Using fNIRS Brain Sensing to Evaluate Information Visualization Interfaces

Evan M Peck
Tufts University
evan.peck@tufts.edu

Beste F Yuksel
Tufts University
beste.yuksel@tufts.edu

Alvitta Ottley
Tufts University
alvittao@cs.tufts.edu

Robert J.K. Jacob
Tufts University
jacob@cs.tufts.edu

Remco Chang
Tufts University
remco@cs.tufts.edu

ABSTRACT

We show how brain sensing can lend insight to the evaluation of visual interfaces and establish a role for fNIRS in visualization. Research suggests that the evaluation of visual design benefits by going beyond performance measures or questionnaires to measurements of the user's cognitive state. Unfortunately, objectively and unobtrusively monitoring the brain is difficult. While functional near-infrared spectroscopy (fNIRS) has emerged as a practical brain sensing technology in HCI, visual tasks often rely on the brain's quick, massively parallel visual system, which may be inaccessible to this measurement. It is unknown whether fNIRS can distinguish differences in cognitive state that derive from visual design alone. In this paper, we use the classic comparison of bar graphs and pie charts to test the viability of fNIRS for measuring the impact of a visual design on the brain. Our results demonstrate that we can indeed measure this impact, and furthermore measurements indicate that there are *not* universal differences in bar graphs and pie charts.

Author Keywords

fNIRS; BCI; visualization; brain sensing; evaluation.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces

General Terms

Human Factors; Design; Measurement.

INTRODUCTION

The quantitative evaluation of visual interfaces has been a significant goal of both the HCI and visualization community for decades. Numerous quantitative and qualitative approaches have been proposed to peek into the user's cognitive processes during interaction. Nevertheless, there are limitations to evaluating performance in a visual interface without directly monitoring the brain's cognitive processes. Evaluations of basic tasks may not generalize to complex tasks using the same visual forms (i.e. bar graphs and pie charts [6, 28, 32]), and psychology research suggests that evaluating performance without workload may lead to incorrect conclusions about the cognitive efficiency of an interface [3, 17, 23, 40]. Finally, cognitive state can change even as performance remains stable, meaning that performance metrics may not always accurately reflect cognitive processes [7, 37].

As a result, there has been a renewed interest in objective methods to evaluate cognitive processes during interaction with a visual interface [1, 26]. In particular, *functional near-infrared spectroscopy (fNIRS)* (Figure 1) has received increased attention as a lightweight brain sensing technology in HCI because in comparison to other brain sensing methods, it is portable [36], resistant to movement artifacts [19], and observes similar physiological parameters to fMRI [5, 35, 39]. While previous fNIRS experiments in HCI have studied cognitive state at various stages of interaction [2, 12, 15, 16, 29, 31], these experiments largely omit a critical component of interface design: How do different visual designs and interfaces affect the user's ability to perform visual judgment at a cognitive level?

The potential of using fNIRS to inform the design of interactive interfaces for visualization is appealing. If fNIRS can successfully measure the impact of visual design on the user, then it can provide access to physiological parameters that have not previously been analyzed in this context. Furthermore, it can do so in ecologically sound settings that allow users to interact naturally with an interface [30].

However, there are concerns as to whether fNIRS may be capable of monitoring brain activity in these scenarios. The physiological parameters which fNIRS monitors (oxygenated and deoxygenated hemoglobin) typically peak 5-7 seconds after interaction, meaning that the signal is slow-moving in comparison to the massively-paralleled processes employed by the brain's visual system. In addition, tasks that leverage the perceptual system may not induce measurable activity in the prefrontal cortex (PFC), the area of the brain most commonly monitored by fNIRS.

CHI 2013, April 27–May 2, 2013, Paris, France.

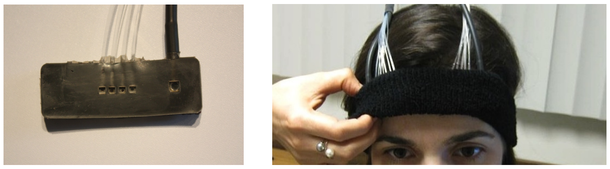

Figure 1. Left: An fNIRS probe with four light source and one light detector. Right: Two fNIRS probes are secured on a participant's forehead using a headband.

In this work, we test the viability of using fNIRS to observe how visual design modifies brain activity in complex tasks. We conducted three experiments to (a) examine how participants process bar graphs and pie charts differently in their brains, (b) determine the efficacy of using fNIRS as a technique for evaluating mental workload in visual tasks, and (c) classify visual tasks that are most suited for using fNIRS in evaluation.

To investigate this, we employ a classical comparison in the field of visualization - bar graphs and pie charts - and ask users to perform a difficult task on the information contained in those graphs. Based on our results, we make three contributions:

- **Our findings suggest that fNIRS can be used to monitor differences in brain activity that derive exclusively from visual design.** We find that levels of deoxygenated hemoglobin in the prefrontal cortex (PFC) differ during interaction with bar graphs and pie charts. However, there are *not* categorical differences between the two graphs. Instead, changes in deoxygenated hemoglobin correlated with the type of display that participants believed was more difficult. In addition, participants reacted differently to pie charts and bar graphs at a cognitive level, but exhibited the same performance characteristics.
- **We propose that the fNIRS signals we observed indicate the amount of cognitive workload induced by interacting with a visual interface.** We conducted an experiment that compares brain activity observed in bar graphs and pie charts with activity from a visuospatial n-back task - a well-characterized task from the psychology literature for modifying load on working memory. Our results are consistent with the existing fMRI literature and agree with participant response data (NASA-TLX), indicating that fNIRS signals correlate with cognitive workload.
- We discuss the benefits of using fNIRS for evaluating visual design and conduct an auxiliary study to identify the limits of using fNIRS in perceptually driven tasks. **We find that fNIRS can provide insight on the impact of visual design during interaction with difficult, analytical tasks, but is less suited for simple, perceptual comparisons.**

BACKGROUND

Brain and Body Sensing in Visualization Evaluation

As Fairclough [10] points out in his seminal review, physiological sensing in HCI has the advantage of having higher temporal fidelity in that it can access data at any time. In contrast, post-hoc questionnaires or recordings of observable behaviors represent discrete and sporadic events that reflect aggregated opinions about a whole experience.

While the field of HCI has seen an increased acceptance of physiological sensing in evaluation, to date, this push has not translated to the evaluation of visual interfaces and visual form. Historically, recording behavioral metrics or administering questionnaires have been used to evaluate visual design. However, Riche [26] notes that the exploratory nature of tasks in infovis systems, coupled with the "the difficulty to decompose [them] into low-level and more easily measured actions" makes analysis problematic. To overcome some of these obstacles, Riche proposes the use of physiological measures to evaluate visual interfaces.

