The Dissertation Committee for Evan M. Peck

certifies that this is the approved version of the following dissertation:

Designing Brain-Computer Interfaces for Intelligent Information Delivery Systems

Committee:

Robert J.K. Jacob, Supervisor

Remco Chang

Benjamin Hescott

Lennart Nacke

Sergio Fantini

Designing Brain-Computer Interfaces for Intelligent Information Delivery Systems

A dissertation

submitted by

Evan M. Peck, B.S., M.S.

In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Computer Science

TUFTS UNIVERSITY

August 2014

Advisor: Robert J.K. Jacob

While traditional interfaces use static designs that are meant to reach as many people as possible, research shows that a user's moment-to-moment state can impact interaction, either positively or negatively. In this thesis, I show that brain sensing can be used as passive input to intelligent information delivery systems. Using functional near-infrared spectroscopy (fNIRS) as a lightweight brain sensor, I present work to design brain-computer interfaces (BCIs) that analyze brain data and classify user state in real time. These systems react to user state by modifying the flow of information and measurably improving user performance. I describe a brain-driven recommendation system that changes which information is prioritized to the user, an intelligent interruption system that uses brain data to determine opportune moments of interruption, and a study that demonstrates the brain's sensitivity to visual design. Finally, I discuss design strategies for building robust online systems that adapt to physiological input. I suggest that someday our computers may have the capability of being personally attentive to us — optimizing when information is delivered, which information is prioritized, and how information is presented.

During the past six years, an enormous number of people have moved through my life. It is impossible to thank them all, but this work is for them.

First and foremost, I want to thank my family. My wife, Lauren, navigated many late nights and lost weekends. Her support was invaluable. Coming home to my son, Grayson, made the longest days end with a smile. I am grateful for the years of support from my parents and sister. I cannot thank them enough.

This would not be possible without my advisor, Robert Jacob. His encouragement and advice made the graduate journey markably easier. I am incredibly fortunate to have him as an advisor. I would also like to thank Remco Chang, who spent more than a few late nights turning my rough ideas into polished projects.

Many have also shared their time at Tufts working with me. They have been both my colleagues and friends, and their support has been irreplaceable. In no particular order: Dan Afergan, Jordan Crouser, Francine Lalooses, Beste Filiz, Eli Brown, Alvitta Ottley, Erin Solovey, Audrey Girouard, Leanne Hirshfield, Michael Horn, Sam Hincks, Tomoki Shabati, Lane Harrison, and more. I am eager to call all of these people my friends and colleagues for years to come.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Brain-Computer Interfaces: From Alternative to Augmentative Input

An interface is said to be a **brain-computer interface** if it allows a user to communicate with external devices through thought processes alone [36]. For much of their history, brain-computer interfaces (BCI) have used brain signals as an *alternative* input to the computer. Rather than moving, clicking a mouse, or typing on the keyboard, *direct-control* (or *active*) BCIs use the brain as a primary input mechanism to the computer, substituting for other input mechanisms. Direct-control BCI is often used for a user without full motor capabilities, making it difficult or impossible for them to use a mouse or keyboard. For these users, the brain may be one of few available input channels to the computer [175]. BCIs have been used as keyboard-substitutes [181], controls for wheelchairs [56], and even to direct robotic limbs [15]. These are powerful applications that directly allow users to accomplish tasks that previously may have been impossible.

As brain sensors have become cheaper and more lightweight, there has a been a push towards integrating them into the consumer space. However, broadening the user base necessitates a reimagining of the use of brain-computer interfaces [18]. Direct-control BCIs are often cumbersome and slow, relying mental triggers that

interfere with normal interaction (i.e., a user thinks about moving their right arm to move the mouse to the right side of the screen). Mouse movement or keyboard use is both easier and faster for users with full motor capabilities. For the short term at least, direct-control BCI has little utility for the everyday worker.

In contrast to direct-control BCI, new work has begun to consider how to use the brain as an *augmentative* input to the computer. Rather than using the brain as a replacement to an existing input modality, *passive brain-computer interfaces (pBCI)* use the natural thought processes of users to supplement existing interactions. For example, instead of performing explicit mental triggers to mute our phones while driving, a passive BCI might detect that user is busy with driving and automatically filter incoming phone calls and text messages.

However, there is currently very little work that considers BCI applications for the everyday user. Most brain sensing technology is not well suited for the working environment, enforcing strict movement constraints on the user in order to reduce the impact of motion artifacts. In addition, monitoring physiological data presents problems to interface designers. Brain data is often noisy and difficult to interpret, making the automatic detection of user states a non-trivial task.

Despite these challenges, using the brain as a passive input has the significant advantage that it does not interfere with users' normal behavior. Acting purely as an augmentative input is particularly beneficial for applications such as information delivery, in which minor interruptions or perterbations may adversely impact a worker's performance. Ideally, a computer could monitor users' brain activity and predict the best information to give them at any given moment, much in the same way that people observe and react to the social cues [128].

1.2 BCIs for Information Delivery

People consume information at many different times and in many different ways. We read news articles, shop for products online, monitor information streams, stay updated with our social networks, receive text messages, sift through emails, and

monitor instant messages. However, a 2012 Pew Internet Survey found that 25% of smartphone owners believe their device makes it more difficult to focus on a task without being distracted, and 67% of cell owners find themselves checking their phone for notifications even when they dont notice their phone ringing or vibrating [149]. While the advent of new information delivery services and devices has expanded our capacity to access timely information, our biologocal capacity to see and understand information has not changed. Attention has become a scarce resource.

The consequences of misappropriating attention are well documented. Consuming too much information, consuming the wrong type of information, or consuming information at the wrong moment can not only lead to a decrease in working performance, but negatively impact stress and anxiety [10, 24, 149]. For example, a recent analysis of 10,000 programming sessions found that the average programmer takes 10-15 minutes to continue editing code after a single interruption [121]. A survey issued in the early 2000s even went so far to suggest that the stress of information overload can negatively impact personal relationships or perceived health [13].

Complicating matters, information accessibility is also rapidly increasing, leaving similarly increasing demands on the user. Wearable computing devices such as Google Glass can move information delivery from locations that are peripheral to the user to a visual location that immediately draws attention. This tension between an increasing quantity of information and the increasing accessibility of that information leaves technology at a crossroads: **As information increases, how can technology prevent users from becoming overwhelmed by information?** *When* is the best time to give someone new information so that it won't be disruptive? *Which* information is most helpful (or unhelpful) to us at any given moment? *How* should information be delivered to users that makes the most sense to them?

But humans are not inherently poor at handling information. In fact, we already navigate and share a wealth of information in social interactions that *don't*

come from their computer devices. Our brains are wired to attend to the people we interact with, and communicate with them in a way that is both effective and appropriate. For example, research has shown that when teachers detect that their students are bored, they change the way they teach: increasing the physical space of their gestures and modifying the tone of their voice to regain the attention of their students [31]. However, this adjustment only works because the instructor observes a less-than-ideal state in his or her students. The students' engagement improves only because their teachers adapt to their needs in real time.

Our devices miss these non-verbal cues. They do not understand when users are bored or excited or happy or frustrated. Consequently, they cannot adapt to keep the user interested or adjust to prevent the user from becoming overwhelmed. If the computer can be considered a collaborator, it is a collaborator which frequently breaches the social rules we apply to each other. While a computer infringing on social norms may seem trivial, research has suggested that people unintentionally treat computers as social entities [113, 112], and that 'polite computers' can significantly improve user interaction and performance [120, 115, 169, 176, 177]. In fact, frameworks that describe emerging interaction techniques frequently suggest that they should be 'natural', or grounded in our real-life experiences [88, 172].

The source of these miscommunications echo a fundamental problem in the field of human-computer interaction (HCI): the amount (or quality) of information that the computer has about us pales in comparison to what people naturally absorb about each other. Put succinctly, the bandwidth of communication between computers and computer users is too small to construct an attentive computer. Just as the speed at which people navigate the web is limited by their internet bandwidth, the information that computers access in order to respond to users is limited by the bandwidth of its input devices. Relying exclusively on input from a keyboard or a mouse or even a touchscreen falls far short of the full descriptive range of human expression. The computer simply does not have enough information about its user to react in the same way that people react to each other. Until the computer sufficiently understands the user and his or her context, the delivery of information will

continue to be a significant challenge.

To help address this problem, researchers have investigated new methods of communicating with the computer. Tangible computing, multitouch screens, and gesture recognition, for example, have all allowed users to communicate with their devices in ways that were not possible in the past. And while these new input modalities have largely been successful, the bandwidth still leaves much to be desired. It is with this motivation in mind that we return to monitoring the brain and a powerful source of input to the computer.

In one of the first published papers that considered the use of brain sensing in Human-Computer Interaction, Cutrell and Tan [39] write the following:

> *In many ways a brain-computer interface is the holy grail of HCI research. The idea of direct neural control of computational systems goes to the heart of what HCI researchers and designers are all about: creating usable and useful systems that are intuitive and just work like a user wants them to. One definition of HCI is to improve the impedance match between computer systems and the people that use them, and BCIs are the epitome of that goal.*

The prospect of building an interface, or an information delivery system, that always delivers information when we expect and and how we would expect is an appealling one. However, as the Cutrell and Tan mention later in the paper, "reality is a bit more sobering; we are a long way from the Matrix"[39]. The brain is complex, neural data from sensors is noisy, and users' natural working environments often introduce complex and difficult-to-control variables. As a result, while brain-computer interfaces have existed for decades, the translation of this technology into everyday life has largely remained an idea for the future.

In this thesis, I investigate *functional near infrared spectroscopy (fNIRS)* as an alternative brain sensing device that circumvents some of the traditional problems of other brain sensing technology, and consider its potential as input to information delivery applications. fNIRS is an optical brain sensing device that has the poten-

tial to lend insight to working-place interactions [26, 129, 167, 174], observing the same physiological parameters as fMRI. It is relatively tolerant to minor movement artifacts and generally valued for its potential to attain ecologically valid measurements. While there are many advantages and considerations to the use of fNIRS (many of which will be communicated in the following chapter), in the context of passive BCI, it remains a largely unexplored technology.

Examples of using fNIRS as passive input to intelligent systems are few and far between, and there has been no work that directly connects fNIRS input to information delivery. As a result, there is a poor understanding of whether fNIRS can serve as input to information delivery systems. Is fNIRS capable of identifying information that is beneficial to the user - *which* should be emphasized, *when* should it be delivered, and *how* should it be presented? How should this unique, but potentially noisy, input be used in the design of intelligent systems?

