This is Your Brain on Interfaces: Enhancing Usability Testing with Functional Near-Infrared Spectroscopy

Leanne M. Hirshfield[1] **Rebecca Gulotta**[2] **Stuart Hirshfield**[1] **Sam Hincks**[1] **Matthew Russell**[1] **Rachel Ward**[1] **Tom Williams**[1] **Robert Jacob**[3]

[1]Computer Science
Hamilton College
Clinton, NY 13323

[2] HCI Institute
Carnegie Mellon University
Pittsburgh, PA 15213

[3]HCI Laboratory
Tufts University
Medford, MA 02155

ABSTRACT

This project represents a first step towards bridging the gap between HCI and cognition research. Using functional near-infrared spectroscopy (fNIRS), we introduce techniques to non-invasively measure a range of cognitive workload states that have implications to HCI research, most directly usability testing. We present a set of usability experiments that illustrates how fNIRS brain measurement provides information about the cognitive demands placed on computer users by different interface designs.

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

Author Keywords: usability, brain, mental workload, fNIRS, brain-computer interaction.

General Terms
human factors, experimentation

INTRODUCTION

Usability researchers attempt to formalize and quantify the process whereby an interface is evaluated, and to measure precisely the degree to which an interface meets the goals of its intended audience. Although one can measure the accuracy with which users complete tasks and the time it takes to complete a task with a user interface (UI), measuring subjective factors such as workload, frustration, and enjoyment is more difficult. These factors are often "measured" by qualitative observation of subjects or by administering subjective surveys to subjects. Such surveys are inherently subjective and they can elicit participant biases, as participants often attempt to please experiment investigators in their responses. Additionally, surveys are often administered after a task has been completed, lacking insight into the users' changing experience as they work with a UI.

Our research addresses these evaluation challenges with respect to mental workload (WL). We use a non-invasive brain sensing technique called functional near infrared spectroscopy (fNIRS) to record real time, objective measurements of users' WL while working with UIs. Users can wear the comfortable fNIRS device (Fig. 1) in working conditions.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.
Copyright 2011 ACM 978-1-4503-0267-8/11/05....$10.00.

Using brain measurement to quantify the level of WL experienced by computer users is a difficult task because "workload" is somewhat of an umbrella term. The brain is a complex structure, and there are many cognitive resources that work in serial and in parallel to process information. Indeed, when we compute arithmetic, compose a poem, or chat with a friend, we are experiencing some form of WL.

Figure 1: A participant wears the comfortable fNIRS.

However, for each task that we take part in, we may use different (and often overlapping) cognitive resources. While there are many definitions in the literature describing WL [13, 14, 33], for the purposes of this paper we define it in functional terms as: *The load placed on various cognitive resources in the brain in order to complete a task that involves the processing of information.*

Since the processing of information is a key element in users' interactions with computers, the field of human-computer interaction (HCI) is derived from and heavily influenced by cognitive psychology. We seek a thorough understanding of the effects that a new UI will have on users' mental resources. Ideally, a UI will be easy to use, allowing users to focus their mental resources on the task at hand. However, there remains a large gap between the high-level references made to these mental resources in HCI research and the low-level ability to pin point and measure these resources in human users. For this reason, Czerwinski and Larson [9] discuss the need to tie together cognition research and HCI design principles. They note that this is not an easy task, as most cognitive research focuses on task manipulations that are on a low-level, with a

cognitive load that is small, and the jump from these specific cognitive tasks to UI tasks administered in real world settings is large.

With these challenges in mind, our work provides two primary research contributions to the field of HCI:

1) We attempt to bridge the gap between HCI and cognition research. We introduce techniques to measure via non-invasive means a range of cognitive WL states that have implications to HCI research. We describe several cognitive resources at a low-level, as they pertain to the brain, and at a high-level, as they relate to the field of HCI.

2) We demonstrate ways that fNIRS brain measurement can be used to complement usability testing by measuring a range of WL states objectively and in real time. This additional information can yield a more thorough, comprehensive understanding about a UI design than can be achieved with standard usability testing.

We begin this paper by describing related work on the fNIRS device and the human brain and go on to describe an experiment protocol designed to measure the level of load placed on users' low-level cognitive resources while working with a UI. We then present a series of usability experiments that apply this protocol to the evaluation of design choices of UIs. Finally, we describe our results and analysis, as well as avenues for future work in this area.

RELATED WORK

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

Measuring the Human Brain with fNIRS

EEG has been used in the HCI and Human Factors domains to measure various aspects of mental workload [4, 12, 24, 28, 34]. While a promising tool for non-invasive brain measurements, EEG has several drawbacks such as low spatial resolution, susceptibility to noise, and long set-up time which can make EEG challenging to use in realistic human-computer interactions. fNIRS has recently been introduced [8] to overcome many of the drawbacks of the EEG and other brain monitoring techniques. 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, as depicted in Figure 2. Deoxygenated (Hb) and oxygenated hemoglobin (HbO) are the main absorbers of near infrared light in tissues during hemodynamic and metabolic changes associated with neural activity in the brain [8]. These changes can be detected measuring the reflected light that has probed the brain cortex [8, 19].

Experimental Studies of Workload in the Brain

To date, most experimental psychology studies explore brain functioning while people conduct simple, highly controlled tasks that have been designed to target a specific cognitive resource, such as visual scanning and perception, working memory (WM), and similar higher-order cognitive functions. Many of these resources relate directly to current research in HCI.

Working Memory

WM refers to an information processing system that temporarily stores information in order to serve higher-order cognitive functions such as planning, problem solving, and understanding language [3]. One of the most common tasks used in cognitive psychology to elicit load on participants' WM resources is the n-back task [11, 23, 30]. The n-back task is depicted in Figure 3.

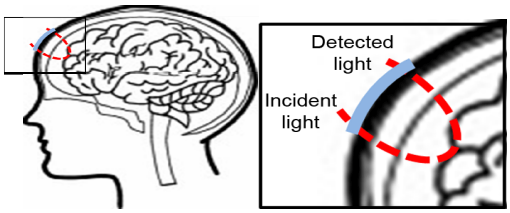

Figure 2: Light in the near-infrared range is pulsed into the brain cortex and the reflected light is detected.