Unfortunately, to our knowledge, there have been only two significant attempts to explore this space. Investigating the impact of visual variables on heart rate, galvanic skin response (GSR), and respiratory rate, Causse and Hurter found that interactions with text v. angle-based visual forms elicited different signals with GSR [4]. Few other significant interactions were observed. Work by Anderson et al. is the most promising example of using physiological signals to evaluate visual interfaces [1]. They used electroecephalography (EEG) to determine that the canonical box plot requires less extraneous load (i.e. the additional load placed on users by the design of a task) than various other box plot designs [1].

However, there are notable caveats to the use of EEG. While EEG has a high temporal resolution, it also has a low spatial resolution, meaning that the origin of recorded electrical activity is difficult to locate. Additionally, EEG has traditionally been considered to be extremely sensitive to movement artifacts, although recent developments have lessened this issue [24].

Brain Sensing with FNIRS

An alternative technology to objectively monitor activity in the brain is *functional near-infrared spectroscopy (fNIRS)*, an optical brain sensing device that has the potential to lend insight to visual interactions [5, 36, 39].

fNIRS uses near-infrared light to measure concentration and oxygenation of the blood in the tissue at depths of 1-3cm [36]. Light is sent into the forehead in the near infrared range (650-900 nm), where it is diffusely reflected by the scalp, skull, and brain cortex. At this wavelength, oxygenated and deoxygenated hemoglobin are the primary absorbers of light. A very small percentage of the light sent into the head returns from the cortex to the detector on the fNIRS probe. By measuring the light returned to the detector, researchers are able to calculate the amount of oxygen in the blood, as well as the amount of blood in the tissue. Since changes in blood flow and oxygenation indicate activation levels in the brain we can use fNIRS to measure activity in localized areas of the brain.

In general, fNIRS is quick to set up and more tolerant of user movement than other brain sensing techniques such as fMRI or EEG - a critical feature for ecologically valid evaluation [30, 35]. Investigating the use of fNIRS in user studies, Solovey et al. [30] found that mouse-clicking, typing, eye

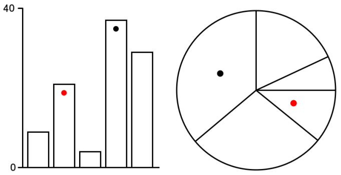

Figure 2. An example bar graph and pie charts from Cleveland and McGill's comparison task. Participants were asked to make a percentage estimation of the smaller section, marked by a red dot, with the larger section, indicated by a black dot.

movement, and blinking do not disrupt the fNIRS signal. Additionally, minor head movement, respiration, and heartbeats are correctable, as they can be filtered out using known signal processing techniques. Only major head and forehead movement (which could be induced by frowning) are disruptive to the signal [30].

FNIRS readings have been validated against fMRI and the device is sensitive to physiological parameters that are not accessible by other brain sensing techniques [35]. Because it measures the relatively slow hemodynamic response of blood to the brain (5-7 seconds), fNIRS has a slow temporal resolution in comparison to EEG, which indirectly measures the electric activity of the brain. However, this light-based sensing technique allows fNIRS to have a spatial resolution of 1-3cm, which is much sharper than EEG (although less precise than fMRI).

As a result, fNIRS has seen increased use for research in HCI as a complementary device to EEG [12]. Hirshfield et al. [16] used fNIRS to create a novel experimental protocol to explore the mental workload of users. They then used that protocol to measure the syntactic workload of users while interacting with two different interfaces. One of the most recent examples is from Solovey et al. [29, 31], who explored the use of fNIRS in designing adaptive interfaces to support multitasking. This led to the system *Brainput* which can identify different brain signals occuring naturally during multitasking and use these to modify the behavior of the interface [29]. FNIRS has also been used to monitor workload changes in air traffic control tasks and piloting a virtual unmanned aerial vehicle (UAV) [2].

These studies have been important in the development of fNIRS within HCI. However, the cognitive effects of different visual displays on the user is a yet unexplored area.

Pie Charts and Bar Graphs

We chose the visualization of bar graphs and pie charts as a suitable testbed for monitoring the user's cognitive processes because it is a familiar, well-studied comparison in the field of information visualization. In this section, we briefly outline the body of research that studies interaction with bar graphs and pie charts.

In Cleveland and McGill's ranking of visual variables, participants were presented with either a bar graph or pie chart (Figure 2) and asked to estimate the proportion percentage of a smaller value in the graph to a larger value [6]. Their results indicated that position judgments (bar graphs) facilitated more accurate visual comparisons than angle judgments (pie charts).

However, Simkin and Hastie found that pie charts and bar graphs performed equally well in part-to-whole comparisons [28]. Spence and Lewandowsky demonstrated that pie charts perform reasonably well in a direct comparison with other basic visual forms [32]. In more complex tasks - when comparisons consist of combined proportions (A+B v. C+D) - pie charts can outperform bar graphs [34]. For a more extensive history of the pie chart, see Spence's article "No Humble Pie: The Origins and Usage of a Statistical Chart" [33].

Recently, there have been a handful of studies that utilize Cleveland and McGill's comparison as a baseline to investigate various dimensions of interaction. Heer et al. replicated Cleveland and McGill's experiment using Mechanical Turk, demonstrating that "crowd sourcing" is a viable mechanism for graphical perception experiments [14]. Using pie charts and bar graphs, Hullman et al. showed that social factors can influence quantitative judgments [18]. For example, showing a user a histogram of previous responses to a visual comparison would dramatically skew the user's own judgment. Finally, Wigdor et al. explored the impact of distortion on angle and position judgments in tabletop displays. They found that varying the orientation of the display surface altered visual comparisons [38].

Despite the sizable body of research that has investigated bar graphs and pie charts, these studies also indicate that as the task or environment change, performance differences between the two forms become less clear. Therefore, we find this familiar comparison to be a sufficient baseline for objectively exploring users' cognitive processes with fNIRS.

RESEARCH GOALS

Our primary goal in this work was to investigate the viability of using fNIRS to evaluate visual design by having participants perform the same complex task on both bar graphs and pie charts. We theorized that in a complex task, bar graphs and pie charts would support the cognitive processes of the user differently. Thus, our principal hypothesis was as follows:

- *Hypothesis:* We will observe different brain signals during interaction with bar graphs and pie charts, indicating that bar graphs are easier to use.