1.3 Purpose and Outline of this Work

The high-level goal of this thesis is twofold: First to expand the bandwidth of interaction between the user the computer through the use of passive brain sensing, and more specifically, functional near-infrared spectroscopy (fNIRS). By using fNIRS to monitor user's mental state, we enable the computer to access cues that are closer to those that humans identify in each other. Second, to demonstrate that our computing devices can utilize these cues to deliver information to us in a way that is specially calibrated to our unique moment-to-moment state.

In this thesis, I attempt to answer the following high-level questions: **Can the brain be used as passive input to an intelligent information delivery system? Can a brain-computer interface personalize how information is presented, which information is prioritized, and when it is delivered?**. To move towards this goal, I structure the remainder of this thesis in the context of the following contributions:

- In Chapter 3, I use the field of information visualization to show that brain

sensing can be used to capture cognitive differences that relate to *how* information is presented [130, 125]. I also discuss the importance understanding individual cognitive differences to evaluate information visualization, proposing a framework of individual cognitive differences [131].

- In Chapter 4, I demonstrate that brain sensing can be used as passive input to a system that personalizes *which* information is presented to the user [126]. In this study, fNIRS is used as the sole input to an information-filtering system which provides users with improved recommendations over time.

- In Chapter 5, I show that fNIRS can be used as passive input to an intelligent interruption system that combines neural markers of *which* information is relevant, and *when* a person is capable of handling an interruption. In addition, I outline the potential of an intelligent information delivery system that optimizes the content, presentation, delivery timing of information.

- Finally, I discuss the strategies that enable brain-computer interfaces to improve user performance despite potentially noisy data. In particular, I focus on applications in which users do not have strong expectations of incoming information, and graded adaptation strategies that depend on the integrity of the brain input.

Taken together, I suggest that brain-computer interfaces, and more generally physiological data, have the potential to one day form the backbone of intelligent information delivery systems.

Chapter 2

Related Work

The design and construction of a brain-computer interfaces often requires drawing on knowledge from multiple fields: neuroscience, biomedical engineering, human factors, computer science, and behavioral psychology. As a result, the space of potentially related work is enormous. However, in this section we focus primarily on the application of brain-computer interfaces to everyday users. In particular, we first survey the existing state of passive brain-computer interfaces alongside some of the challenges of constructing implicit interfaces. Finally, we review the analysis and application of functional near infrared spectroscopy within the context of human-computer interaction [127].

2.1 Passive Brain-Computer Interfaces

Contrary to BCIs that are used as explicit (and often alternative) input, passive brain-computer interfaces take advantage of naturally occuring signals in the brain. Zander et al. [183] identifies three significant characteristics that distinguishes passive approaches from direct-control BCI:

- **Complementarity:** passive input does not interfere with the natural thought processes of users.
- **Composability:** an application can use passive brain input in parallel with

other passive inputs with no conflict. For example, brain sensing, heart rate monitoring, and keyboard input can all be monitored and integrated in an intelligent system.

- **Controlled Cost:** since there is no cognitive cost to operate a passive BCI, the cost of using a passive BCI is impacted only by system misclassifications. As a result, it is possible to design systems that provide zero benefit in the worst case (as opposed to a negative impact) [183]. Throughout this thesis, we frequently reference this challenge to design implicit adaptations that are no cost in the worst case.

Despite the differences between direct-control BCI and passive BCI, their fundamental architecture (or *biocybernetic loop*) has much in common with each other, as well as with other systems that are driven by physiological input. The biocybernetic loop is initiated by sensor information (in this case, from the brain), where the system applies filters to reduce noise and extract relevant features of the signal. Then the filtered data is mapped to a user state such as workload or emotion. This can be done either through simple threshholds, or more complex algorithmic solutions. Finally, the system chooses to respond to this mapping (the exact conditions for whether the system responds is called the *adaptive trigger*) through a designed *adaptive mechanism*. Once the system's response is experience by users, their second-order experience may be translated to a physiological response, thus completing the biocybernetic loop. Later in this chapter, we expand on the details of this process as it is related to fNIRS.

Returning back to the bandwidth problem in HCI, a passive source of information can serve as an augmentative source of input that improves and expands the bandwidth between computers and users. In fact, Cutrell and Tan[39] suggest that brain sensing may epitomize a central goal of HCI: creating an "impedence match" between people and their computer devices.

2.1.1 Examples and Applications of Passive BCI

While the notion of using the brain as passive input has been discussed and referenced for decades, their proper use and application is still largely considered to be an open question. In particular, there are few unifying frameworks or design guidelines that tie together existing prototypes. In 2010, George and Lecuyer [57] surveyed existing research on passive BCI and classified them into four primary categories:

1. **Adaptive Automation:** Refers to a switch in control of some aspect of the system from manual to automatic (and back). In Solovey et al. [150], a robot switched from manual navigation to automatic navigation based on neural metrics related to multitasking. We use this as a case study for fNIRS in BCI in a following section. For other examples, see [3, 93, 133].

2. **Multimedia Content Tagging:** Neural predictions of user state (often emotion) are used to tag multimedia content, such as music or video for later use. This approach has also been used in contexts in which analysts are looking for a specific objects in images. Identifying the object triggers a neural response that allows a system to automatically tag a response faster than manual tagging would allow [89, 99].

3. **Video Games:** Brain activity is mapped to some variable in a video, such as music [59] or the shape of an avatar [65].

4. **Error Correction:** Taking advantage of the quick temporal response of EEG, these BCIs attempt to identify moments when users realize an error and automatically correct it for them. For example, [104, 144, 165, 182]

Other categorizations of passive BCI shift the focus from application domains to user state detection. For example, Molina et al. [107] proposes *emotion BCI* in which computing devices react to naturally occuring emotions by users. Zander proposes a similar brand of passive BCIs that detect *affective covert user states* [182]. While BCIs that perform multimedia content tagging often take advantage of emotion state detection, these definitions broaden the scope of emotion BCIs beyond

this specific application area to include systems that respond to user frustration [182], or attempts to more accurately predict user behavior [107].

It's worth noting that passive BCI shares a close neighbor to systems that are driven by physiological measures. In fact, the umbrella term *physiological computing systems (PCS)* is increasingly being used to reference applications that accept either brain or body sensors as input to the system. This broadening of the space encorporates a number of systems that have investigated the impact of physiological metrics on games [46, 110, 111], the presentation of content [69, 178], and adaptive automation [19, 97] just to list a few. In addition, other categorizations arise when considering physiological input. For example, Gilleade and Dix [58] proposed the following classification of game adaptations that can broadly be applied to other applications as well:

- **'Assist-Me':** adaptations attempt to limit the user's frustration with a game.
- **'Challenge-Me':** adaptations attempt to optimize the level of challenge a user experiences in a game (not unlike the adaptive automation BCIs desribed above)
- **'Emote-Me':** adaptations that are meant to induce an emotional response (such as fear in a horror game).

We can also consider physiological computing systems within the framework of this thesis: when information is delivered, how it is presented, and which information is delivered. Adaptive automation mediates *when* the system hides a layer of complexity to the user (through automation). Similarly, content tagging moves towards interfaces that manipulate *which* information is presented. However, none of the referenced work uses its neural tags to improve the user's experience with the system, leaving an incomplete biocybernetic loop. We will present the first full implementation and evaluation of this idea in chapter 4. Finally, Szafir et al. [159] constructs of one of the few examples of a BCI that manipulates *how* information is presented. In their work, students are tutored by a robot that adapts its pre-

sentation of a story based on the students' level of engagement (deteced by EEG). We propose the application of this idea to information visualization, a field that places significant emphasis on optimizing the representation of information, and in chapter 3, show that fNIRS can detect changes in cognitive state that derive from visual design.

2.1.2 Building Implicit, Adaptive Systems

To this point, we have discussed passive BCI primarily from the perspective of application domains without discussing the details of the system's adaptations. However, designing how the system responds to passive input is not trivial [47]. Critics of adaptive interfaces contend that changing a system's behavior 'on the fly' may result in making the system unpredictable and inconsistent, negatively impacting performance [49, 146]. Fairclough warns that if researchers ignore the complexity of these challenges, "the performance of prototype systems will be erratic and unreliable, which runs the risk of premature abandonment" [47].

As a result, design needs to minimize these disruptions to a user's mental model. For example, Gajos et al. [55] found that using split interfaces to duplicate (rather than move) functionality can be used as an adaptive mechanism that offers 'medium to high benefits while causing minimal confusion'. We adopt the phrase *implicit interfaces* to suggest system adaptations that are subtle and gentle, nudging the user towards a predicted goal.

Although we have motivated information delivery as potentially a compelling application of brain-computer interfaces, it is also is well suited to these design goals. Incoming data streams (whether it is email, Twitter, etc.) consistently deliver new information to users. Since the timing and content of this information is largely outside of the user's control, there are characteristics about the data that can be modified without severely disrupting the user's mental model of the system.

Assessing the impact of an adaptive system can also be difficult. Most commonly, they are judged on three primary variables: usability, perceived usefulness, and appropriateness of adaptation [163]. However, in a survey of 63 adaptive sys-

tems, Van Velsen found that there was significant diversity in the measures that researchers collected, organizing them in five broader categories [163]:

- **Attitude and Experience:** appreciation, trust and privacy issues, user experience, user satisfaction
- **Actual Use:** usability, user behavior, user performance
- **System Adoption:** intention to use, perceived usefulness
- **System Output:** appropriateness of adaptation, comprehensibility, unobtrusiveness

In fact, many studies find usability tradeoffs in adaptive interfaces [124]. For example, Gajos explored tradeoffs between accuracy and *predictability* (what did the user expect to happen?) in adaptive menus [55]. While many studies rely on *accuracy* (how well the system inferred the user's state) as a primary measure of success, it is a measure that does not necessarily reflect the user's experience or efficiency of the system [95].

To further complicate evaluation, it is not clear how to obtain objective sources of data about the system. Questionnaires may be unreliable, and interviews or think-aloud protocols can provide shallow sources of information [163]. As a result, it is suggested that data logs are triangulated with other data sources in order to obtain a clear portrait of the system's success. In this thesis we frequently compare the fNIRS signal with survey responses and behavioral data to construct a clear portrait of the user's interaction.

2.2 Introduction to Functional Near Infrared Spectroscopy

Functional near infrared spectroscopy (fNIRS) is an optical brain sensing technique developed in the 1990s that is portable, resistant to movement artifacts, and observes similar physiological parameters to functional magnetic resonance imaging

(fMRI) [26, 79, 152]. These characteristics have made it an attractive alternative for researchers seeking to observe the brain in natural working environments.

fNIRS uses near-infrared light to measure concentration and oxygenation of the blood in the tissue at depths of 1-3cm [167]. 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.