A series of letters is presented to participants, one letter at a time, on the computer screen. For each letter, the participant indicates whether or not that letter matches the letter that she saw *n* letters previously. During the 3-back condition, participants must store and manipulate, three items at a time in WM, and for the 1-back task, they only manipulate and store one item at a time.

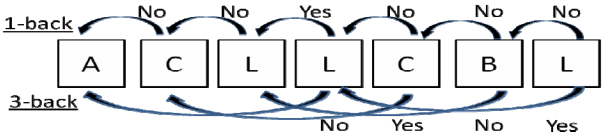

Figure 3: Depiction of the 0-back and 3-back task.

Parasuraman [27] found that WM and executive functioning tasks activate areas in the prefrontal cortex, and that the amount of activation increases as a function of the number of items held in WM. The presence of WL activation and the relative load (of holding *n* items in WM or of making *n* updates to WM) can be quantified using fNIRS [27]. WM is at the foundation of all interactions between humans and computers. If we can determine that a UI design elicits high WM load on users, we can modify the UI to alleviate this load. For example, a poorly constructed menu may require users to keep several menu items in their WM while searching for the correct selection. Perhaps this menu can be redesigned to alleviate this load on the user's WM.

Visual Perception and Search

Another WL related set of tasks involves visually searching for items within a set of distracter items. There are two kinds of visual search; efficient and inefficient. If a target item is saliently different as compared to the distracter items surrounding the target item, it can be found immediately, regardless of the number of distracter items [2, 26, 32]. Inefficient search occurs when the search item is not highly salient as compared to the distracter items. In this

case, a serial visual search must be performed in order to locate the item of interest.

One task that has been used to induce inefficient visual search is called the "finding A's" task. This task can be found, among a variety of other experimental tasks, in the Kit of Factor Referenced Cognitive Tests [10]. In this task, users are instructed to scan through a list of words and to cross off every word that contains the letter 'a'.

fNIRS has been used to measure the resources responsible for visual search [22, 25]. If we can determine that users are conducting inefficient visual searches while looking for items in a UI, we can modify the UI to alleviate these demands. For example, a poorly structured web page may not direct users' visual attention to the relevant content that they are likely to search for. The page can be restructured to provide salient visual cues directing people to the most relevant semantic content on the web page.

Executive Processes and Response Conflict

Pinpointing functional brain regions is difficult to do with our executive processes, as they are involved in high order processing that involves the recruitment of a number of overlapping cognitive resources [3]. While we still have much to learn about executive processing, functions related to response conflict have been empirically validated. Response conflict deals with the central executive's suppression of automatic, but incorrect responses. We use response conflict throughout our daily lives. For example, while driving a car we may see a squirrel run across the road. Our initial, automatic reaction may be to swerve the car away from the squirrel. However, a quick look around may show oncoming traffic on one side of us and a cliff on the other side. Our central executive helps us to inhibit the automatic response of swerving in order to choose the response of staying in our lane (sorry, squirrel!).

A common test used in experimental psychology research to induce response inhibition in the brain is the Stroop test [29]. In this task, a color name, which is written in a font of a particular color, is presented to participants. Participants must say the color that each word is written in out loud. In the congruent condition (Figure 4a), the name of the word and the color that the word is written in are one and the same. In the incongruent condition (4b), the name of the word and the color that the word is written in are different. People's ability to name a color is slightly slower than their semantic ability to read a word. Thus, the incongruent condition of the Stroop test requires people to use their response conflict resources, suppressing the automatic response of saying the name of the word and answering correctly, with the color of the word. fNIRS has been used to measure the brain activity induced by the Stroop test [29].

If we determine that a user has high response conflict while working with a UI, this may indicate that something about the UI is unintuitive. For example, some un-intuitive video games may require the user to press awkward keys in order to navigate in the game; pressing the 'A' key when one wants to turn right may cause response conflict as the user may automatically want to physically move the controller to the right, as is done with the successful Nintendo Wii.

a) BLUE
RED

b) YELLOW
GREEN

Figure 4: Say the color that each word is written in out loud. a) congruent Stroop test, and b) incongruent Stroop test.

TRANSITIONING FROM LOW-LEVEL EXPERIMENTAL PSYCHOLOGY TO HIGH-LEVEL USABILITY TESTING

In order to enhance usability testing with fNIRS, we aim to measure which low-level resource(s) are being taxed and the level of load (i.e., high or low) that is placed on each of these resource(s) while users work with UIs. Although there is some disagreement as to the effects of multitasking in the brain [1, 20], most agree that combining several low-level, simplified tasks has an additive effect in the brain. fNIRS has been used to measure the spatiotemporal changes occurring when different cognitive resources are recruited to complete a high-level mental task [17, 18]. This suggests that we can use fNIRs to measure different patterns of activation associated with the various cognitive resources targeted in this experiment.

The brain research presented in this section shows that: 1) There are unique signatures of brain activation relating to the level of load placed on many low-level cognitive resources. 2) Brain activation increases in a given region as the load on that resource increases. 3) Combining several low-level, simplified tasks has an additive effect in the brain, which can be measured with fNIRS. Thus, we expect to be able to measure different levels of load placed on users' WM, visual search, and response inhibition resources.

CONDUCTING USABILITY EXPERIMENTS WITH FNIRS

To make connections between the users' brain activity while completing high-level UI tasks and the low-level demands placed on their cognitive resources, we developed an experiment protocol, described in detail in [16], for use in our usability tests. The general protocol is as follows:

Usability Experiment Protocol

1) Researchers gather benchmark tasks from cognitive psychology that elicit high and low-levels of WL on a range of target cognitive resource(s) such as visual search, WM, and response inhibition. We refer to these exercises as *cognitive benchmark tasks.*

2) Researchers create a set of tasks that users will complete with the UI to be evaluated. We refer to these as *UI tasks.*

3) An experiment is run where users complete the *cognitive benchmark tasks*, yielding a measure of their brain activity while experiencing high and low WL levels in their various cognitive subsystems. Users also work with the *UI tasks*. Brain activity is measured throughout the experiment.