Depending on the outcome of our experiments, our secondary goal was to further specify the use of fNIRS in visualization research. First, we compared fNIRS signals from participants in a well-established psychology task (n-back task) to those observed in bar graphs and pie charts. We combined those observations with previous fMRI literature and participant survey responses to surmise the underlying cognitive processes associated with our fNIRS signal. Additionally, we performed an auxiliary study using simple comparisons on bar graphs and pie charts to identify a lower bound for using fNIRS in visualization research. We present these results below, after the main experiment.

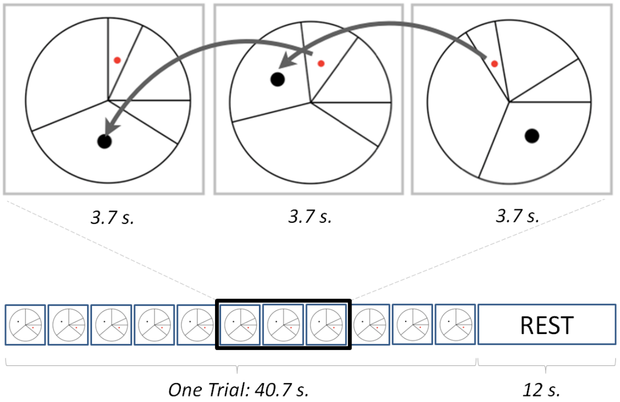

Figure 3. In our modified comparison task, participants compare a slice in the current pie chart to a slice from the previously seen pie chart.

In the following sections, we outline the methodology used for our bar graph v. pie chart experiment, discuss the results of that experiment, and finally, generalize our study to visualization research.

METHODS

Although originally inspired by Cleveland and McGill's classical position v. angle experiment, we modified the complexity of their task in order to reconstruct the memory-intensive, analytical reasoning that is performed on high-performance visual interfaces. For that reason, we modeled our task loosely after the n-back task, a well-characterized psychology task that is meant to increase load on working memory.

In this task, participants were presented a series of slides, each displaying either a bar graph or pie chart, to view sequentially. They were instructed to estimate the size difference to the nearest ten percent of a smaller section of the graph (marked by a red dot) in the current slide to a larger section (marked by a black dot) in the *previous* slide. Estimates were entered using a single keystroke on the keyboard ('1' for 10 percent, '2' for 20 percent, etc). Figure 3 shows an example of three slides using the pie chart condition.

Each trial lasted 40.7 seconds and consisted of 11 slides (or 10 comparisons with the previous slide), with each slide being presented for 3.7 seconds. Participants viewed 8 trials where the task depended on bar graphs and 8 trials where the task depended on pie charts. Trials were shown in random order.

To construct the graphs, 88 datasets (8 trials x 11 slides) were randomly generated at the time of the experiment using the same constraints as those outlined in Cleveland and McGill's classical angle v. position experiment. Accordingly, the same datasets were used for both bar graphs and pie charts. Comparisons were chosen at run-time by randomly selecting one of the largest two graph elements in the current slide and one of the smallest three elements in the next slide. This final constraint was necessary to guarantee that the two marked segments of each graph would not overlap and that percentage estimates would not exceed 100%.

Measures

Questionnaire: NASA TLX

We used an unweighted NASA-TLX questionnaire [20], a subjective rating that has been successfully used to capture workload since the 1980s [13]. The questionnaire collects six components of workload - *mental demand, physical demand, temporal demand, performance, effort,* and *frustration*. In total, we collected two surveys reflecting the two conditions - bar graphs and pie charts. We focus primarily on the questionnaire's mental demand dimension.

Brain Sensing: fNIRS Signal Analysis

We used a multichannel frequency domain OxyplexTS from ISS Inc. (Champaign, IL) for fNIRS data acquisition. Two fNIRS probes were placed on the forehead in order to measure the two hemispheres of the PFC (Figure 1). The source-detector distances were 1.5, 2, 2.5, and 3cm. Each distance measures a difference depth in the cortex. Each source emits two light wavelengths (690 nm and 830 nm) to detect and differentiate between oxygenated and deoxygenated hemoglobin. The sampling rate was 6.25Hz. For each of the two fNIRS probes, we selected the fNIRS measurement channels with source-detector distances of 3cm, as the light from these channels is expected to probe deepest in the brain tissue, while the closer channels are more likely to pick up systemic effects and noise.

To remove motion artifacts and optical changes due to respiration and heart beat we applied a folding average filter using a non-recursive time-domain band pass filter, keeping frequencies between 0.01Hz and 0.5Hz. The filtered raw data was then transformed into oxygenated hemoglobin and deoxygenated hemoglobin concentrations using the modified Beer-Lambert Law [5]:

$$\Delta A = \varepsilon \times \Delta c \times d \times B \tag{1}$$

where ΔA is the change in attenuation of light, ε is the molar absorption coefficient of the absorbing molecules, Δc is the change in the concentration of the absorbing molecules, d is the optical pathlength (i.e., the distance the light travels), and B is the differential pathlength factor. The attenuation of light is measured by how much light is absorbed by oxygenated and deoxygenated hemoglobin (which are the main absorbers of near infra-red light at these wavelengths). As the attenuation of light is related to the levels of hemoglobin, given ΔA, we can derive the changes in the levels of oxygenated and deoxygenated hemoglobin [5]. Finally, to remove noise artifacts, we smoothed the data by fitting it to a polynomial of degree 3 and applied a low-pass elliptical filter [31].

Performance: Speed and Accuracy

We logged all key-strokes and response times. We defined response time as the number of milliseconds from a graph's appearance to the final keystroke (user judgment) before the next graph. For accuracy, we used Cleveland and McGill's log absolute error measures of accuracy [6]:

$$\text{error} = \log_2(|\text{judged percent} - \text{true percent}| + .125) \tag{2}$$

