these time series to classify the remaining trials. We did this for each trial, switching the trial used to choose the relevant channels.

Single Trial Analysis Results

We used our KNN classifier with k = 3, (Table 3) to make comparisons between brain activity induced by our conditions on a single trial basis. Results in 3a reflect research question 1, comparing each condition to the *no workload* conditions. Results in 3b reflect our second research question, where we distinguish between our *no workload*, *LS,* and *HS* exercises and between our *LS* and *HS* exercises. Table 3c shows our accuracy at classifying our interface exercises. The last row of Table 3 shows average accuracy for each comparison across subjects.

a)	NL v 0WL	DL v 0WL	LS v 0WL	HS v 0WL	b)	LS v HS	LS v HS v 0WL	c)	DL v NL
S1	67%	87%	81%	78%	S1	*78%*	*63%*	S1	64%
S2	67%	77%	89%	67%	S2	*87%*	*62%*	S2	72%
S3	86%	58%	83%	78%	S3	*74%*	*63%*	S3	70%
S4	100%	93%	83%	88%	S4	*56%*	*48%*	S4	64%
S5	74%	64%	64%	65%	S5	*61%*	*47%*	S5	64%
S6	77%	79%	81%	66%	S6	*69%*	*52%*	S6	57%
S7	100%	91%	70%	91%	S7	*78%*	*64%*	S7	62%
S8	72%	58%	71%	88%	S8	*40%*	*33%*	S8	83%
S9	91%	94%	72%	95%	S9	*78%*	*60%*	S9	72%
S10	56%	48%	92%	61%	S10	*61%*	*48%*	S10	67%
avg	*79%*	*75%*	*79%*	*78%*	*avg*	*68%*	*54%*	*avg*	*68%*

Table 3: Percentage of instances classified correctly when comparing various conditions. s1 = subject 1, etc. a) and b) address the first and second research questions, respectively. c) compares the *DL* and *NL* conditions.

We see variation between subjects for these comparisons. However, average results are promising. We can distinguish between low and high spatial WM load with nearly 70% average accuracy (3b), between each condition and the 0WL condition with nearly 80% (3a) average accuracy, and between our two interface exercises with 68% average accuracy (3c). When distinguishing between the three classes of LS, HS, and 0WL (3b), we see that subject 8 had accuracy no better than random (33%), but the other subjects' comparisons show promise with respect to differentiating between these three classes.

CONCLUSION

We presented a novel experiment protocol and a set of analysis algorithms that can help UI evaluators, or designers of adaptive systems, to gain information about the workload experienced by users in the various cognitive resources in their brains while they work with computer systems. We attempted to push workload measurement out of the lab, where workload is manipulated as an independent variable, and into the realm of UI evaluation, where users' workload is a dependant variable, changing in unknown ways based on the UI and task.

We also designed two simplified UIs and a task that were intended to map directly to users' spatial and verbal WM, respectively. We ran an experiment to acquire quantitative, real time measures of the syntactic (i.e., spatial) workload of our UIs. The experiment tied Shneiderman's theory on syntactic and semantic workload to quantifiable brain measurements. We believe that this is the first case of separating syntactic and semantic workload using fNIRs, even though we use a specially constructed interface.

We chose our UIs and the IR task to show how our experiment protocol and analysis procedures can be used to measure the syntactic workload of our UIs. These tasks allowed us to target separate cognitive resources and to change the syntactic workload (i.e.,spatial WM) while keeping the semantic workload (i.e., verbal WM) constant, but most UIs and tasks are more complex, overlapping in the cognitive resources they require. While separating semantic and syntactic workload may not be possible in more complex UIs, evaluators can make informed changes to UI designs based on the level of workload measured in users' various cognitive resources. In more complex UIs, our novel protocol will work in a similar manner. Although the established cognitive psychology tasks will not parallel our UIs as closely as they did in this experiment, the fNIRs analysis algorithms will still show similarities and differences between brain activity induced by the UIs and by the cognition exercises.

Also, the experiment protocol and single trial analysis used in this experiment can serve as a step toward adaptive UIs. One can picture a future adaptive system that trains users on a set of known cognition tasks that elicit no, low, and high levels of workload on various cognitive resources. Then the UI can adapt appropriately based on the user's current workload in each resource.

ACKNOWLEDGMENTS

We acknowledge the contributions of Michel Beaudouin-Lafon, Wendy Mackay, and the HCI group at Universite Paris-Sud. We thank the National Science Foundation (grant nos. IIS-0713506 and IIS-0414389), NSERC, and AFOSR for support of this research Any opinions, findings, and conclusions or recommendations expressed in

this article are those of the authors and do not necessarily reflect the views of NSF, AFOSR, or NSERC.