4) fNIRS data from the *cognitive benchmark tasks* are used as training data to build a machine learning classifier. The fNIRS data from the *UI tasks* is input into the machine learning classifier as testing data for the classifier. The classifier outputs the level of cognitive load experienced by each user while working with a given *UI task*.

USABILITY EXPERIMENTS

We conducted a usability experiment to evaluate the design of two UIs and to demonstrate our experimental protocol.

Cognitive Benchmark Tasks

Participants completed the following benchmark tasks:

Finding A's

In the Finding A's task, participants were instructed to look at a matrix of words and click on those that contained the letter 'a'. In one version of the task, the a's in the words were highlighted. In the other version, the a's were not highlighted. There were 14 words with a's in them on each screen presented to participants. If a participant clicked on all of the words containing a's, a new screen was shown, though there was a maximum of two screens per 50 second task. These two tasks induced benchmark levels of high and low load on users' visual search resources.

Stroop

The Stroop task had two variations, an incongruent Stroop and a congruent Stroop, which are depicted in Figure 5. In both variations, adapted from previous fNIRS research [29], one of the words BLUE, RED, YELLOW, or GREEN appeared on the screen for .5 seconds before another of those four words was added to the screen beneath the first. Participants determined whether the bottom word correctly described the color of the top word. In the congruent version, the top word was always colored to correspond with its meaning. In the incongruent version, the top word did not have to match its own meaning. These tasks induced benchmark levels of high and low load on users' response inhibition resources.

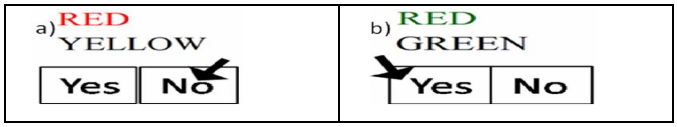

Figure 5: a) congruent Stroop and b) incongruent Stroop task

n-back

We used the 0-back and 3-back variations of the n-back task. In the 0-back condition, participants were first shown a single letter. As the task progressed, participants were presented with new letters on the screen. They were asked to identify whether the letters they were seeing were the same or different than the first letter. In the 3-back task, they were asked to judge whether the letter they were looking at was the same as the letter they saw three letters before. These two tasks induced benchmark levels of high and low load on users' WM resources.

UI Tasks

The first UI was a driving simulator and the second UI was for conducting web searches. We chose these UIs because the design of driving/video game [6, 21] and web search UIs [5, 32] have been popular areas of HCI research.

Car Driving Interface

The driving tasks were completed using a software program called 3D Driving School (Fig. 6). The task took place in a parking lot where there were three lines of cones. Participants were instructed to drive around the cones in a slalom pattern using the arrow keys on the keyboard to navigate. There were two variations of the task. In the first, the commands associated with the arrow keys performed as expected. That is, pressing the ← and → arrow keys caused the car to turn to the left and the right, respectively. In the second variation, the functions of the ← and → arrow keys were reversed. Pressing the ← key turned the car to the right, and pressing the → key turned the car to the left.

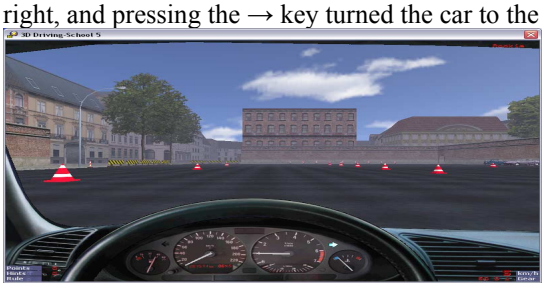

Figure 6: The Free Drive course used in the driving tasks.

Web Search Interface

The web search was completed using a Google search engine embedded in an online site for financial news articles (www.fdi-magazine.com). Participants were presented with a question which could be answered by searching the financial site and reading the articles (Fig. 7).

Figure 7: The web search task with highlighting.

Participants were instructed to highlight their answer using the mouse and then to proceed to the next question (this eliminated noise caused by participants speaking or writing their answers in the middle of the task). For each instance of the web search, there was a maximum of two questions. There were two versions of the web search task. In one version, the user's search terms were highlighted in the search results and the articles. In the other version there was no highlighting of search terms.

Completion Time

We controlled for completion time in all experiment conditions. During the nback and Stroop conditions, new questions were presented to subjects every 3 seconds. There was only one speed enabled when subjects moved forward during the driving tasks, placing a limit on the number of cones users could access in each task. During the finding a's and web search tasks, subjects only had a set number of task-items to complete in each 50-second period of time. Subjects were instructed (with the help of a progress bar that was visible during the experiment) to work at a rate that would result in the completion of all tasks by the end of the task period. By controlling for time in this way, we were able to focus more on the difference in brain activity during different tasks rather than on differences caused by working at different paces.

Usability Experiment Setup and Protocol

In the rest of this paper, we use the following terminology:

Cognitive benchmark tasks: Experiment tasks that have been pulled from experimental psychology research. In this experiment, these tasks target the following resources*: low and high WM, low and high visual search, and low and high response inhibition tasks.*

UI tasks: Experiment tasks that represent the UIs to be evaluated. These include: *driving with correct mapping, driving with incorrect mapping, web search with highlighting,* and *web search with no highlighting tasks*.

Methodology

Ten participants completed the experiment. After providing informed consent, participants were instructed to complete a tutorial. Participants were instructed to keep body movement to a minimum, and to only move when using the keyboard and mouse while working with the experiment tasks. Previous research has shown that these minimal hand movements do not add detrimental amounts of noise to the fNIRS data [31]. Experiment tasks were presented to participants in a randomized order, with a 23 second rest period between tasks. There were six trials and each trial consisted of 10 tasks (the six benchmark tasks and the four UI tasks).. Each task lasted 50 seconds. As the experiment progressed, the answers provided by the participants were recorded. After the experiment was complete, participants completed a post-experiment survey where they rated the tasks on a 1-7 Likert scale, with 1 representing the lowest and 7 representing the highest level of difficulty.