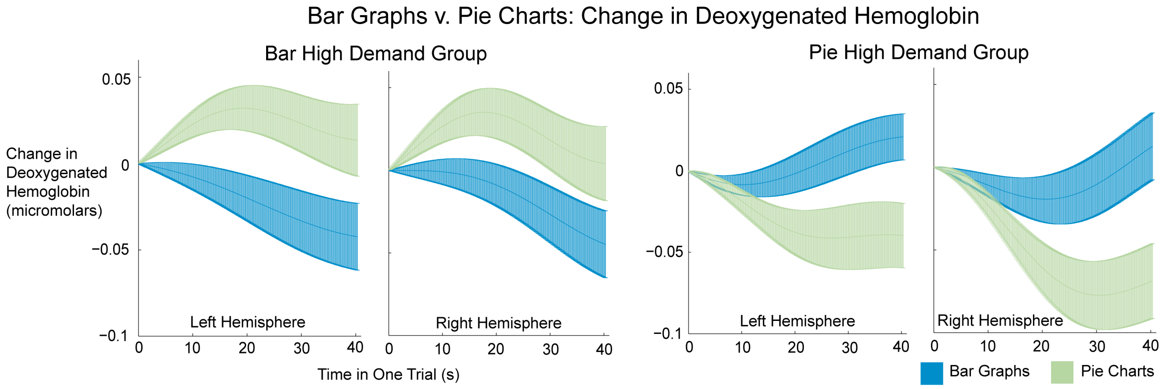

Figure 4. In a user study involving bar graphs (blue) and pie charts (green), we found that a group of participants that subjectively rated bar graphs as more mentally demanding than pie charts (left) exhibited reserved fNIRS signals from those who rated pie charts as more mentally demanding than bar graphs (right). The differences between signals in each graph demonstrate that brain sensing with fNIRS can monitor neural activity derived exclusively from visual. The plots represent the mean change in deoxygenated hemoglobin across all trials of each condition. The width of the line represents the standard error at each time point.

Experimental Design

16 participants took part in the study (7 male, 9 female). Participants had a mean age of 20 years (SD 2.4) and were incentivized $10 for participation. The study used a within-subjects design. All participants completed a fifteen minute bar graph v. pie chart task in which the independent variable was the data visualization technique: *bar graphs*, *pie charts*. Participants also completed a fifteen minute visuospatial n-back task in which the independent variable was the number of slides the participant needed to remember at once: *1-back*, *3-back* (we discuss the results of this experiment in our investigation of fNIRS signals and workload). At the conclusion of each section, participants completed an unweighted NASA-TLX questionnaire for each condition. The order of sessions (n-back, angle vs. position) was counterbalanced and the order of conditions (1-back vs. 3-back, bar graph vs. pie chart) in each session was randomized. The study was conducted in a lab setting, with stimuli presented on a single monitor under controlled lighting conditions.

RESULTS

For the purpose of analyzing the fNIRS signal, we calculated the mean change in deoxygenated hemoglobin ($\overline{\Delta Hb}$) across the duration of each trial (omitting the first 10 seconds[1]) for each participant as shown in equation (3):

$$\overline{\Delta \mathrm{Hb}} = \frac{\sum_{\mathrm{t}=0}^{\mathrm{n}}(\mathrm{Hb_t} - \mathrm{Hb_0})}{\mathrm{n}} \tag{3}$$

where n is the number of time-points, Hb_0 is the level of deoxygenated hemoglobin at the first recorded point (time zero), and Hb_t is the level of deoxygenated hemoglobin at time-point t of a trial. The change in deoxygenated hemoglobin (ΔHb) is calculated by subtracting Hb_0 from the level of deoxygenated hemoglobin at each time-point t. This is one of many techniques that have been used in the fNIRS literature to evaluate changes in oxygenated and deoxygenated hemoglobin [2]. While there may be boundary cases in which this measure is not sensitive to differences between signals, in this case, it captures the clear distinction between conditions.

fNIRS Signal: Bar Graphs v. Pie Charts

Addressing our initial hypothesis, we found no significant differences in deoxygenated hemoglobin between the bar graph ($M = -.0292, SD = .0471$) and pie chart ($M = -.0249, SD = .0679$) conditions ($t(15) = -.280, p = .784$). Contrary to our initial belief, these results indicate that there were no categorical differences in brain activity between the two visual forms. However, during the examination of data from NASA-TLX questionnaires, we encountered an interesting trend. In this section, we discuss and analyze this.

NASA-TLX Results

Isolating the mental demand dimension of the NASA-TLX survey, we found that 7 out of 16 participants believed pie charts to be more mentally demanding than bar graphs while an additional 7 participants expressed that bar graphs were more mentally demanding than pie charts (a remaining 2 participants found the graphs to require equal amounts of mental effort). These responses were largely unexpected, as our hypothesis indicated that we would likely find a categorical difference between bar graphs and pie charts. For the sake of clarity, those who thought pie charts to be more mentally challenging will be referred to as **pie high demand** and those who thought bar graphs to be more mentally demanding will be referred to as **bar high demand**.

fNIRS Signal: Bar High Demand v. Pie High Demand

Investigating the differences in these two groups, we found that the levels of deoxygenated hemoglobin exhibited by participants who found bar graphs more mentally demanding were the *reverse* of those participants who found pie charts more mentally demanding. Figure 4 shows that in the *bar high demand group*, we observed a decrease in deoxygenated hemoglobin in both the left and right hemisphere during tasks completed on bar graphs. In comparison, these same interactions induced a slight increase in deoxygenated hemoglobin in the *pie high demand group*.

[1] Omitting the first 10 seconds of the trial is due to the delayed physiological response of sending oxygen to the brain

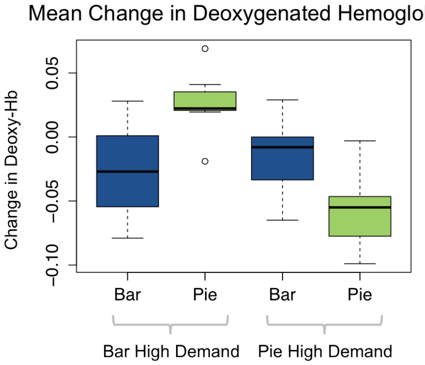

Figure 5. The mean change in deoxygenated hemoglobin for each graph shows that the visual design that partipants found to be more difficult resulted in larger decreases in deoxygened hemoglobin.

Thus, we performed an ANOVA on the mean change in deoxygenated hemoglobin using a 2 (task) x 2 (group) split plot design. The ANOVA revealed a significant difference between groups ($F(1,12) = 9.95, p < .01$), as well as a significant interaction between groups (*pie high demand* and *bar high demand*) and task ($F(1,12) = 16.49, p < .01$). This finding shows that participants in the *pie high demand group* and the *bar high demand group* showed significantly different patterns of deoxygenated hemoglobin while performing the two tasks (Figure 5). Note that while the mean provides a suitable metric for analysis, it can miss some trends in time-series data. Specifically, Figure 5 suggests that both groups recorded similar changes in deoxygenated hemoglobin while interacting with bar graphs. However, Figure 4 shows that the fNIRS signal was trending in opposite directions.