Figure 2.1: **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.

Biologically, when a region of the brain is active, there is an increase of blood flow to that region [44]. This increase of blood flow is typically coupled with decreased levels of deoxygenated hemoglobin and increased levels of oxygenated hemoglobin. Thus, fNIRS can be used to measure activity in localized areas of the brain.

To make this calculation, raw data can be transformed into deoxygenated hemoglobin concentrations using the modified Beer-Lambert Law:

$$\Delta A = \varepsilon \times \Delta c \times d \times B \tag{2.1}$$

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 infrared 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 [26].

For the sake of focus, we will primarily discuss fNIRS probes that are placed on the forehead, measuring brain activity in the anterior prefrontal cortex. This placement offers a significant advantage for researchers: fNIRS measurements can be made without the interference of hair follicles, which can absorb light and disrupt the signal. As a result, probes that monitor activity in the prefrontal cortex are often more unobtrusive, and therefore, more interesting for researchers searching for ecologically sound measurements.

2.2.1 Comparison with other brain sensing techniques

In the past, various brain sensing technologies have been proposed to observe a user's response to activities in a lab setting. Electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) are two of the most prevalent and both have been successful at measurement and classification of brain activities. fMRI requires a person to lie motionless inside a large, loud chamber in which small movements (those larger than 3mm) often result in discarded data (i.e. [20, 40]). fMRI scanners are also expensive to purchase and maintain, requiring technical staff and purpose-built rooms or buildings.

EEG has recently seen commercial success because it is portable, less invasive, and relatively inexpensive. While EEG has a high temporal resolution, it also has a low spatial resolution, which makes it difficult to pinpoint the origin of neural activity. Although EEG is easier to set up and use than fMRI, many configurations require applying gel into a person's hair to create a conductive contact with the skin. Finally, movement artifacts can be problematic with the use of EEG. With-

out proper filtering methods, minor movements, such as facial muscles can disrupt incoming data. Despite these limitations, EEG has gained popularity because of its quick temporal response (1 ms), the strong existing body of EEG research, and the availability of well-supported commercial setups.

2.2.2 fNIRS Advantages

While fNIRS preserves some of the core features that make EEG a popular brain sensing technology, most notably its ease of use and portability, fNIRS also has a few unique properties that are worth considering. For example, fNIRS has a short setup time and is generally resistant to movement artifacts [78]. Mouse-clicking, typing, eye movement, and blinking in normal computing environments are acceptable during the use of fNIRS [152], and minor head movement, respiration, and heartbeats can be filtered using known signal processing techniques. More major head or forehead movements (which could be induced by frowning) are disruptive to the signal

fNIRS also has a spatial resolution on the order of 2-3cm, and readings have been validated against fMRI [157]. Furthermore, fNIRS provides access to hemodynamic and metabolic parameters that are not accessible with EEG (which is sensitive to electrical signals and not to blood flow or tissue oxygenation) and fMRI (which is only sensitive to deoxygenated hemoglobin and not oxygenated hemoglobin).

2.2.3 fNIRS Considerations

It's important to note that there are both caveats and considerations to the use of fNIRS as well. Once a region of the brain becomes active, the biological response to support this increase in activity takes several seconds to reach the cortex [63]. As a result, changes in blood oxygenation that reflect user state cannot be detected immediately using fNIRS.

The impact of this limitation is two-fold. First, we are more likely to measure signal differences in short-term and long-term cognitive states rather than instantaneous one-time events. Second, because the slow biological response impacts how

quickly we can identify or classify user state, it also impacts the design of adaptation mechanisms in biocybernetic systems that respond to fNIRS user state. In a later section, we will discuss some design considerations that can help circumvent this issue.

Because light from fNIRS reaches depths of 1-3 cm, activity in deeper areas of the brain is not directly accessible. Additionally, hair can obstruct light, so sources and detectors must be maneuvered to maintain contact with the skin [36]. Although there are several variants of full-head fNIRS devices [51], most are noticeably less comfortable than probes designed exclusively for the forehead.

In an HCI setting, it is important to maximize comfort without sacrificing signal quality [67]. Recent research suggests the implementation of brush optrodes [173] to help maintain a more comfortable contact. In addition, proof-of-concept studies have begun to investigate non-contact fNIRS, using light sources that do not maintain direct contact with the skin [141]. While these projects suggest an evolved trajectory of fNIRS sensors into the future, many fNIRS researchers limit their investigation of neural activity to the prefrontal cortex (PFC), an area of the brain situated behind the forehead, to circumvent these issues. For the remainder of this thesis, the studies discussed will refer to the use of fNIRS on the PFC.

2.3 Analysis and Classification of fNIRS Data in HCI

Although the history of fNIRS analysis and classification is deeply embedded into neuroscience and biomedical science, we focus primarily on applications to Human-Computer Interaction. This is largely because the high-level objectives in HCI differ from those in other domains, which may seek to study the functionality of the brain. These studies place strict restrictions on participants in an effort to prevent muscle movement from contaminating the signal, and use extremely controlled experimental designs that are typically divorced from realistic use-cases. Finally, analysis is always performed offline, or after the experiment, as there is no incentive to obtain results during experimentation.

Conversely, a major objective for using fNIRS in HCI is obtaining ecologically valid evaluations of the user, which necessitate a less-controlled environment in which movement artifacts are more likely to be present. Investigating the inner-workings of the brain often is not the primary objective of physiological studies in human-computer interaction. Instead, researchers focus on trying to obtain reliable, generalizable correlations between physiological signals and user state. Finally, a necessary component of constructing adaptive systems is the classification of fNIRS data in real-time, further nudging analysis away from offline, heavily controlled environments.

In the sections below, we survey different analysis methods either directly from the HCI literature or from studies that analyze fNIRS data with similar goals in mind (for example, for use in brain-computer interfaces).

2.3.1 Statistical Analysis of Oxy-Hb and Deoxy-Hb

One method for using fNIRS to investigate changes in cognitive state is the statistical analysis of changes in oxy-Hb and deoxy-Hb. Recall that as brain activity increases, we generally observe increases in oxy-Hb and decreases in deoxy-Hb. By analyzing the changes in these parameters during a user's interaction with a complex task, we can hypothesize the level of activity in the user's prefrontal cortex. Here, we share two examples of studies that have performed offline analysis of changes in oxy-Hb or deoxy-Hb to investigate workload levels.

Ayaz et al. [7] used this approach to detect the level of workload for participants piloting unmanned air vehicles (UAVs). In this task, participants were asked to sit at workstations and direct simulated air traffic, trying to prevent accidents. The number of UAVs was varied (6, 12, 18) between trials and the mean change in oxy-Hb was calculated over the course of each trial.

As users were forced to keep track of more UAVs, fNIRS detected increased levels of oxy-Hb in the PFC. Ayaz found these changes to be comparable to those observed during interacton with the n-back task - a well-characterized psychological task for increasing working (or short-term) memory load. Increased levels of oxy-Hb

also correlated with self-reported NASA-TLX workload measures, further validating the detection of signals that point to workload.

In another translation of fNIRS measures to real world environments, activity was recorded as participants were engaged as part of a human-robot team [153]. In this task, participants engaged in a multi-tasking assignment that could not be accomplished by the human nor the robot alone. The study investigated three classifications of multi-tasking — delay, dual-task, and branching. While *branching* required participants to maintain the context of a primary task while exploring a secondary task, users did not have to maintain this context in the *dual-task* condition, and they completely ignored the secondary task in the *delay* condition. Analyzing the mean combined hemoglobin (deoxy-Hb + oxy-Hb) for each participant, the branching condition was found to have higher levels of combined hemoglobin than either the dual-task or the delay conditions. In a second experiment that compared changed in deoxy-Hb in random interruptions with interruptions that could be predicted by the user, Solovey found that random interruptions provoked sharper decreases in deoxy-Hb.

Although offline statistical analysis is limited in its direct application to brain-computer interfaces, this methodology serves as common first-step in establishing a foundation for using fNIRS to detect new user states or apply fNIRS to different domains. In chapter 3, we use a similar technique to validate the use of fNIRS for evaluation of information visualization interfaces, or analysis of *how* information is presented to users.

2.3.2 Automatic Detection of User State

If fNIRS is to become a viable tool for analyzing mental state during interaction with an interface, it would be ideal for the analysis of fNIRS signals to move from a manual to an automated process. In this section, we discuss work that has employed the use of predictive models to objectively (and automatically) classify user state.

2.3.3 Classifying Known User State

When the user's intended state is known, some researchers have used predictive models as a statistical method to show that multiple user states is separable [74, 60, 98, 106, 143, 142]. However, the true potential of these models is in the automation of classifying fNIRS signals. This allows evaluators to avoid non-automated analysis - a potentially time-consuming task if fNIRS is to be used as a tool in real user studies. In these circumstances, researchers check the accuracy of their model by using cross-validation techniques.

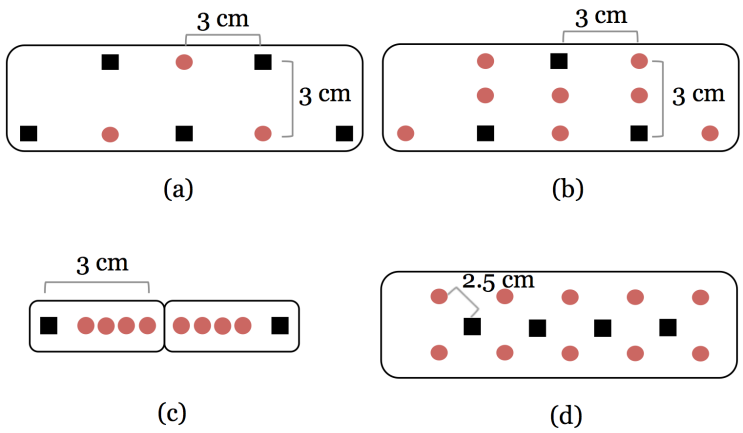


Figure 2.2: Probes with different numbers/locations of sources and detectors may impact classification accuracy of user state. We show four probe configurations used in studies described in section 5. Red circles are light sources and black squares are detectors. In each case, probes would be centered on the participant's forehead (a) Moghimi et al. [106] (b) Luu and Chau [98] (c) Girouard et al. [60], Hirshfield et al. [74], Solovey et al. [153, 150] (d) Ayaz et al. [7].