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 were sampled at 6.25Hz.

As brain activity differs widely on a person-by- person basis, we ran all preprocessing of data separately for each participant. We normalized the fNIRS light intensity raw data in each channel by their own baseline values. We then applied a moving average band pass filter to each channel and we use the modified Beer-Lambert Law [8, 19] to convert our light intensity data to measures of the change in oxygenated hemoglobin (ΔHbO) and deoxygenated hemoglobin (ΔHb) in the brain. Therefore, we had a recording of ΔHbO and another recording of ΔHb 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). Both ΔHbO and ΔHb contribute to what is known as the blood oxygen level dependent (BOLD) signal, which is correlated to brain activity. For a review on the BOLD signal see [7] and for a review on the measurement patterns seen in ΔHbO and ΔHb in fNIRS studies see [15]. We cut off the first 4 seconds of each task, as blood takes several seconds to move to areas of activation. Next, we extracted the following features from our preprocessed fNIRS data: *largest value, smallest value, average, slope, time to peak*, and *full width at half maximum;* for the first and second half of each task. We used a cross validation scheme that takes into account the block design of the experiment, as traditional cross validation produces higher classification accuracy that is not representative of real world HCI-relevant applications [12]. Once we had partitioned our data into training and testing data, we used CfsSubsetEval(), a feature subset selection algorithm from the Weka open source toolkit [35] on our training data. The function selects feature subsets that are highly correlated with the class and have a low correlation to one another. We used these features to classify our test data.

USABILITY EXPERIMENT RESULTS AND ANALYSIS

We used the usability experiment protocol described above in our usability studies. Our results and analyses are intended to (in)validate the following research hypotheses:

Experiment Hypotheses

1) We can distinguish between three cognitive resources (visual search, WM, and response inhibition).
2) Not only can we determine which resource is being taxed, but within a given cognitive resource, we can distinguish between high and low-levels of load placed on that resource.
3) We can use our benchmark cognition tasks as training data to build a machine learning classifier. We can use this classifier to determine the user's workload while working with a driving UI and a web search UI.
4) The accuracy and survey results will support the results from the fNIRS data analysis, and the fNIRS results will provide information above and beyond the information provided by the more traditional usability metrics of accuracy and survey data.

In the rest of this section we describe our analysis techniques and we relate our results to each hypothesis. Then we make recommendations to the UIs based on the results.

Experiment Hypothesis 1: *We can distinguish between which cognitive resource is being taxed.*
To address our first hypothesis, we used a Naïve Bayes classifer from Weka's open source toolkit [35] to make pair-wise comparisons between the *cognitive benchmark tasks* (high WM *vs* high VS *vs* high RI). Average classification accuracy across all subjects is reported in Fig. 8.

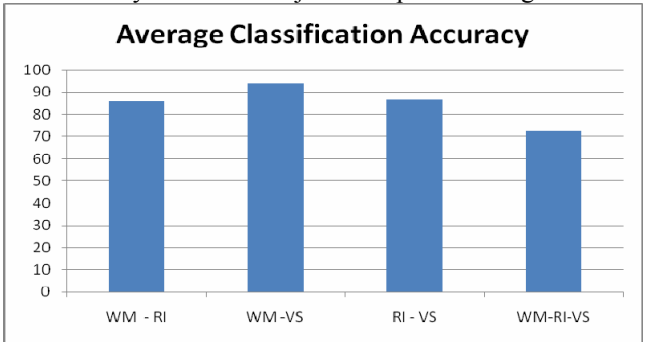

Figure 8: Average classification accuracy across 10 subjects distinguishing between the cognitive benchmark tasks.

As the figure shows, we were able to distinguish between WM and response inhibition (RI), WM and visual search (VS), and response inhibition and visual search with over 80% average accuracy across subjects. Also, we were able to distinguish between the three classes of WM, response inhibition, and visual search with over 70% accuracy. These classification accuracies support our first hypothesis.

Experiment Hypothesis 2: *Not only can we determine which resource is being taxed, but within a given resource, we can distinguish between high and low-levels of load placed on that resource.*
To address our second hypothesis, we used our Naïve Bayes classifer to distinguish between the load (low or high) placed on each cognitive resource. Mean classification accuracies for all subjects are in Fig. 9. As the figure shows, for each resource, we were able to distinguish between low and high-levels of load on that resource with average accuracies ranging from 76-94%. These promising classification accuracies support our second hypothesis.

Experiment Hypothesis 3: *We can use our benchmark tasks as training data to build a machine learning classifier. We can use this classifier to determine the user's workload while working with a driving and a web search UI.*
We used the process described in our usability experiment protocol to predict the WM, response inhibition, and visual search load of the UI variations. Next we will illustrate, for each participant, how we built three distinct machine learning classifiers. Each of the three classifiers was responsible for predicting that participant's level of 1) WM load, 2) response inhibition load, and 3) visual search load while working with the driving and web search tasks.

Construction of Classifiers
For each participant, we built three distinct classifiers. Figure 10 depicts our process for creating a machine learning classifier to predict the WM load (high or low) associated with our driving UI tasks. First, the data from each of the low and high benchmark WM tasks was used as training data to build a Naïve Bayes classifier (see the *fNIRS Data Analysis* section for more detail). This resulted in a WM classifier that predicted whether or not a given test instance had a high or low WM load.

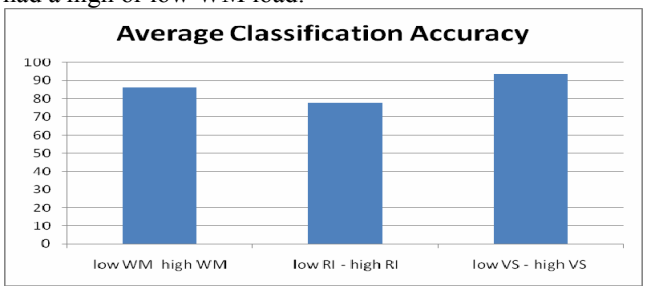

Figure 9: Mean classifier accuracy for 10 subjects.