Performance: Bar High Demand v. Pie High Demand

In light of these group differences, we performed another analysis on response times by running a similar ANOVA on mean response time using a 2 (task) x 2 (group) split plot design. After ensuring that the data fit a normal distribution, we found no significant interaction between groups and tasks ($F(1,12) = 2.425, p = .145$). Similarly, an ANOVA on log error as shown in equation (2) found no significant difference in the interaction between group and task ($F(1,12) = .51, p = .4907$). We display a box-plot of log error and response time for each of the two groups in Figure 6.

These results suggest that although there were significant differences in brain activity between bar graphs and pie charts, there was no observable differences in performance, either categorically (bar graphs v. pie charts) or between group (bar high demand v. pie high demand). This is a very different result from those observed by Cleveland and McGill [6], in which position judgments (bar graphs) were found to significantly outperform angle judgments (pie charts). However, given the complex nature of the task, it is not surprising that performance corresponds more closely to findings from Spence and Lewandowsky that pie charts can perform as well, or better than bar graphs in difficult tasks [32, 34].

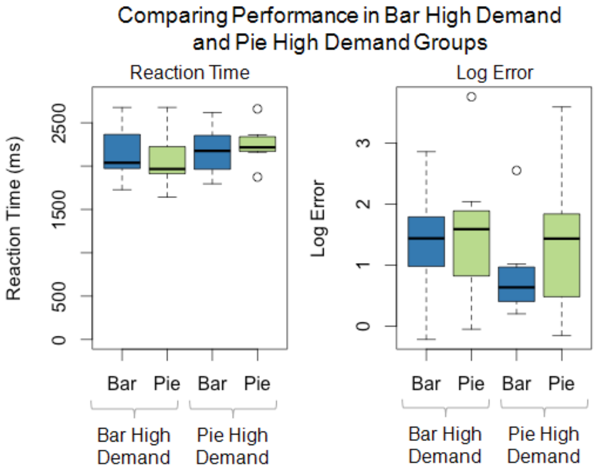

Figure 6. Despite a clear separation in brain activity between the bar high demand group and the pie high demand group, we observe very little difference in response time and error. The whiskers represent the max/min values, excluding outliers. Outliers are assigned by being more/less than 1.5 times the value of the upper/lower quartiles.

DISCUSSION OF BAR GRAPHS AND PIE CHARTS

Our results show that changes in deoxygenated hemoglobin during the use of bar graphs in a complex task are statistically different from those observed during the use of pie charts. However, this distinction was not categorical. Instead, brain activity depended on the individual and correlated with reports of mental demand in a NASA-TLX questionnaire. These differences between participants may call into question the conventional wisdom to *always* use bar graphs instead of pie charts.

Differences in Perceived Mental Demand

In the background, we outlined studies that used performance metrics of speed and accuracy to compare the use of bar graphs and pie charts. We expected that self-reports of mental demand would roughly resemble performance trends, and following previous research, one visual form would be categorically favored over the other. However, we discovered that 14 out of 16 participants found one chart to be more mentally demanding than the other. **Therefore, we reject our initial hypothesis that brain signals would indicate that bar graphs are easier to use for most people**.

Subjectively, there was no indication that either bar graphs or pie charts were superior across all participants on this particular task. 7 participants reported pie charts to be more mentally demanding and 7 participants reported bar graphs to be more mentally demanding (the final 2 reported no noticeable difference). Although we did not investigate the underlying cause of this observation, we suspect that this is due to either differences in cognitive traits (e.g. spatial ability), strategies employed to complete the task, or previous experience with bar graphs and pie charts.

Survey Responses and fNIRS Signals

While surveys can be found to be affected by bias or an inability to accurately externalize cognitive state, we found a surprising correlation between fNIRS readings and mental demand reports on NASA-TLX. **The graph that participants reported to be more mentally demanding recorded decreased levels of deoxygenated hemoglobin, validating the use of fNIRS to procure meaningful information about cognitive state**. Additionally, the results indicate that participants were generally well-tuned to their own cognitive processes and accurately externalized their cognitive load. We discuss the implications of this observation in the following section.

Indistinguishable Performance Between Graphs

A comparison of NASA-TLX responses and speed and accuracy demonstrates a dissociation between performance and cognitive state during the use of bar graphs and pie charts. Performance measures on both graphs were statistically indistinguishable from each other, regardless of whether participants found one graph to be more mentally demanding. However both questionnaire responses and fNIRS readings showed that the two designs influenced brain activity differently.

Given these results, it is possible that participants were exerting different amounts of mental effort on a given graph to achieve the same levels of performance. Furthermore, this observation suggests that evaluating performance metrics without considering cognitive state might have led to different conclusions about the efficacy of bar graphs and pie charts in this experiment. In the next section, we investigate whether the fNIRS signals we observed reflect levels of mental demand.

N-BACK TASK: DETECTING MENTAL WORKLOAD

During the course of this paper, we have been intentionally ambiguous about assigning a specific cognitive state to our fNIRS readings. The brain is extremely complex and it is dangerous to make unsubstantiated claims about functionality. However, for fNIRS to be a useful tool in the evaluation of visual design, there also needs to be an understanding of *what* cognitive processes fNIRS signals may represent. In our experiment, we have reason to believe that the signals we recorded correlate with levels of mental demand. We share three legs of evidence that support this claim:

1. fMRI studies have suggested that decreases in deoxygenated hemoglobin are indicative of increased brain activity [9]. Active regions of the brain require more oxygen to function. Thus, as levels of oxygenated hemoglobin increase to meet these demands, levels of deoxygenated hemoglobin decrease.

2. Self-reports of mental demand from the NASA-TLX results during the bar-graph and pie chart task correlated with levels of deoxygenated hemoglobin. Graphs that were reported to require more mental effort were accompanied by lower levels of deoxygenated hemoglobin.

3. We ran each participant on a well-characterized working memory task from the psychology literature - the visuospatial n-back test - and found that brain activity in the more mentally demanding graph mirrored activity in the more demanding n-back condition. We discuss the details of this experiment in the next section.

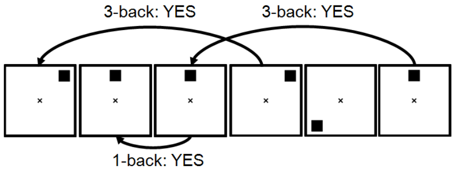

Figure 7. In the visuospatial n-back task, participants view a series of slides and respond whether the current pattern matches the pattern from n slides ago. We show positive answers for both the 1-back and 3-back conditions.