For example, Luu and Chau [98] used predictive models to distinguish between fNIRS signals during periods of low and high preference. In their experiment, participants viewed pictures of soft-drinks that they either highly-preferred or did not like at all. After identifying sensors that highly correlated with user preference, they were able to predict the preferred drink with over 80% accuracy.

Similarly, Moghimi et al. [106] used fNIRS to measure participants' emotional responses to music, attempting to capture both valence (positive or negative feelings) and arousal. They used linear discriminant analysis (LDA) to build a clas-

sifier and found that they could distinguish positive and negative valence with an average accuracy of 71.94%. They also found they could distinguish between high arousal music and brown noise (low arousal) with an average accuracy of 71.93%.

Finally, Girouard et al. [60] measured users as they played a game of Pacman, interacting with game modes that were both very easy and very difficult. They used a k-nearest neighbor (kNN) algorithm to classify game difficulty levels. While Girouard found that they could distinguish between periods of play and non-play with accuracy levels above 90%, distinguishing between easy and hard difficulty levels yielded classifications just over 60%.

As Figure 2.2 suggests, the discrepancies in classification accuracy in each study may stem from the various configurations of fNIRS probes as well as the differing analysis methods; the first two studies used configurations with significantly more source-detector pairs. This provides two distinct advantages. First, increasing the number of information channels decreases the potential for one noisy channel to adversely impact a model. Second, these configurations provide better coverage of the prefrontal cortex. For example, Moghimi et al. [106] showed that using information from a source-detector pair on the anatomical midline of the prefrontal cortex yielded the best overall accuracy in their model for capturing emotion. In the next section, we go into further detail about how fNIRS is used as input to these models.

In this thesis, I rely on the same set up used by Hirshfield [74], Solovey [153, 150], and Girouard et al. [60]. Although this may limit the signal coverage in the prefrontal cortex, it serves as a suitable baseline for fNIRS usage, and opens the possibility that use of a different system may increase classification accuracy.

2.3.4 Selecting fNIRS Features

When predictive models are created, we need to determine which features of the signal are fed into the models. Choosing too many features with too few training examples may result in the "curse of dimensionality" and low classification rates. Choosing too few features, or incorrect features of the signal, may lead to a set of

features that is not truly descriptive of the signal, also resulting in low classification accuracy. Currently, there is no standardized approach to feature extraction. We give four examples from current fNIRS literature:

- Solovey et al. [150] used the signal value from each time point and each channel over the entire trial as individual features to a support vector machine (SVM).
- Luu et al. [98] used the average signal value, estimated from a specific channel over a specific time interval within a trial (for example, 15-45 seconds).
- Hirshfield et al. [73] extracted the max signal value, min signal value, mean signal value, slope, time to peak, and full width at half maximum.
- Moghimi et al. [106] used the mean and slope of the signal during each trial. They also used the coefficient of variation, mean difference between signal and noise, and a handful of laterality features.

As these examples suggest, there is yet to be a prevailing consensus about which features of the fNIRS signal potentially result in the highest levels of accuracy. However, there are at least two dominant approaches to feature selection used in current fNIRS literature.

The first is to manually select a set of features based on an expert's knowledge or personal experience. For example, based on the changes of deoxy-Hb in response to load on working memory that we observed in Figure 3.7, the mean change in deoxy-Hb would appear to be a good indicator of low/high load on working memory. In a finger-tapping task, Cui et al. [37] show that including both oxy- and deoxy-Hb information to a predictive model improves accuracy. They also found that increasing the number of information channels improves accuracy. Broadly speaking, because the mean change in oxy-Hb is often used in the statistical comparison of fNIRS signals, it stands to reason that this feature is a good starting point for input to a model.

Using feature selection, context is important. Cui et al. [37] note that the features they chose were "necessarily dependent" on the classification technique (in

their case, support vector machines), and may be dependent on the task. One interesting distinction is that each of the previous examples is event-related — they observe how the fNIRS signals respond following a discrete moment in time. However, to serve as input to biocybernetic systems, it is often desirable for evaluators to view a moment-by-moment picture of workload, introducing new challenges. For example, task starting times, end times, and length may be undefined.

Given these challenges, an alternative approach is to select a large feature space. Then, using the participant's data, automatically determine which features yield the most information for each individual (for example, Luu et al. [98]). While this method often results in higher cross-validation classification rates, there is a danger that we may be building a model that succeeds on a particular dataset rather than one that represents a more general user state. We must take care not to overfit the model to the user's data, making it less flexible and robust for real world environments, not to mention less meaningful for HCI. In the next section, we describe how research is attempting to construct more generalized models of user state for real-time monitoring.

2.3.5 Classifying Periods of Unknown State

Unfortunately, in normal user evaluation, researchers often do not know in advance the true label (or user state) of a given period of fNIRS data, or even when that user state may begin. To help solve this problem, Hirshfield et al. [73] proposed a methodology for using machine-learning techniques to classify user state in these scenarios.

1. Choose cognitive benchmark tasks from the psychology literature that are known to induce specific user states. For example, if we are investigating the level of verbal working memory that a visual environment might induce, we might run a participant on a demanding 0-back and 3-back task, representing low and high levels of verbal working memory.

2. Next, we build a machine learning classifier to identify and store a cognitive

footprint of the fNIRS signal during each level of the benchmark task. We make the assumption that we have stored an accurate representation of verbal working memory for this particular user.

3. Finally, the participant performs a set of tasks in a more complex environment. We run the fNIRS data from those tasks on the classifier we built in the previous steps. The idea is that we are comparing the fNIRS data from this complex environment to the patterns we identified in the cognitive benchmark tasks. Our machine learning classifier returns whether the signal most closely matches low, medium, or high verbal working memory.

Hirshfield used this methodology to explore the working memory demand in a driving task in which the steering controls were reversed [73]. In comparison to more natural steering controls, the model classified incoming fNIRS data during the reversed control condition as requiring high working-memory. In the same way, Hirshfield used the Stroop test to detect response inhibition during interaction with an interfaces. We use this general structure as a foundation to move towards adaptive brain-computer interfaces.

2.4 Using fNIRS in Adaptive Interfaces

One of the primary reasons we focus on algorithms that automatically classify the fNIRS signal is that they enable fNIRS to be used as input to intelligent, adaptive systems. Because fNIRS is lightweight and does not place any unreasonable restrictions on the user, it can feasibly be used to augment many operator stations. In a real-time adaptive system, fNIRS data may be used as an additional implicit or explicit input, relaying information about the user to the computer without any further work for the user.

However, while fNIRS devices typically have a sharp temporal resolution, as we have discussed, the biological signal is sluggish. As a result, there are limitations to the systems that can be constructed. In this section, we discuss adaptive systems

that use fNIRS as input, the limitations of such systems, and also identify domains where these systems may thrive.

2.4.1 Calibration and Training

Similar to the training method previously described [74], in order to create a successful biocybernetic system, users must first perform a task with a known cognitive effect in order for the system to calibrate to the characteristics of brain patterns of each individual. However, the model must be very cautious since users are often in a different mental state during offline calibration and online feedback [166].

Ideally, we strive to minimize training time and maximize classification accuracy. Unfortunately, these two objectives typically compete with each other. Many machine learning algorithms are traditionally designed with the assumption that there are hundreds or thousands of training examples. But in an experimental setting, training the user for hours on end is unreasonable. As a result, researchers typically train users for as long as the ordinary time constraints of a user study allows. However, as more research is done in this field, we may find universal patterns that allow us to circumvent the training period. For example, using fNIRS, Herff et al. [71] found neural responses to speaking modes to be consistent enough to construct a general classifier of accuracy of 71%. We suspect that similar general models may be constructed for classification of other user states.

2.4.2 Brainput: A Real-Time fNIRS System

In this section, we give a concrete example of previous work that uses passive fNIRS input to an intelligent system. Solovey et al. [150] created a system, *Brainput*, which was able to adapt a scenario where an interactive human-robot system changed its state of autonomy based on whether it detected a particular state of multitasking. We pay special attention to this study because it serves as the primary example of using fNIRS as input to a passive interface.

To train the system to detect these states, Solovey used a well-validated multi-tasking exercise that had previously been explored using fMRI. Participants

were shown either lower-case or upper-case letters of the word 'tablet'. Depending on the case of the letter, participants were instructed to perform different actions, thereby resembling a multi-tasking environment. In a previous study (which we described earlier), Solovey et al. [153] showed that this task could be used to identify multi-tasking scenarios with fNIRS.

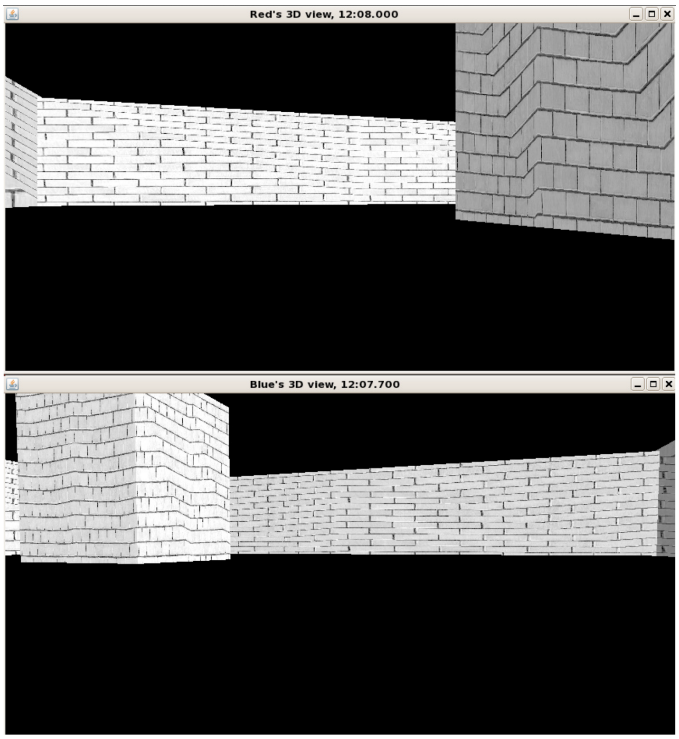


Figure 2.3: 3D View from the robots' perspective in Solovey's multi-tasking navigation environment. Automation was turned on or off based on the user's workload.

In the testing task, users were instructed to direct two robots through a virtual environment to search for areas with strong transmission strengths, and were told to not let the robots go idle or collide with walls or objects in the environment. fNIRS signals from the participants' prefrontal cortex were collected and classified as one of two user states that described the multi-tasking load associated with navigation. The second robot was autonomously controlled whenever the system detected a state of branching, where the user must hold in mind goals while exploring and processing secondary goals [150]. The changes occurred in real-time, allowing the system to dynamically respond to the user's individual, situational needs.