Next, the same participant's data from the driving UI and web search UI tasks was fed into the classifier as testing data. The classifier predicted the level of WM load (high or low) associated with each UI task. For each subject, we refer to this first classifier as the *WM classifier*.

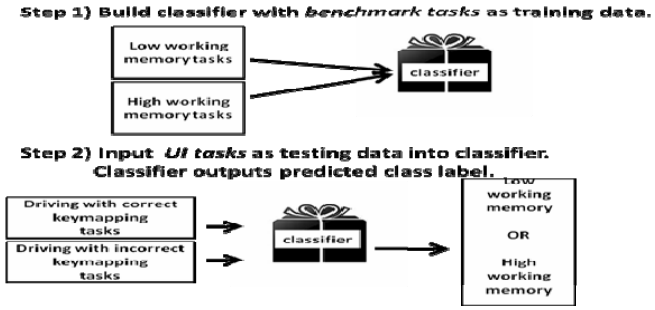

Figure 10: The *WM classifier* is built using WM benchmark tasks. The UI tasks are fed into the classifier as test instances.

Next, for each participant, the low and high response inhibition benchmark tasks were used as training data to build another Naïve Bayes classifier. The same process as that depicted in Figure 10 was followed, however, this time the training data input into the classifier were the response inhibition benchmark tasks rather than the WM benchmark tasks. This resulted in a response inhibition classifier that predicted the load of response inhibition (high or low) placed on users while working with the UI variations. For each subject, we refer to this second classifier as the *response inhibition classifier*.

Lastly, a third Naïve Bayes classifier was built using just the visual search benchmark tasks as training data. This resulted in a visual search classifier that predicted the visual search load (high or low) placed on users while working with the UI variations. For each subject, we refer to this third classifier as the *visual search classifier*.

Composition of Classifiers
For each of our 10 subjects we built three Naïve Bayes classifiers; a WM classifier, a response inhibition classifier,

and a visual search load classifier. Thus, we built 30 distinct classifiers using 30 completely separable sets of training data. While training each classifier for each subject, we had eight sensor locations (four on the left side of the forehead and four on the right) and each sensor generated values of the rate of change of oxygenated blood (HbO) and deoxygenated blood (Hb). Thus, we had 16 timeseries, and we generated 12 features from each of these timeseries (resulting in a total of 192 features that were generated from the raw fNIRS data).We used the CfsSubsetEval feature selection algorithm to prune our features (*see fNIRS Data Analysis section*). The feature selection algorithm pruned the features greatly, and the number of features selected by each classifier ranged from two to 15 features. On average, 7.6 features were selected for classification when building the WM classifiers while an average of 5.2 and 9.1 features were selected while building the response inhibition and visual search classifiers, respectively. Figure 11 displays the composition of the three classifier types across subjects.

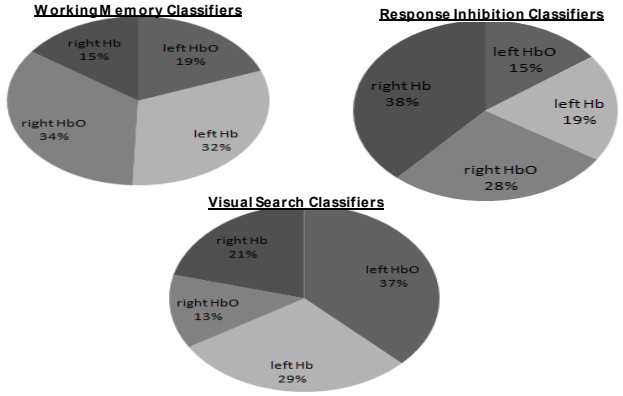

Figure 11: Composition of the classifier types across subjects.

We created a WM classifier, a response inhibition classifier, and a visual search classifier for each of our 10 participants. Thus, we created 10 distinct WM classifiers which were suited to the individual. We tallied the features selected by each of these 10 classifiers and grouped the features based on the sensor location of that feature, and on whether or not the feature represented ΔHbO or ΔHb data. We did the same for the response inhibition and for the visual search classifiers. We made four groups to describe the features selected by each of the classifiers across all subjects: 1) ΔHbO on the left side of the forehead, 2) ΔHbO on the right side of the forehead, 3) ΔHb on the left side of the forehead, and 3) ΔHb on the right side of the forehead.

As shown in the figure, each classifier type differed in structure from the other two classifiers. Participants' ΔHbO on the right side of the head and ΔHb on the left side of the head were the most predictive for the WM tasks. Participants' ΔHbO and ΔHb on the right side of the head were the most predictive for response inhibition tasks. The ΔHbO and ΔHb on the left side of the head were the most predictive for visual search tasks. This supports the hypothesis that the benchmark WM, response inhibition, and visual search tasks used different (though probably overlapping) cognitive resources, and that these differences were measurable with fNIRS.

Classifier Predictions Across Subjects

We ran machine learning predictions for 10 participants x 3 classifiers x 4 UI variations x 6 trials, which resulted in 720 distinct machine learning predictions. We describe our classifier predictions for all participants next. We've broken down our results into four graphs (Figure 11 and Figure 12). Each graph shows a tally of the predictions made by each of the 10 subject's three classifiers while each subject worked on one of the UI variations. Thus, each graph contains a tally of 10 subjects x 6 trials x 3 classifiers = 180 classifier predictions. For each driving UI variation we report the total number of high and low load predictions made by each of the three classifiers across all subjects (10 subjects x 6 trials = 60 instances total . These results are depicted in Figure 12.