REFERENCES

1. Baddeley, A. and Della Sala, S. Working memory and executive control. *Philosophical Transactions of the Royal Society of London*, *351*, 1996.
2. Boechler, P. How Spatial Is Hyperspace? Interacting with Hypertext Documents: Cognitive Processes and Concepts. *CyberPsychology and Behavior*, *4* (1), 2001.
3. Chance, B., et al. A novel method for fast imaging of brain function, non-invasively, with light. *Optics Express*, *10* (2). 411-423, 1988.
4. Czerwinski, M. and Larson, K. Cognition and the Web: Moving from Theory to Web Design. in *Human Factors and Web Development*, Ratner, J. (Ed.), Erlbaum: NJ, 2002, 147-165.
5. Eckstrom, R., French, J., Harman, H. and Derman, D. Kit of factor-referenced cognitive tests. 1976.
6. Gevins, A., Smith, M., McEvoy, L. and Yu, D. High-Resolution EEG Mapping of Cortical Activation Related to Working Memory: Effects of Task Difficulty, Type of Processing, and Practice. *Cerebral Cortex*, 1997.
7. Gratton, G., Fabiani, M., Friedman, D., Franceschini, M., Fantini, S., Corballis, P. and Gratton, E. Rapid Changes of Optical Parameters in the Human Brain During a Tapping Task. *Journal of Cognitive Neuroscience*, *7*. 446-456, 1995.
8. Grimes, D., Tan, D., Hudson, S., Shenoy, P. and Rao, R., Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph. in *CHI Conference on Human Factors in Computing Systems*, (2008).
9. Hart, S.G. and Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theorical research. in Hancock, P., Meshkati, N. ed. *Human Mental Workload*, Amsterdam, 1988, pp 139 - 183.
10. Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K. and Chance, B. Functional Optical Brain Imaging Using Near-Infrared During Cognitive Tasks. *International Journal of Human-Computer Interaction*, *17* (2). 211-231, 2004.
11. John, M.S., Kobus, D., Morrison, J. and Schmorrow, D. Overview of the DARPA Augmented Cognition Technical Integration Experiment. *International Journal of Human-Computer Interaction*, *17* (2). 131-149, 2004.
12. Keogh, E. and Pazzani, M., Scaling up dynamic time warping for datamining applications. in *Proc. of the Sixth ACM SIGKDD*, (2000).
13. Kohlmorgen, J., et al. Improving Human Performance in a Real Operating Environment through Real-Time Menatal Workload Detection. in *Toward Brain-Computer Interfacing*, MIT Press, 2007, 409-422.
14. Larson, K. and Czerwinski, M., Web Page Design: Implications of Memory, Structure and Scent for Information Retrieval. in *Proc. of the SIGCHI Conference*, (1998).
15. Lee, J.C. and Tan, D.S., Using a Low-Cost Electroencephalograph for Task Classification in HCI Research. in *ACM Symposium on User Interface Software and Technology*, (2006).
16. Leung, H., Oh, H., Ferri, J. and Yi, Y. Load Response Functions in the Human Spatial Working Memory Circuit During Location Memory Updating. *NeuroImage*, *35*, 2007.
17. Muller, K.T., M., Dornhege, G., Krauledat, M., Curio, G. and Blankertz, B. Machine learning for real-time single-trial EEG-analysis: From Brain-computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, *167* (1). 82-90, 2008.
18. Parasuraman, R. and Caggiano, D. Neural and Genetic Assays of Human Mental Workload. in *Quantifying Human Information Processing*, Lexington Books, 2005.
19. Sassaroli, A., Zheng, F., Hirshfield, L.M., Girouard, A., Solovey, E.T., Jacob, R.J.K. and Fantini, S. Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy. *in the Journal of Innovative Optical Health Sciences*, 2009.
20. Shneiderman, B. and Plaisant, C. Designing the User Interface: Strategies for Effective Human-Computer Interaction, Fourth Edition, Addison-Wesley, Reading, Mass., 2005.
21. Smith, E. and Jonides, J. Storage and Executive Processes n the Frontal Lobes. *Science*, *283*, 1999.
22. Son, I.-Y., Guhe, M., Gray, W., Yazici, B. and Schoelles, M. Human performance assessment using fNIR. *Proceedings of SPIE The International Society for Optical Engineering*, *5797*. 158–169, 2005.
23. Wickens, C., Lee, J., Liu, Y. and Becker, S. *An Introduction to Human Factors Engineering*. Pearson, 2004.
24. Witten, I.H. and Frank, E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.
25. Zhang, Q., Brown, E. and Strangman, G. Adaptive filtering for global interference cancellation and real-time recovery of evoked brain activity: a Monte Carlo simulation study. *Journal of biomedical optics*, *12*, 2007.