Solovey found that more participants completed the task with fewer collisions in this adaptive condition. However, to demonstrate that adaptation mechanism was indeed reacting to correctly classified fNIRS data, Solovey also introduced a maladaptive condition. This condition caused the system to intentionally perform the *opposite* response that should aid the user. In this maladaptive state, users did worse than in a nonadaptive condiation, and had a lower overall average transmission strength than either other condition. This experiment provides a successful example of using fNIRS as input to a real-time system that intelligently adapts to the user.

2.4.3 Applications and Opportunities

At least two other BCI implmentations exist that use fNIRS as input. Girouard et al. [59] constructed the first working passive BCI that relied on fNIRS as input, applying it to a simple game manipulation. Their work in building *OFAC* (online fNIRS and analysis and classification) set the groundwork for the subsequent Brainput system [150]. As a proof of concept to test OFAC, users were instructed to alternate between playing a game of Pacman and watching relaxing videos. Building on earlier work that differentiated the signal in these two tasks, a musical score was manipulated to increase it's pace during real-time detection of game-playing and classification of video-watching (or relaxing).

Afergan et al. [3] used fNIRS indices of workload to improve user performance in the navigation of unmanned aerial vehicles. This system used the *n-back task* for calibration, a psychology task that has been validated to induce varying loads of working memory load (we provide a more detailed description of the n-back task in Chapter 3, where it is used in the experimental methods). In the adaptive task, users performed path planning of multiple UAVs around numerous obstacles (unmanned aerial vehicles) while their workload levels were monitored using fNIRS. When the operator was detected to be in a high state of workload for an extended period fo time, the system decreased the user's workload, removing a UAV from the screen. When the operator was detected to be in a low state of workload for a long period of time, the system increased the workload by adding a UAV. In comparison with

a randomly adapting condition, Afergan found that the adaptive system resulted in fewer operator failures, fewer obstacles entered, and fewer neglected UAVs [3].

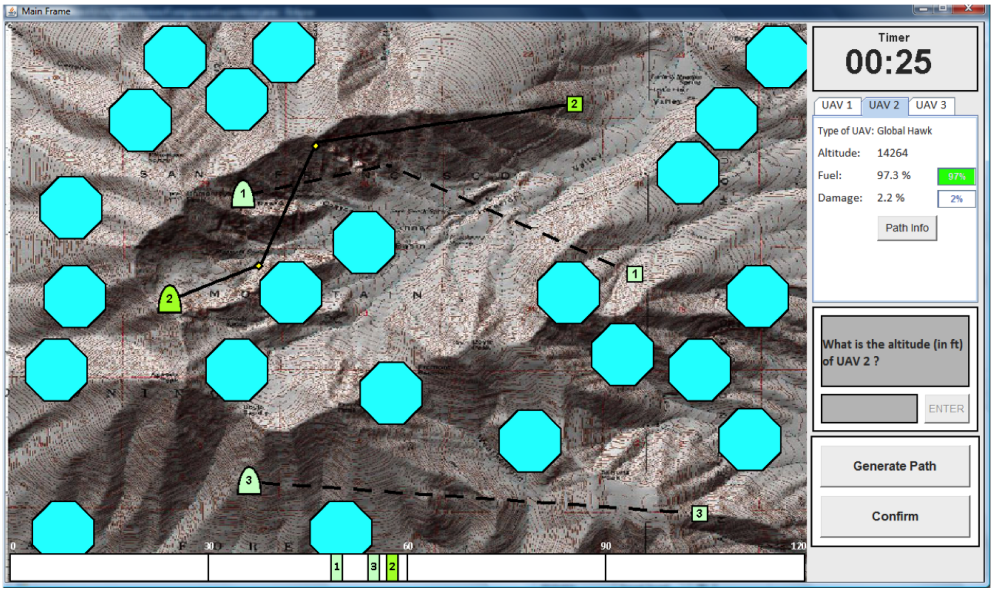


Figure 2.4: Afergan et al.'s [3] dynamic difficulty BCI. Participants navigated UAVs in a simulation to goals while avoiding obstacles. Based on fNIRS indices of workload, the system attempted to provide and optimal number of UAVs to manage.

Placing the previous studies in the context of this thesis, they use fNIRS to manipulate a temporal aspect of information delivery - *When* would a user benefit from the robot acting autonomously [150]? *When* should the user hear music during interaction [59]? *When* is the optimal time to give a user more or less work [3]? While they serve as strong examples of adaptive systems that can measurably improve users' interaction with their computer, this categorization also reveals open opportunities.

In this work, I attempt to expand the use of fNIRS beyond temporal manipulations of information delivery to those that also optimize the content or presentation of information. In addition, I discuss adaptive strategies that enable these systems to improve the user's experience despite relatively low classification rates from fNIRS.

Chapter 3

How: Using fNIRS to Evaluate Information Visualization

In order to build towards a system that can optimize its information delivery to the user, I constructed a conceptual framework of questions that need to be answered: *when* should the the computer deliver information, *which* information should the system deliver, and *how* should the system deliver information? In this first chapter, we begin by investigating what is perhaps the least explored of these three from a physiological perspective - *how*.

To explore this question as it related to fNIRS, and eventually to adaptive systems, we ground this work in the field of information visualization for two reasons: first, it is a discipline that is centered around the representation of information. Second, there have been numerous calls for stronger evaluation methods. That is, current techniques for optimizing the visual representation to a person, or even a group of people, remain unconvincing [23].

Finally, 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 [152].

In this chapter, I provide the first exploration of fNIRS in the context of information visualization, and motivate the potential of personalized systems [131, 125, 2, 116, 126].

3.1 Motivation

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 [34, 147, 154]), and psychology research suggests that evaluating performance without workload may lead to incorrect conclusions about the cognitive efficiency of an interface [16, 81, 119, 179]. Finally, cognitive state can change even as performance remains stable, meaning that performance metrics may not always accurately reflect cognitive processes [41, 170].

As a result, there has been a renewed interest in objective methods to evaluate cognitive processes during interaction with a visual interface [4, 136]. While previous fNIRS experiments in HC have studied cognitive state at various stages of interaction [7, 61, 73, 74, 150, 153], 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?

Finally, brain sensing opens the potential for systems that personalize information delivery to the user. Research increasingly suggests that users may differ in the way that they interact with information visualization [64, 184]. If a system is able to monitor user state during interaction, it may be possible to either provide the user with support on the fly or personalize future visual forms to better fit their cognitive profile.

However, there are concerns as to whether fNIRS may be capable of monitoring brain activity in visualization tasks since the physiological parameters which fNIRS monitors 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).

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 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.**

3.2 Background

3.2.1 Brain and Body Sensing in Visualization Evaluation

As Fairclough [47] 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 [136] 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 [25]. Few other significant interactions were observed. Work by Anderson et

al. is the most promising example of using physiological signals to evaluate visual interfaces [4]. 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 [4].

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 [134].

As a result, we explore the use of fNIRS as an alternative brain sensing technology. Recall from Chapter 2 that 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 [152, 157].

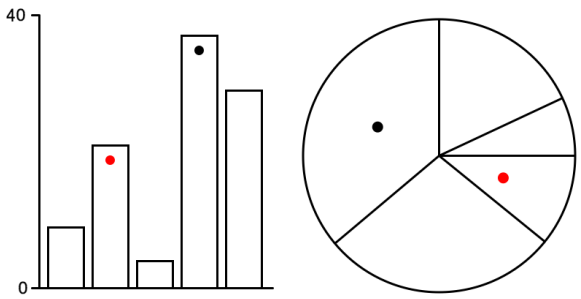


Figure 3.1: An example bar graph and pie charts from Cleveland and McGill's comparison task. Participants were asked to make a percentage estimation of the smaller section, marked by a red dot, with the larger section, indicated by a black dot.

3.2.2 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 pre-

sented with either a bar graph or pie chart (Figure 3.1) and asked to estimate the proportion percentage of a smaller value in the graph to a larger value [34]. 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 [147]. Spence and Lewandowsky demonstrated that pie charts perform reasonably well in a direct comparison with other basic visual forms [154]. In more complex tasks - when comparisons consist of combined proportions (A+B v. C+D) - pie charts can outperform bar graphs [156]. For a more extensive history of the pie chart, see Spence's article "No Humble Pie: The Origins and Usage of a Statistical Chart" [155].

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 [70]. Using pie charts and bar graphs, Hullman et al. showed that social factors can influence quantitative judgments [83]. 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 [171].

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.

3.3 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.

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.

3.4 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.

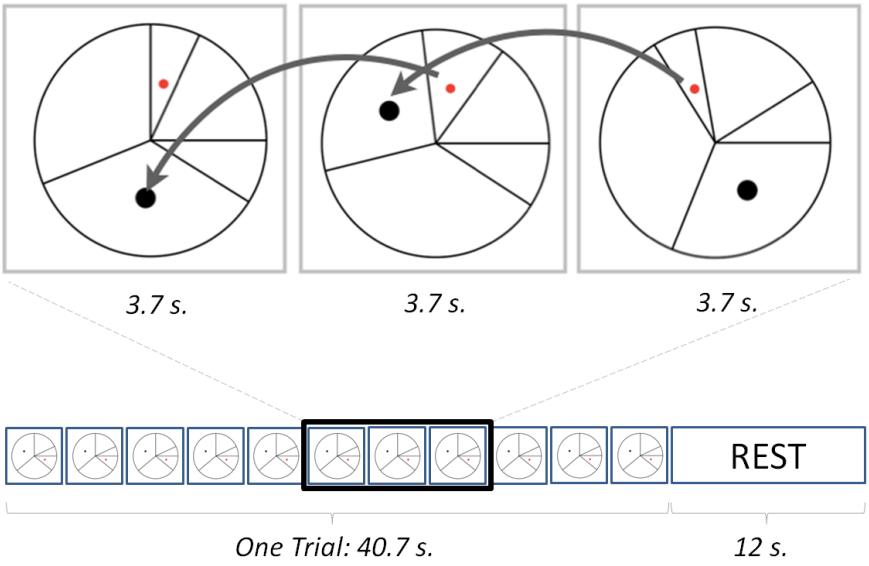


Figure 3.2: In our modified comparison task, participants compare a slice in the current pie chart to a slice from the previously seen pie chart.

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 3.2 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%.

3.4.1 Measures

3.4.1.1 Questionnaire: NASA TLX

We used an unweighted NASA-TLX questionnaire [109], a subjective rating that has been successfully used to capture workload since the 1980s [68]. 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.