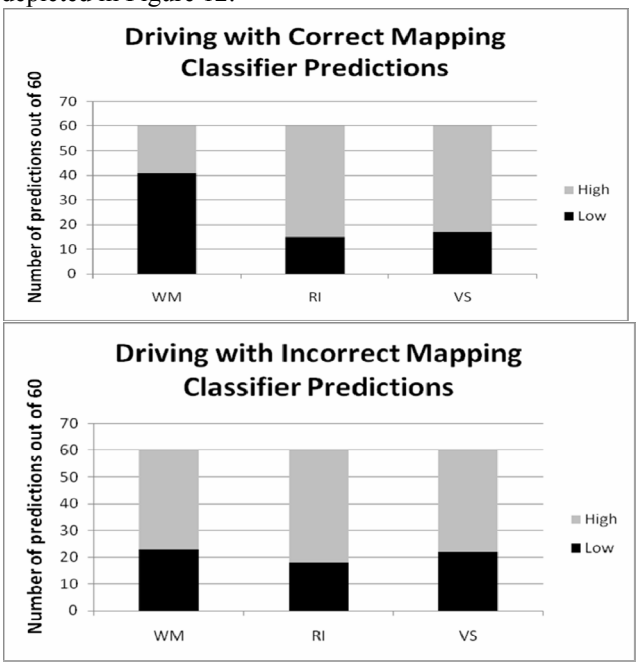

Figure 12: The total number of high and low load predictions made by each of the three classifiers across all subjects..

The results show that across all subjects most of the driving with correct mapping instances were classified as having a low WM, high response inhibition, and high visual search load. The majority of the driving with incorrect mapping instances were classified as having a high WM, high response inhibition, and high visual search load.

It is not surprising that driving with the correct mapping caused users to exert lower levels of WM than driving with the incorrect mapping. It is likely that when using the controls in the incorrect mapping condition, users had to keep the unusual mapping stored in their WM throughout the driving task. It is also not surprising that both driving con-

ditions caused a high visual search load, as users were continually scanning the cones ahead of them while steering. Interestingly, both keyboard mappings (correct and incorrect) caused a high level of response inhibition. We wonder if the choice of using a keyboard (as opposed to a joystick or steering wheel) placed restraints on users' response inhibition by forcing them to inhibit natural steering movements in order to press the right keys. Post-experiment interviews support this speculation.

As shown in Fig. 13, the results showed that across all subjects the majority of the web search with highlighting instances were classified as having a high WM, high response inhibition, and low visual search load. The majority of the web search with no highlighting instances were classified as having a high WM, low response inhibition, and high visual search load. It is not surprising that both web search UIs caused users to have high-levels of WM. After all, users had to remember the content of the question that was asked of them while they searched for the solution in the articles in front of them. It is also not surprising that the level of visual search load was higher when highlighting of search terms was not available than when the highlighting was available for users. We expected that users would have to conduct an inefficient visual search of the articles for the keywords when highlighting was not available.

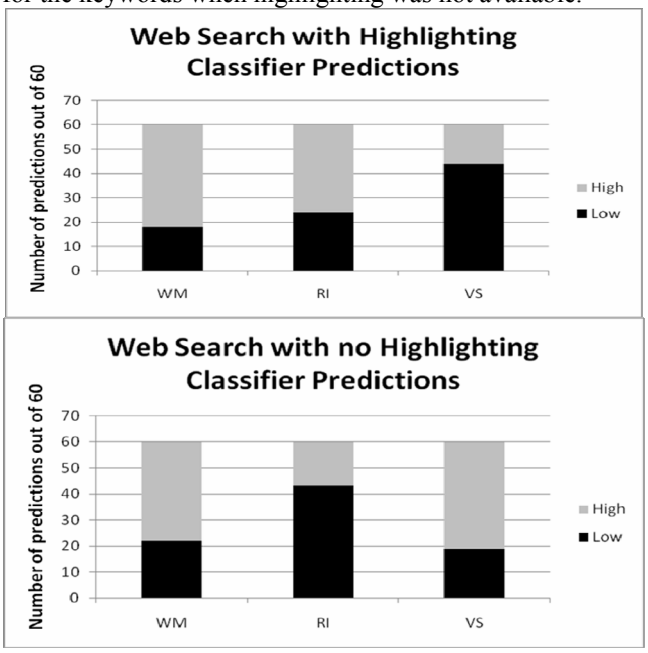

Figure 13: the total number of high and low load predictions made by each of the three classifiers across all subjects.

We did not expect, however, the web search with highlighting to be associated with a higher load of response inhibition than the web search with no highlighting. Since the use of highlighting enabled our users to avoid an inefficient visual search of search items, they were able to spend more time processing the verbal information presented to them in the research articles than in the no highlighting condition. A thorough review of recent experimental psychology literature shows that people use response conflict resources while processing sentences and semantic information [36]. Therefore, we conjecture that the highlighting of search terms enabled users to minimize their visual search load and focus on processing the text in front of them.

This resulted in users having a higher level of response inhibition that reflected the added load that they were able to place on semantic, task related, processing. We hypothesize that the added effort that users were able to place on semantic (or task related) processing rather than on syntactic (or UI related) processing resulted in the higher accuracy that users achieved in the web search with highlighting condition. Our classification results support our third hypothesis; the cognitive load classifications depicted in Fig. 11 and 12 are in line with what we would expect users' cognitive resource load to be while working with each UI.

Experiment Hypothesis 4:
The accuracy and survey results will support the results from the fNIRS data analysis, and the fNIRS results will provide information above and beyond the information provided by the more traditional usability metrics of accuracy and survey data.

We recorded the number of correct and incorrect responses made by participants during the experiment. A repeated measures ANOVA was used on this accuracy data to make comparisons between conditions. A Kruskal-Wallis test was used to analyze the Likert scale survey data for each condition. Table 3 provides a summary of these results. Highlighted cells indicate significance with 95% confidence.Not surprisingly, all accuracy data (except the Stroop tasks) and all survey data indicated that the low benchmark workload (WL) tasks were more difficult than the high benchmark WL tasks. Additionally, the accuracy data supported the fNIRS findings that *driving with the correct mapping* was easier than *driving with the incorrect mapping*, and that web searches with highlighting were easier than searches with no highlighting.