Methods

In the n-back task, participants were shown a series of slides, each with a distinct visual pattern, and asked whether the current slide matched the pattern from either 1 slide previously (1-back) or 3 slides previous to the current slide (3-back). Thus, the 3-back task strains the participant's visuospatial working memory by forcing him or her to constantly remember (and update) 3 images at once. By comparison, the 1-back task is relatively simple, requiring participants to remember only visual pattern from the previous slide.

Figure 7 shows an example of 6 slides from the n-back test. For each slide, the visual pattern remained on the screen for 300ms followed by a blank response screen for 1550ms in which participants answered 'yes' or 'no' using a single keystroke. Participants were given 8 trials of each condition with each trial consisting of 22 slides. Each trial lasted for 40.7 seconds and trials were separated by 12-second rest periods. This experimental timing mirrors the timing in the bar graphs/pie charts task, enabling us to compare equal slices of time for the fNIRS data.

Results

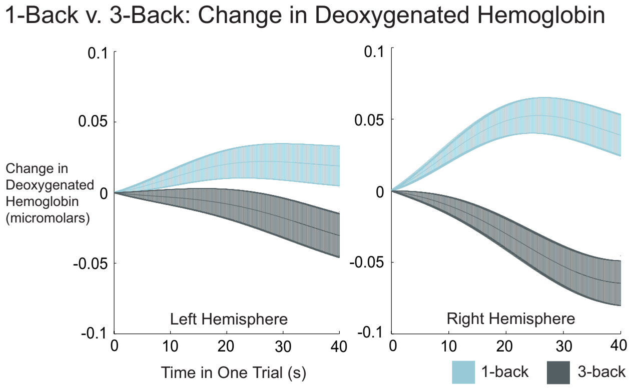

Figure 8. The mean fNIRS signal across all 16 participants in the Baseline Task. We see a clear separation between the 1-back and 3-back conditions participants. The more demanding 3-back condition mirrors signals from the graph design that participants believed was more mentally demanding.

Looking at the results, Figure 8 shows that there is a clear distinction between 1-back (blue) and 3-back (black) trials. These results are expected and resemble previous studies of the n-back task [21]. Additionally, the 3-back task induced lower levels of deoxygenated hemoglobin, agreeing with other observations of deoxygenated hemoglobin from the fMRI literature.

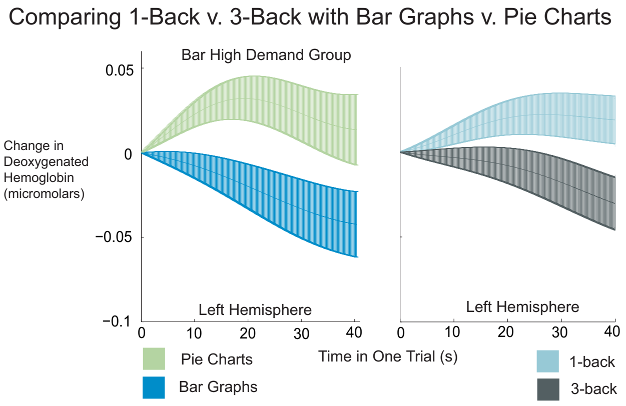

Figure 9. An example of comparing the n-back signal with those recorded in the bar graph v. pie chart experiment. The signal recorded during the more demanding 3-back resembles the signal recorded during bar graphs for the bar high demand group - participants who found bar graphs to be more mentally demanding than pie charts.

When placed side-by-side with the fNIRS readings from out bar graph/pie chart task, we notice that signals from the more mentally demanding 3-back resemble those from the graph that participants identified as requiring more mental effort (Figure 9). Similarly, the signal recorded from the less-demanding 1-back task resembles those observed in the graph that participants identified as requiring less mental effort (Figure 9).

Given these three legs of evidence - previous observations noted in fMRI studies, correlations with survey data, and correlations with signals observed in the n-back task - we feel confident that the fNIRS signals observed during use with bar graphs and pie charts correlate with mental demand in the brain. Furthermore, these results suggest that fNIRS can be used to monitor mental demand in other visual interfaces.

FNIRS: CONSIDERATIONS FOR EVALUATION

We have shown that we can successfully differentiate fNIRS signals during the interaction of bar graphs and pie charts in a complex task and that these signals likely indicate workload in the brain. In this section, we synthesize our results, previous literature, and an auxiliary study to explore *when* fNIRS is an appropriate tool for the evaluation of visual design.

Are Surveys Good Enough?

Cognitive state is often overlooked in evaluation, partially because it is difficult or cumbersome to quantify. We found that a simple survey agreed with fNIRS readings and accurately captured the participant's mental workload. This is good news for simple evaluations of mental demand. Questionnaires do not require an unreasonable time investment, and the strength of our observations were based on a single dimension in the NASA-TLX questionnaire. If more objective measures are not available, questionnaires can provide insight into a user's cognitive state. Nonetheless, questionnaires can be problematic as they depend on the assumption that people can sense and externalize their subjective feelings without being biased by external influences [8, 22]. In comparison, brain sensing provides an objective snapshot of cognitive state and short-cuts the rating process by directly measuring the brain *during* interaction. As opposed to post-hoc questionnaires, neurophysiological measures require no additional effort or time from the participant. Furthermore, physiological measures can be used in more complex or time-consuming tasks for fine-grained observations of cognitive processes. Instead of a single workload metric for the entirety of a task, physiological measures can provide time-sensitive evaluations, potentially identifying periods of mental demand. We recommend that visualization researchers carefully weigh the nature of their comparison to select an appropriate technique.

Lending Insight to Complex, Analytical Tasks

Given the results of our study, we suggest that fNIRS may be well-suited for the analysis of complex interactions that are common in visual analytic systems. In this section, we highlight three other factors that point to fNIRS being well-suited for analytical tasks:

- The extended timeline of complex tasks mitigates the slow temporal resolution of fNIRS, which occurs because of the delayed (5-7 seconds) physiological response to brain activity.
- The PFC - the region of the brain that fNIRS most easily measures - has been posited to "integrate the outcomes of two or more separate cognitive operations in the pursuit of a higher behavioural goal" [25]. These higher-level cognitive functions typically drive analytical thought and include (but are not limited to) selection, comparison, the organization of material before encoding, task switching, holding spatial information 'online', and introspective evaluation of internal mental states [25, 27].
- The successful examples of applying fNIRS measures to interface evaluation have traditionally leveraged mentally demanding scenarios such as multi-tasking the navigation of multiple robots [29], increasing the difficulty of a video game [11], or reversing the steering mechanism in a driving task [15].