3.4.1.2 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. 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 [26]:

$$\Delta A = \varepsilon \times \Delta c \times d \times B \tag{3.1}$$

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 infrared 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 [26]. 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 [153].

3.4.1.3 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 [34]:

$$\text{error} = \log_2(|\text{judged percent} - \text{true percent}| + .125) \tag{3.2}$$

3.4.2 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.

3.5 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 (3.3):

$$\overline{\Delta \mathrm{Hb}} = \frac{\sum_{\mathrm{t}=0}^{\mathrm{n}} (\mathrm{Hb_t} - \mathrm{Hb_0})}{\mathrm{n}} \tag{3.3}$$

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 [7]. 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.

3.5.1 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

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

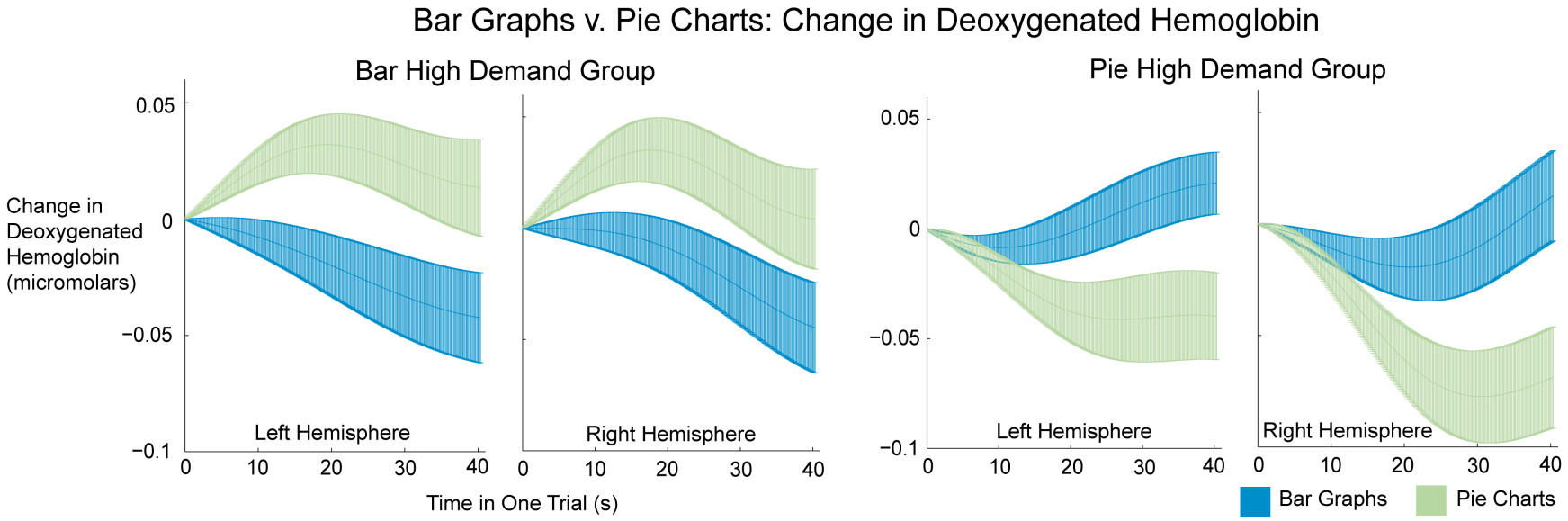


Figure 3.3: In a user study involving bar graphs (blue) and pie charts (green), we found that a group of participants that subjectively rated bar graphs as more mentally demanding than pie charts (left) exhibited reversed fNIRS signals from those who rated pie charts as more mentally demanding than bar 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.

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.

3.5.2 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 (the 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**.

3.5.3 fNIRS Signal: Bar High Demand v. Pie High Demand

We found that the levels of deoxygenated hemoglobin (at the 3cm source-detector distance) exhibited by participants who found bar graphs more mentally demanding were the *reverse* of those participants who found pie charts more mentally demanding.

Figure 3.3 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*.

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 signifi-

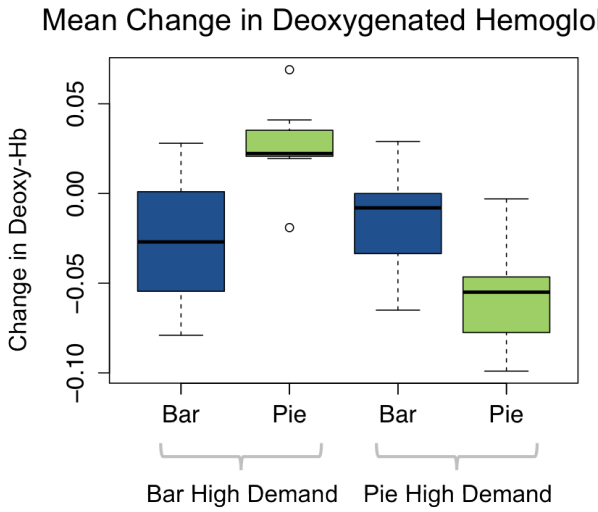


Figure 3.4: The mean change in deoxygenated hemoglobin for each graph shows that the visual design that partipants found to be more difficult resulted in larger decreases in deoxygened hemoglobin.

cant 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 3.4).

While the mean provides a suitable metric for analysis, it can miss some trends in time-series data. Specifically, Figure 3.4 suggests that both groups recorded similar changes in deoxygenated hemoglobin while interacting with bar graphs. However, Figure 3.3 shows that the fNIRS signal was trending in opposite directions.

Finally, it's worth noting that the decreases in deoxy-Hb we observed occurred at levels that are lower than those typically observed in fNIRS studies (less than 0.1). However, we observed them consistently across participants. Coupled with our experimental design, which used a random ordering of trials (8 bar graph vs. 8 pie chart) and a counterbalanced ordering of sessions (n-back vs. graph), we find these differences to be lend insight into participants' interaction with graphs. For a more detailed view, in Appendix A we show the fNIRS plots from all source-detector pairs.

3.5.4 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 (3.2) 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 3.5.

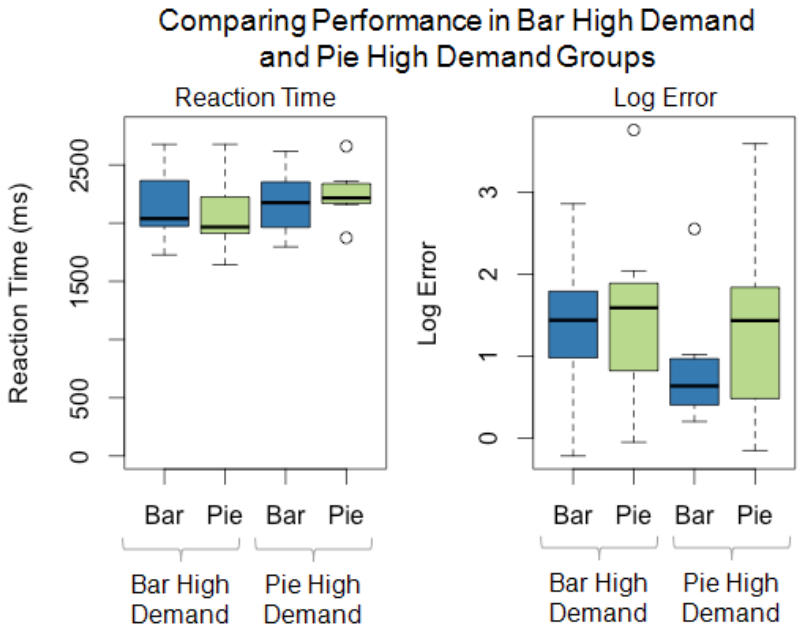


Figure 3.5: Despite a clear separation in brain activity between the bar high demand group and the pie high demand group, we observe very little difference in response time and error. The whiskers represent the max/min values, excluding outliers. Outliers are assigned by being more/less than 1.5 times the value of the upper/lower quartiles.

These results suggest that although there were significant differences in brain activity between bar graphs and pie charts, there were 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 [34], 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 [154, 156].

3.6 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.

3.6.1 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.

3.6.2 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.

3.6.3 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.

3.7 N-Back Task: Detecting Mental Workload

During the course of this chapter, we have been 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 pieces of evidence that support this claim:

1. fMRI studies have suggested that decreases in deoxygenated hemoglobin are indicative of increased brain activity [44]. 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.

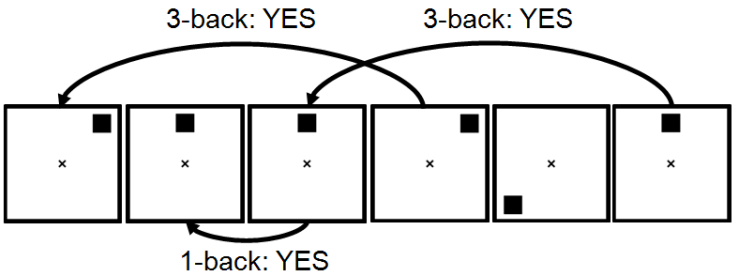


Figure 3.6: In the visuospatial n-back task, participants view a series of slides and respond whether the current pattern matches the pattern from n slides ago. We show positive answers for both the 1-back and 3-back conditions.

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.

3.7.1 Methods

In the visuospatial 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 3.6 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.

3.7.2 Results

Looking at the results, Figure 3.7 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 [117]. Additionally, the 3-back task induced lower levels of deoxygenated hemoglobin, agreeing with other observations of deoxygenated hemoglobin from the fMRI literature.

When placed side-by-side with the fNIRS readings from the 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 3.8). Similarly, the signal recorded from the less-demanding 1-back task

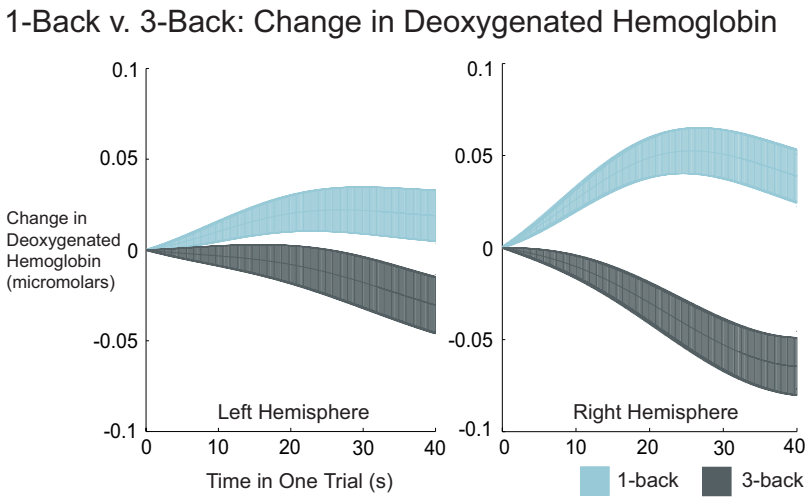


Figure 3.7: The mean fNIRS signal across all 16 participants in the Baseline Task. We see a clear separation between the 1-back and 3-back conditions participants. The more demanding 3-back condition mirrors signals from the graph design that participants believed was more mentally demanding.