Interestingly, users did not report a difference in difficulty between the UI tasks in the Likert survey. This is in contrast to the accuracy data and to the data from the fNIRS results. Some common, and well known, issues with self-report surveys are that users may not be aware of subtle differences in their own user experiences, and post-surveys may lack insight into the user's real time experiences while working with a given task. One of the primary benefits of non-invasive brain measurement during usability studies is to overcome the short coming of self report surveys.

In general, the results from the accuracy and survey data supported the fNIRS findings. Furthermore, the fNIRS findings provided information about the load placed on users' cognitive resources that was above and beyond that

which could be acquired with the behavioral results alone. This supports our fourth hypothesis.

Implications to UI Designs

The work presented here would not stay true to the premise of a realistic usability study if it didn't shed light on the

particular design choices of each UI design. Therefore, based on the information acquired in our study, we discuss the implications of our findings on the design of the web search and driving UIs next:

Driving UIs

As reflected in Table 3, the driving with correct keymapping appears to be a preferable UI design than the driving with incorrect keymapping as it is associated with fewer errors. Furthermorre, the results in Figure 11 show that the driving with correct keymapping was associated with a lower level of WM load for our subjects than the driving with incorrect keymapping. However, the level of response inhibition and visual search were high for both UI variations. We would suggest using an actual driving wheel or joystick to alleviate the demands placed on users' response inhibition resources while driving. We are currently conducting a follow-on experiment to compare the level of response inhibition exerted by users driving with the keyboard and the response inhibition exerted by users driving with a steering wheel. We expect to see a lower level of response inhibition when users work with the steering wheel than when they use the keyboard.

Table 3: Summary of behavioral results. Shaded cells indicate significance, and we accept the stated hypothesis.

Condition		Hypothesis	Result
Low benchmark tasks	*High benchmark tasks*	*Users' accuracy during low benchmark task < accuracy during high benchmark task*	*Users rated low benchmark task more difficult than high benchmark task*
Congruent Stroop	Incongruent Stroop	$F(9, 81) = 1.589$, $p = .1324$	$H = 10.41$, 1 d.f., $P < .05$
0back	3back	$F(9,81) = 4.004$, $p < .05$	$H = 14.2$, 1 d.f., $P < .05$
Finding a's with highlighting	Finding a's with no highlighting	$F(9,81) = 84.926$, $p < .05$	$H = 13.14$, 1 d.f., $P < .05$
UI version 1	*UI version 2*	*Users' accuracy during UI 1 > accuracy during UI 2*	*Users rated UI version 1 as less difficult than UI version 2*
Driving with correct mapping	Driving with incorrect mapping	$F (9,81) = 6.409$, $p < .05$	$H = 1.36$, 1 d.f., $P = .2443$
Web search with highlighting	Web search with no highlighting	$F (9,81) = 2.094$, $p < .05$	$H = 1.26$, 1 d.f., $P = .2643$

Web Search UIs

As reflected in Table 3, the web search with highlighting UI seems to be a preferable UI design than the web search with no highlighting UI as it is associated with fewer errors. Furthermore, the results in Figure 12 show that the web search with highlighting still causes high levels of WM and response inhibition (due most likely to the task related effort involved in reading through an article for solutions to a specific query). One possible UI enhancement could show the users search terms in a semi-transparent window that follows the users mouse across the computer screen. This could alleviate users' WM demands by enabling them to see their search terms within the context of the article they are reading instead of requiring them to recall the search terms while searching through articles.

CONCLUSION

Our experiment results showed that we were able to use brain measurement to measure high and low-levels of load experienced by users' various cognitive resources while working with our driving and web search UI variations. The usability metrics provided by our survey and accuracy data yielded results that are in line with our fNIRS results. The protocol presented in this paper, combined with the fNIRS data acquired during the usability experiment, provided us with information above and beyond the knowledge gained by the more traditional survey and accuracy usability metrics. With the fNIRS data, we were able not only to determine which tasks were 'more difficult' for participants, but to shed light on the low-level cognitive resources in the brain that were more heavily taxed by a given UI design choice.

FUTURE WORK

While our fNIRs device only provided measurements on the left and right side of the participant's forehead, there are devices that can acquire more measurements across participant's cortex. In order to find regions of the brain that are activated while load is placed on various cognitive resources, we are using a new 52-channel fNIRS device to explore the use of more sensor locations across the forehead, enabling the measurement of small spatiotemporal changes that occur when different cognitive resources are taxed. In the future, we foresee brain data as an additional metric gathered in usability tests. This cognitive state information, combined with more traditional usability metrics such as speed, accuracy, and survey results, can provide in-depth evaluations of a range of UIs.

REFERENCES

1. Adcock, R.A., Constable, R.T., Gore, J.C. and Goldman-Rakic, P.S. Functional neuroanatomy of executive processes involved in dual-task performance. *Proceedings of the National Academy of Sciences of the United States of America*, *97* (7). 3567-3572.
2. Anderson, E.J., Mannan, S.K., Husain, M., Rees, G., Summer, P., Mort, D.J., McRobbie, D. and Kennard, C. Involvement of prefrontal cortex in visual search. *Experimental Brain Research*, *180* (2). 289-302.
3. Baddeley, A. and Della Sala, S. Working memory and executive control. *Philosophical Transactions of the Royal Society of London*, *351*.
4. Berka, C., Levendowski, D., Cvetinovic, M., Petrovic, M., Davis, G., Lumicao, L., Zivkovic, V., Popovic, M. and Olmstead, R. Real-Time Analysis of EEG Indexes of Alert-