Given these factors, we believe that fNIRS will provide the most insight to visual interfaces that require complex, analytical thought. However, fNIRS is not without its limitations; as we demonstrate in the next section, short, low-level tasks are difficult to detect using fNIRS.

Perceptually-Driven Tasks are Difficult to Monitor

To explore the limits of using fNIRS to evaluate visual interfaces, we constructed an experiment that is closer to Cleveland & McGill's original comparison of position v. angle, which is based on more perceptually-driven interactions. Whereas trials in our previous experiment required participants to make percentage comparisons in graphs across slides, a trial in this modification consisted of 4 percentage comparisons (3.75 seconds per comparison) on the same

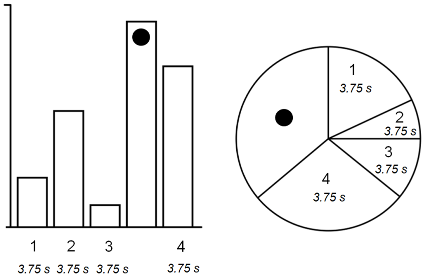

Figure 10. Participants sequentially compared elements of a graph to the largest element of the graph.

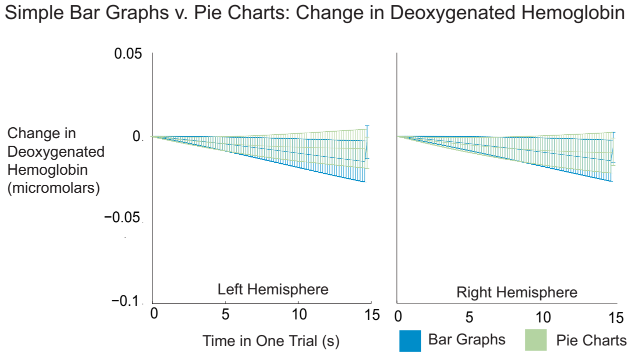

Figure 11. The mean fNIRS signal across all 8 participants in a simple bar graphs and pie charts task. The lack of activation shows that fNIRS may be less suited for simple, perceptual comparisons.

graph and participants interacted with 12 trials of bar graphs and 12 trials of pie charts. Thus, for each trial, four small pieces on a graph were sequentially compared to the largest piece in the graph (Figure 10).

To compare the changes in deoxygenated hemoglobin with our previous study, we ran an additional 8 participants and plotted the fNIRS signal using the axis of the same scale as the complex task. Looking at Figure 11, we can see that both pie charts and bar graphs caused very little activation in the PFC, with little to no differentiation between signals.

These results are not surprising. Quick visual and perceptual tasks are not likely to be observed by fNIRS. Tasks that rely heavily on preattentive processing use very little of the processing power of the PFC. Additionally, it takes a couple of seconds to observe the hemodynamic response resulting from brain activity, and 5-7 seconds in total for the oxygen levels to peak in the brain. This means that we are unlikely to observe quick and subtle interactions with a visualization. We therefore recommend that fNIRS will lend the most insight during more complex analytical interactions.

FINDINGS AND FUTURE WORK

We have demonstrated that fNIRS is a viable technology for investigating the impact of visual design on a person's cognition processes. Using the classical comparison of bar graphs and pie charts, we found that decreasing levels of deoxygenated hemoblogin correlated with the visual form that participants found to be more mentally demanding. We suggest that these changes in deoxygenated hemoglobin, detected in the PFC, indicate the amount of mental effort associated with the visual design. As we demonstrated in our study, these differences in workload are not necessarily reflected in traditional performance metrics.

Exploring the use of fNIRS in visualization research, we suggested that fNIRS is well suited for the evaluation of visual interfaces that support analytical reasoning tasks. This advantage should be particularly appealing for interface designers, as the complexity of visual analytic systems often make it difficult to apply traditional performance metrics. Additionally, the resistance of fNIRS sensors to movement artifacts allows users to interact naturally with an interface, resulting in more ecologically sound evaluations.

Lowering the barrier to monitor cognitive state increases the opportunity to develop adaptive applications that specially calibrate the display of information to the individual. Recently, Solovey et. al [29] used fNIRS to determine *when* the user should be interrupted with new information and built a system that adapted the level of automated assistance in a virtual robot navigation task. While recent work in visualization has begun to pay careful consideration to the impact of a user's personality and cognitive traits, using tools like fNIRS, we hope that visual interfaces can be designed to also be attentive to the user's current cognitive state.

The strengths of fNIRS are appealing, however, there are also limitations. While we identified periods of high or low workload, more specific mappings of fNIRS signals to cognitive states are needed to promote fine-grained evaluations of visual interfaces. Additionally, we found that fNIRS is less suited for quick visual tasks that are driven by the user's perceptual system. Despite these drawbacks, fNIRS provides a suite of benefits that are distinctive and complimentary to those offered by other physiological sensors. With the decreasing cost of brain sensing technology and its increasing use in HCI, we believe that the door has finally opened to directly explore the impact of visual design on cognitive state.

ACKNOWLEDGMENTS

We thank Dan Afergan, Andrew Jenkins, Francine Lalooses, and Garth Griffin who are students in the HCI group at Tufts; Sergio Fantini and Angelo Sassaroli from the Biomedical Engineering Department at Tufts; Jordan Crouser from the VALT group at Tufts; and Erin Solovey from the Humans and Automation Lab at MIT. We thank the National Science Foundation (grant nos. IIS-0713506, IIS-1065154, IIS-1218170) for support of this research.

REFERENCES

1. Anderson, E., Potter, K., Matzen, L., Shepherd, J., Preston, G., and Silva, C. A User Study of Visualization Effectiveness Using EEG and Cognitive Load. In *EuroVis 2011*, vol. 30, Wiley Online Library (2011), 791–800.

2. Ayaz, H., Shewokis, P. a., Bunce, S., Izzetoglu, K., Willems, B., and Onaral, B. Optical brain monitoring for operator training and mental workload assessment. *NeuroImage 59*, 1 (2012), 36–47.

3. Bertini, E., Perer, A., Plaisant, C., and Santucci, G. BEyond time and errors: novel evaLuation methods for Information Visualization. *BELIV* (2010), 4441–4444.

4. Causse, M., and Hurter, C. The physiological users response as a clue to assess visual variables effectiveness. *Human Centered Design* (2009), 167–176.

5. Chance, B., Anday, E., Nioka, S., Zhou, S., Hong, L., Worden, K., Li, C., Murray, T., Ovetsky, Y., Pidikiti, D., and Thomas, R. A novel method for fast imaging of brain function, non-invasively, with light. *Optics express 2*, 10 (1998), 411–23.