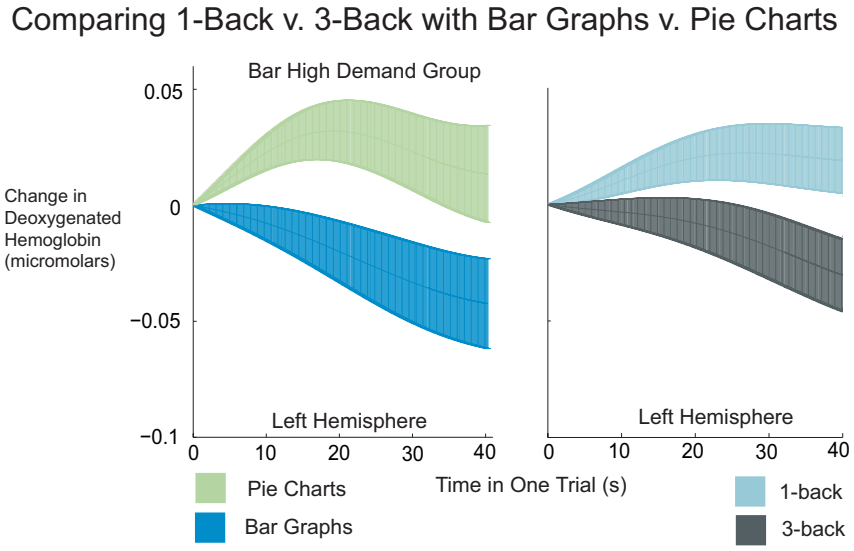


Figure 3.8: An example of comparing the n-back signal with those recorded in the bar graph v. pie chart experiment. The signal recorded during the more demanding 3-back resembles the signal recorded during bar graphs for the bar high demand group - participants who found bar graphs to be more mentally demanding than pie charts.

resembles those observed in the graph that participants identified as requiring less mental effort (Figure 3.8).

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.

3.8 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.

3.8.1 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 [42, 118]. 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 question-

naires, 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.

3.8.2 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" [135]. 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 [135, 139].
- 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 [150], increasing the difficulty of a video game [60], or reversing the steering mechanism in a driving task [73].

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.

3.8.3 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 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 3.9).

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 3.10, 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.

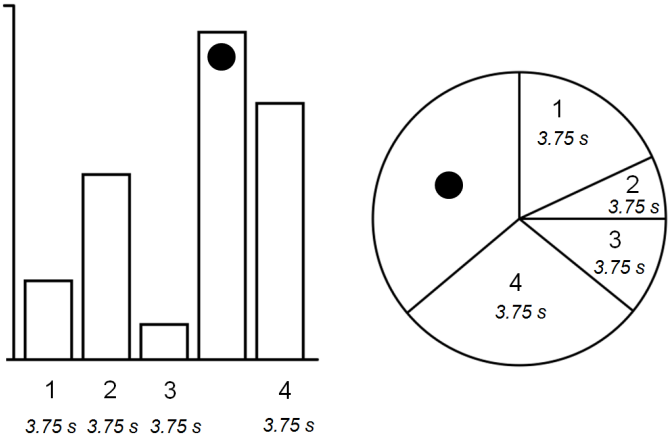


Figure 3.9: Participants sequentially compared elements of a graph to the largest element of the graph.

3.9 Findings

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

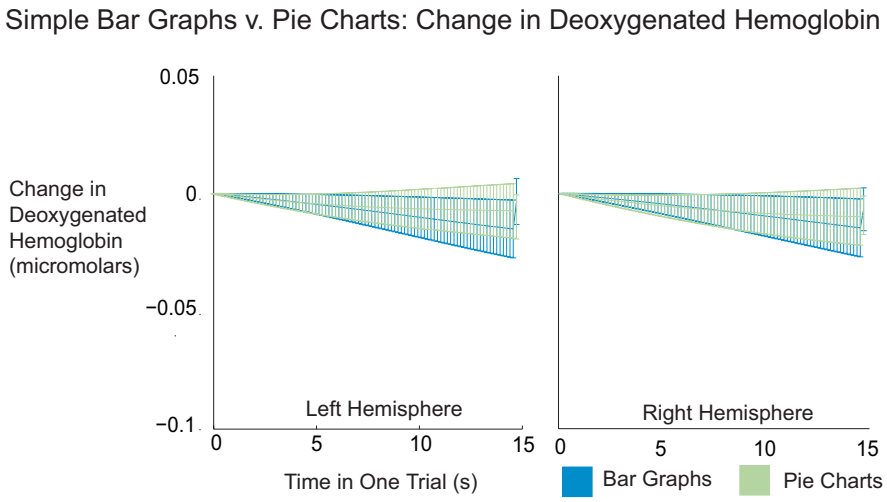


Figure 3.10: The mean fNIRS signal across all 8 participants in a simple bar graphs and pie charts task. The lack of activation shows that fNIRS may be less suited for simple, perceptual comparisons.

the individual. Recently, Solovey et. al [150] 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.

3.10 Towards Understanding Individual Cognitive and Mixed Initiative Systems

In this chapter, we discovered that the neural 'footprint' of interaction between bar graphs and pie charts differed between people. This result suggests that the best visual representation for one person may be different than that for another person. Recent work in visualization has made similar suggestions [90, 162, 184]. However, to move towards a scenario where the computer can maximize engagement (or understanding) by providing a personalized visualization, we need to obtain a better of understanding of the cognitive factors that result in performance differences. In this section, I use previous work in visualization to motivate a framework for considering and investigating the impact of individual cognitive differences.

3.10.1 States, Traits, And Experience/Bias

In recent years, strides have been made toward understanding the impact of individual differences on performance when interacting with visual analytic systems. Research has shown that factors such as personality [64, 184], spatial ability [27], biases [105, 185, 186] and emotional state [6, 52, 91, 132, 145, 138] impact a user's performance. Though progress is undeniable, a common limitation is that every cognitive factor that affects visualization performance is not considered or properly controlled. For instance, studies that focus on personality factors alone do not consider how differences in working memory, perceptual ability, and previous experience can also affect visualization performance. These as well as numerous situational differences make it difficult to not only design systems, but performing evaluations that are generalizable and replicable.

As stated by Yi in his position statement in 2010, the visualization community has yet to employ a comprehensive and standardized model for measuring individual differences such that researchers can better understand how factors in individual differences interact with each other and with existing evaluation techniques [180]. While one conceptual framework cannot solve all the problems de-

scribed above, we believe that individual differences can be categorized into three major dimensions: cognitive traits, cognitive states, and experience/bias.

Cognitive traits are user characteristics that remain constant during interaction with a visual analytic system. Factors such as personality, spatial visualization ability, and perceptual speed are all examples of cognitive traits. These have been shown to correlate with a user's ability to interact with a visualization [28, 35, 64, 164, 184] and can be generalized to predict the behavioral patterns of users with different cognitive profiles.

Cognitive states, on the other hand, are the aspects of the user that may change during interaction and include situational and emotional states, among others. Research has shown that a user's performance can be significantly altered by changes in their emotional state [6, 52, 91, 132, 138, 145], and the importance of combining workload with performance metrics has been noted for decades [81, 119, 179]. Although cognitive states are difficult to measure because of their volatility, they provide important contextual information about the factors affecting user performance that can not be described through cognitive traits alone.

Cognitive states and traits can describe a significant portion of a user's cognitive process but they are not comprehensive; *experience and biases* can also affect cognition. Intuitively, we think of experience and bias separately, but they both describe learned experiences that can affect behaviour when familiar problems arise, and are therefore not orthogonal. Although there has been little work about the impact of experience/bias on interaction with visual analytics systems [5, 32, 45, 186], previous studies have shown that learned behavior such as confirmation bias can significantly affect performance and decision-making [72].

3.10.2 Towards Adaptive Visualization Systems

One important advantage of understanding individual users' cognitive states, traits, and biases as a cohesive structure is that this opens up the possibility of developing adaptive, mixed-initiative visualization systems [161]. As noted by Thomas and Cook in *Illuminating the Path* [161], an important direction in advancing visual

analytics research is the development of an automated, computational system that can assist a user in performing analytical tasks. However, with few exceptions, most visualization systems today are designed in a "one-size-fits-all" fashion without the ability to adapt to different users' analytical needs into the design.

There is mounting evidence that successful adaptive systems can significantly improve a user's ability in performing complex tasks. Ziemkiewicz et al. [184] demonstrate that the impact of locus of control (LOC) on visualization can be significant. When the user is given a hierarchical visualization that correlates with the user's LOC, a user's performance can be improved by up to 52% in task completion time, and 28% in accuracy. Returning to our previous experiment, we used fNIRS to measure the cognitive state of users as they interacted with an information visualization. However, during this process we identified differences in the fNIRS signal between users. While the reason behind these differences is unknown, it is possible that we could use this information to present an optimal chart and reduce the cognitive load.

It is clear that adaptive systems offer new possibilities for visualization research and development, but more work is necessary to model *how* and *when* a system should adapt to a user's needs. As noted earlier, only emphasizing one or two of the three proposed dimensions can lead to a system incorrectly assessing the user's analysis process and provide the wrong adaption. By examining all three dimensions in a cohesive fashion, it becomes possible for a system to predict a user's performance and realize the potentials of an adaptive, mixed-initiative system as proposed by Thomas and Cook. In the the next chapter, we move away from visualization to give an example of how fNIRS can be used in an adaptive system to help optimize *which* information is presented to the user.

Chapter 4

Which: Investigation of fNIRS Brain Sensing as Input to Information Filtering Systems

In the previous chapter, we used fNIRS to detect signals that were dependent on the visual display of information. While this demonstrated the capability of fNIRS to be sensitive to *how* information is delivered to the user, we did not use the signal as input to an adaptive system. In this chapter, I complete the biocybernetic loop and show that fNIRS can be used as input to systems that modify *which* information should be filtered or prioritized to the user [126]. Since this is one of the first fNIRS systems that utilized a biocybernetic loop, it also serves as an initial exploration (after [150]) of the use of fNIRS in a passive BCI.