ness, Cognition, and Memory Acquired With a Wireless EEG Headset. *International Journal of Human Computer Interaction*, *17* (2). 151–170.
5. Blackmon, M.H., Kitajima, M. and Polson, P.G., Tool for accurately predicting website navigation problems, non-problems, problem severity, and effectiveness of repairs. in *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (Portland, Oregon, USA, 2005).
6. Brumby, D.P., Salvucci, D.D. and Howes, A., Focus on driving: how cognitive constraints shape the adaptation of strategy when dialing while driving. in *In Proc. of the 27th international Conference on Human Factors in Computing Systems*, (Boston, MA, USA, 2009), ACM.
7. Buxton, R. *Introduction to functional magnetic resonance imaging*. Cambridge University Press, Cambridge, United Kingdom, 2002.
8. Chance, B., Anday, E., Nioka, S., Zhou, S., Hong, L., Worden, K., Li, C., Murray, T., Ovetsky, Y. and Thomas, R. A novel method for fast imaging of brain function, non-invasively, with light. *Optics Express*, *10* (2).
9. 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.
10. Eckstrom, R., French, J., Harman, H. and Derman, D. Kit of factor-referenced cognitive tests.
11. Gevins, A. and Smith, M. Neurophysiological Measures of Working memory and Individual Differences in Cognitive Ability and Cogntive Style. *Cerebral Cortex*, *10*.
12. Grimes, D., Tan, D., Hudson, S., Shenoy, P. and Rao, R., Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph. in *CHI 2008 Conference on Human Factors in Computing Systems*, (Florence, Italy., 2008).
13. Guhe, M., Liao, W., Zhu, Z., Ji, Q., Gray, W.D. and Schoelles, M.J., Non-intrusive measurement of workload in real-time. in *49th Annual Conference of the Human Factors and Ergonomics Society*, (2005), 1157-1161.
14. 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.
15. Hirshfield, L.M. Enhancing Usabiltiy Testing with Functional Near Infrared Spectroscopy *Computer Science*, Tufts University, Medford, MA, 2009.
16. Hirshfield, L.M., Solovey, E.T., Girouard, A., Kebinger, J., Jacob, R.J.K., Sassaroli, A. and Fantini, S., Brain Measurement for Usability Testing and Adaptive Interfaces: An Example of Uncovering Syntactic Workload in the Brain Using Functional Near Infrared Spectroscopy. in *Proc. of SIGCHI*, (2009).
17. Hoshi, Y. and Tamuraa, M. Near-Infrared Optical Detection of Sequential Brain Activation in the Prefrontal Cortex during Mental Tasks. *NeuroImage*, *5*. 292-297.
18. Hoshi, Y., Tsoub, B., Billockc, V., Tanosakia, M., Iguchia, Y., Shimadaa, M., Shinbaa, T., Yamadad, Y. and Odae, I. Spatiotemporal characteristics of hemodynamic changes in the human lateral prefrontal cortex during working memory tasks. *NeuroImage*, *20*. 1493-1504.
19. 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).
20. Jaeggi, S.M., Seewer, R., Nirkko, A.C., Eckstein, D., Schroth, G., Groner, R. and Gutbrod, K. Does excessive memory load attenuate activation in the prefrontal cortex? Load-dependent processing in single and dual tasks: functional magnetic resonance imaging study. *NeuroImage*, *19* (2). 210-225.
21. Jancke, L., Brunner, B. and Esslen, M. Brain activation during fast driving in a driving simulator: the role of the lateral prefrontal cortex. *Neuroreport*, *19* (11).
22. Jasdzewski, G., Strangman, G., Wagner, J., Kwong, K., Poldrack, R. and Boas, D. Differences in the hemodynamic response to event-related motor and visual paradigms as measured by near-infrared spectroscopy. *NEUROIMAGE*, *20* (1). 479-488.
23. Joanette, Y., Ansaldo, A., de Mattos Pimenta Parente, M., Fonseca, R., Kristensen, C. and Scherer, L. Neuroimaging investigation of executive functions: evidence from fNIRS. *PSICO*, *39* (3).
24. 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).
25. Meek, J., Elwell, C., Khan, M., Romaya, J., Wyatt, J., Delpy, D. and Zeki, S. Regional Changes in Cerebral Haemodynamics as a Result of a Visual Stimulus Measured by Near Infrared Spectroscopy. *Proc. Roy. Soc. London*, *261*. 351-356.
26. Muller-Plath, G. Localizing subprocesses of visual search by correlating local brain activation in fMRI with response time model parameters. *Journal of Neuroscience Methods*, *171* (2). 316-330.
27. Parasuraman, R. and Caggiano, D. Neural and Genetic Assays of Human Mental Workload. in *Quantifying Human Information Processing*, Lexington Books, 2005.
28. Scerbo, M., Frederick, G., Freeman, F. and Mikulka, P. A brain-based system for adaptive automation. *Theoretical Issues in Ergonomics Science*, *4* (1).
29. Schroeter, M.L., Zysset, S., Kupka, T., Kruggel, F. and Yves von Cramon, D. Near-Infrared Spectroscopy Can Detect Brain Activity During a Color-Word Matching Stroop Task in an Event-Related Design. *Human Brain Mapping*, *17* (1). 61-71.
30. Smith, E. and Jonides, J. Storage and Executive Processes n the Frontal Lobes. *Science*, *283*.
31. Solovey, E., Girouard, A., Chauncey, K., Hirshfield, L., Sassaroli, A., Zheng, F., Fantini, S. and Jacob, R., Using fNIRS Brain Sensing in Realistic HCI Settings: Experiments and Guidelines. in *ACM UIST Symposium on User Interface Software and Technology*, (2009), ACM.
32. Tamborello, F. and Byrne, M.D., Information search: The intersection of visual and semantic space. in *CHI 2005 Extended Abstracts*, (Portland, OR, USA, 2005), ACM, 1821-1824.
33. Wickens, C., Lee, J., Liu, Y. and Becker, S. *An Introduction to Human Factors Engineering*. Pearson, 2004.
34. Wilson, G.F. and Fisher, F. Cognitive task classification based upon topographic EEG data. *Biological Psychology*, *40*. 239-250.
35. Witten, I.H. and Frank, E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.
36. Ye, Z. and Zhou, X. Conflict control during sentence comprehension: fMRI evidence. *Neuroimage*, *48* (1). 280-290.