6. Cleveland, W. S., and McGill, R. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association 79*, 387 (1984), 531–554.

7. De Waard, D. *The measurement of drivers' mental workload*. PhD thesis, 1996.

8. Dell, N., Vaidyanathan, V., Medhi, I., Cutrell, E., and Thies, W. Yours is Better! Participant Response Bias in HCI. In *ACM CHI 2012* (2012), 1321–1330.

9. D'Esposito, M., Zarahn, E., and Aguirre, G. Event-Related Functional MRI: Implications for Cognitive Psychology. *Psychological bulletin 125*, 1 (1999), 155–164.

10. Fairclough, S. H. Fundamentals of Physiological Computing. *Interacting with Computers 21* (2009), 133–145.

11. Girouard, A., Solovey, E., Hirshfield, L., Chauncey, K., Sassaroli, A., Fantini, S., and Jacob, R. J. Distinguishing difficulty levels with non-invasive brain activity measurements. *Human-Computer Interaction INTERACT 2009* (2009), 440–452.

12. Girouard, A., Solovey, E. T., Hirshfield, L. M., Peck, E. M., Chauncey, K., Sassaroli, A., Fantini, S., and Jacob, R. J. From Brain Signals to Adaptive Interfaces : using fNIRS in HCI. 2010, 221–237.

13. Hart, S. G., and Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload* (1988), 139–183.

14. Heer, J., and Bostock, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. *ACM CHI 2010* (2010), 203–212.

15. Hirshfield, L. M., Gulotta, R., Hirshfield, S., Hincks, S., Russel, M., Ward, R., Williams, T., and Jacob, R. J. K. This is Your Brain on Interfaces : Enhancing Usability Testing with Functional Near-Infrared Spectroscopy. In *ACM CHI 2011* (2011), 373–382.

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 with Functional Near Infrared Spectroscopy. In *ACM CHI 2009* (2009), 2185–2194.

17. Huang, W., Eades, P., and Hong, S.-H. Measuring effectiveness of graph visualizations: A cognitive load perspective. *IEEE Transactions on Information Visualization 8*, 3 (2009), 139–152.

18. Hullman, J., Adar, E., and Shah, P. Benefitting InfoVis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (2011), 2213–22.

19. Lloyd-Fox, S., Blasi, A., and Elwell, C. E. Illuminating the developing brain: the past, present and future of functional near infrared spectroscopy. *Neuroscience and biobehavioral reviews 34*, 3 (Mar. 2010), 269–84.

20. Moroney, W. F., Biers, D. W., Eggemeier, F. T., and Mitchell, J. A. A Comparison of Two Scoring Procedures with the NASA Task Load Index in a Simulated Flight Task. *National Aerospace and Electronics* (1992), 734–740.

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

22. Paas, F., Tuovinen, J. E., Tabbers, H., and Van Gerven, P. W. M. Cognitive Load Measurement as a Means to Advance Cognitive Load Theory. *Educational Technology 38*, 1 (2003), 63–71.

23. Paas, F. G., and Van Merriënboer, J. J. The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 35 (1993), 737–743.

24. Presacco, A., Goodman, R., Forrester, L., and Contreras-Vidal, J. Neural decoding of treadmill walking from noninvasive electroencephalographic signals. *Journal of Neurophysiology 106*, 4 (2011), 1875–87.

25. Ramnani, N., and Owen, A. M. Anterior Prefrontal Cortex: Insights into Function from Anatomy and Neuroimaging. *Nature reviews: Neuroscience 5*, 3 (2004), 184–94.

26. Riche, N. Beyond system logging: Human logging for evaluating information visualization. In *BELIV* (2010).

27. Rowe, J. B., Toni, I., Josephs, O., Frackowiak, R. S., and Passingham, R. E. The prefrontal cortex: response selection or maintenance within working memory? *Science 288*, 5471 (2000), 1656–60.

28. Simkin, D., and Hastie, R. An Information-Processing Analysis of Graph Perception. *Journal of the American Statistical Association 82*, 398 (1987), 454–465.

29. Solovey, E., Schermerhorn, P., Scheutz, M., Sassaroli, A., Fantini, S., and Jacob, R. Brainput: Enhancing Interactive Systems with Streaming fNIRS Brain Input. In *ACM CHI 2012* (2012), 2193–2202.

30. Solovey, E. T., Girouard, A., Chauncey, K., Hirshfield, L. M., Sassaroli, A., Zheng, F., Fantini, S., and Jacob, R. J. K. Using fNIRS Brain Sensing in Realistic HCI Settings: Experiments and Guidelines. In *ACM UIST 2009* (2009), 157–166.

31. Solovey, E. T., Lalooses, F., Chauncey, K., Weaver, D., Scheutz, M., Sassaroli, A., Fantini, S., Schermerhorn, P., and Jacob, R. J. K. Sensing Cognitive Multitasking for a Brain-Based Adaptive User Interface. In *ACM CHI 2011* (2011), 383–392.

32. Spence, I. Visual psychophysics of simple graphical elements. *Journal of Experimental Psychology: Human Performance and Perception 16* (1990), 683–692.

33. Spence, I. No Humble Pie: The Origins and Usage of a Statistical Chart. *Journal of Educational and Behavioral Statistics 30*, 4 (2005), 353–368.

34. Spence, I., and Lewandowsky, S. Displaying proportions and percentages. *Applied Cognitive Psychology 5* (1991), 71–77.

35. Strangman, G., Culver, J. P., Thompson, J. H., and Boas, D. a. A Quantitative Comparison of Simultaneous BOLD fMRI and NIRS Recordings during Functional Brain Activation. *NeuroImage 17*, 2 (2002), 719–731.

36. Villringer, A., and Chance, B. Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neurosciences 20*, 10 (1997), 435–42.

37. Wickens, C., and Hollands, J. *Engineering Psychology and Human Performance*. Prentice-Hall, Upper Saddle River, NJ, 1999.

38. Wigdor, D., Shen, C., Forlines, C., and Balakrishnan, R. Perception of elementary graphical elements in tabletop and multi-surface environments. *ACM CHI 2007* (2007), 473–482.

39. Wolf, M., Ferrari, M., and Quaresima, V. Progress of near-infrared spectroscopy and topography for brain and muscle clinical applications. *Journal of Biomedical Optics 12*, 6 (2012), 062104.

40. Yeh, Y., and Wickens, C. D. Dissociation of Performance and Subjective Measures of Workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society 30*, 1 (1988), 111–120.