4.1 Motivation

User attention is a scarce resource in modern computing. Mental resources are often divided among disparate but concurrent streams of information. Twitter updates, text messages, and emails, for example, draw the attention of a user and pull cognition away from primary working tasks. In the wake of such pervasive distractions, research has shown that focusing on the wrong information or consuming

information at the wrong moment can not only lead to a decrease in performance during work, but negatively impact work satisfaction, and increase stress and anxiety [10, 62, 100].

To address some of these problems, researchers have suggested that physiological measures of workload or attention should be used to deliver information at an opportune time. For example, Bailey et al. used pupil dilation as a measure of workload for interruption [11], and Solovey et al. built an interactive system that adapts robot automation to a human operator's working memory load [150]. However, while physiological computing has been used to manipulate *when* information is delivered to the user, very little work has focused on *which* information should be delivered.

In this paper, **we explore the use of fNIRS to classify preference judgments and drive information filtering systems.** Given recent neuroscience literature, these parameters may allow the detection of preference judgments that extend beyond emotional response by incorporating the reasoning processes of the brain [14, 98]. Thus, if there is any correlation between fNIRS signals and preference judgments, fNIRS could potentially augment current practices by being used as an additional source of passive information to filtering systems.

However, there are significant challenges to the use of fNIRS in information filtering systems. Previous fNIRS work has analyzed preference judgments exclusively in offline environments [98]. Additionally, signals that correlate with preference are often subtle and may not translate to real world use cases. For these reasons, a primary goal of this paper is to explore whether fNIRS preference measures can be used in a real-time environment.

To investigate the use of fNIRS in information filtering systems, **we present an automated recommendation system that suggests new movies based on fNIRS measures alone.** Using fNIRS to monitor the prefrontal cortex, our system classifies brain data in real-time and iteratively updates a model of user preference to recommend movies that are personalized to the individual user. To evaluate our system, we ran a user study and found that fNIRS can contribute

information to a recommendation environment by outperforming a no-input control condition. In addition, we observed that the system's model of user preference improved the longer the user interacted with the device. Finally, we found that recommendations were uniquely catered to the individual — 45% of the movies each participant viewed were not recommended to any other participant — showing that we were responding to individual preference and not overall popularity.

We suggest that this brain recommendation system acts as a proof-of-concept for the use of fNIRS as input to information filtering systems. We argue that eventually by integrating *when* information should be delivered along with *which* information should be emphasized, brain-computer interfaces have the potential to automatically provide users with the right information at the right moment.

4.1.1 Contributions

We make the following contributions:

- **We show that fNIRS brain sensing can be used as input to information filtering systems.** We construct and evaluate a real-time movie recommendation system that is driven by brain signals that correlate with preference. We find that our system recommends higher-rated movies with fNIRS input than without it, and that the underlying model of user preference improves over time.
- **We discuss the implications of using fNIRS measures in information filtering systems.** We suggest that brain sensing can someday augment current recommendation systems, support the creation of recommendation systems in new domains, and unify recommendations across disparate information sources.

4.2 Background and Related Work

In this work, we focus primarily on *preference judgments* as a key input to information filtering. When integrated with *information filtering algorithms*, preference

enables users to allocate attention through recommendations of personalized information or products [148]. Thus, preference helps determine which information should be presented to the user. In addition, preference has been integrated into applications such as personal search [160], prioritizing incoming text and voice messages [101], optimizing user interfaces [54], calculating interruption costs [75], and guiding conceptual design [12], among others.

4.2.1 Measuring Preference: Explicit v. Implicit

In current information filtering systems, eliciting preference involves a tradeoff between accuracy and obtrusiveness. *Explicit measures* require users to record their own preferences through a rating scale. While explicit ratings are generally accurate representations of what the user prefers, they can act as attention-sinks by disturbing normal behavior with an interface and introducing an introspective cognitive step [114]. Additionally, responses depend on the assumption that users can sense and externalize their subjective emotions, which may not be true [82].

Implicit measures predict user preference by observing the user during natural interactions with a system, and are based on viewing history, purchase history, view time, or other behavioral measures [33, 53, 108, 122]. These ratings are essentially elicited for free as they require no additional effort on the part of the user. However, implicit ratings are widely considered to be less accurate than explicit ratings because they are based on prediction models that might not reflect the user's preference and can be affected by a number of other variables [114]. For that reason, new methods are often proposed to increase the accuracy of implicit ratings.

4.2.2 Physiological Measures of Preference

One approach to increase the effectiveness of implicit measures is to incorporate physiological sensors into preference prediction models. For example, combinations of galvanic skin response (GSR), electromyograms (EMG), blood pressure, respiration pattern, and electroencephalography (EEG) have been used to capture emotional responses to videos [21, 92]. Following this work, there have been several

attempts to use affective signals as input to recommendation engines. For example, Healey et al. [69] constructed an "Affective DJ" that dynamically constructed "energizing" or "relaxing" musical play lists. Similarly, Wu et al. [178] built a system that recommends multimedia with similar emotional content. In each of these cases, recommendations were grounded in emotional responses to content, and analysis of each system was largely preliminary.

4.2.3 Preference Judgments in the Brain

While preference judgments and emotional reactions are often linked, previous work indicates that there are two separate processing chains that combine to influence preference judgments: emotion and reason [14]. This is because preference (and specifically, economic decisions) may be based on various competing factors, such as price, usefulness, branding, and availability. For example, viewing a high-end sports car may elicit positive emotions, but a small, fuel-efficient hybrid car may elicit higher preference values.

The prefrontal cortex offers information about preference judgments that physiological sensors focusing on affective state may not detect. We ground our measures of preference in several studies that investigate the neural correlates of preference using functional magnetic resonance imaging (fMRI) and positron emission tomography (PET). In a study by Deppe et al., fMRI showed increased activation in the prefrontal cortex during *economic decisions* involving a preferred brand name [43]. Paulus et al. recorded similar results in *preference judgments* of drink categories [123] and McClure et al. compared activation in a blind Coke v. Pepsi test, finding that neural responses in the PFC were consistent with behavior [102]. Finally, in an example of emotional processing, Blood et al. found that intensely *pleasurable music experiences* resulted in blood flow changes in the PFC in comparison to a control music condition [17].

In this paper, we apply fNIRS to the detection of preference in the PFC, observing similar physiological parameters to the fMRI studies noted above. Additionally, we base our research on a recent study by Luu and Chau (extending

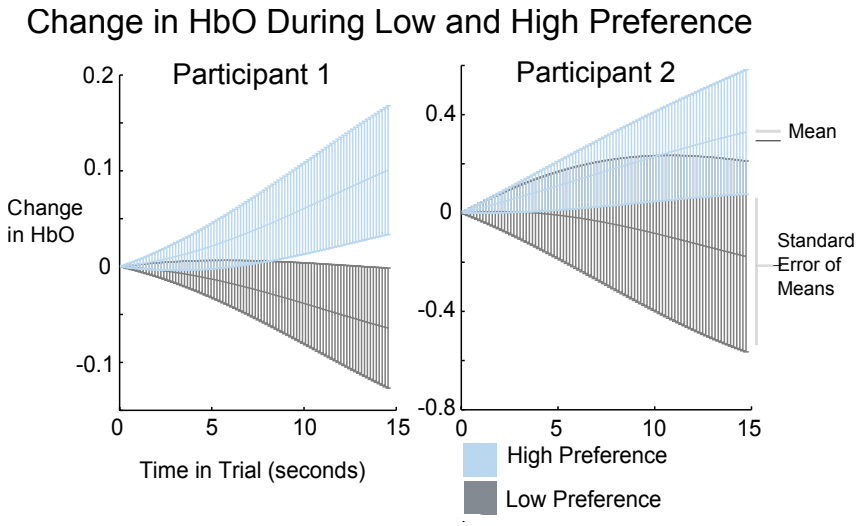


Figure 4.1: From a pilot study, an example of the fNIRS signal from the right hemisphere during periods of high preference and low preference. The plots show the mean change in oxygenated hemoglobin during 8 trials of high preference and 8 trials of low preference.

Paulus et al. [123]), where fNIRS measures of the PFC were used to classify extreme preferences for various beverage categories (e.g. milk, soda, water) [98]. In a pilot study of our own, we confirmed the results of Luu and Chau by observing different activation patterns in participants as they viewed pictures of digital music devices that they liked and disliked (Figure 4.1).

4.2.4 Physiological Input to Adaptive Systems

A key feature of our brain-recommender system is that it monitors preference without any specific effort from the user. The system reads passive information about users during natural interaction, and then adapts to their current state. While most work in BCI has focused on active brain-computer communication, more recent research has suggested the use of implicit neural parameters as input to adaptive systems [39, 47, 61, 183]. For example, Kohlmorgen et al. used EEG to measure mental workload in a multi-tasking driving scenario, where the secondary task would be removed during periods of high-workload [93].

George and Lecuyer survey current passive BCI literature and categorize them into four application areas: 1) adapting the level of automation, 2) implicit

multimedia content tagging, 3) video games, and 4) error correction and detection [57]. Although our work is most closely related to implicit multimedia content tagging, it extends that research by adding an adaptive element (new recommendations).

4.3 The Brain Recommender

In order to explore whether fNIRS can provide useful input to information filtering systems, we constructed a movie recommender that is driven exclusively by fNIRS signals and compared it to a system that does not include passive user input (a no-recommendation environment). By controlling for behavioral indicators of preference, we test the ability of fNIRS to add information to recommendation systems beyond traditional behavioral metrics of viewing time and history. Thus, if a brain-driven recommender provides intelligent recommendations, we believe that implicit fNIRS measurements can be used to augment current techniques.

- **Hypothesis:** Observing the brain with fNIRS will allow the brain recommender system to construct a preference model of the user, suggesting movies that cater to each user's interests.

In the following sections, we discuss the technical details of our system, report our experimental methods, and analyze the data from our experiment. Finally, we discuss the implications of our results, outlining a vision for a information filtering systems that are driven by the brain.

4.4 System Details

Constructing a fully-functional recommendation system based on brain input requires the coordination of a number of technological pieces. To provide a technical overview, we refer to Figure 4.2 and briefly discuss the flow of information in our system.

First, light sensing data is sent from our fNIRS data acquisition software to an analysis program built in our lab, where the signal is filtered to remove noise and movement artifacts. There, we partition the fNIRS data into segments of identical length to training examples we provided during an earlier training period. These segments are sent to Weka, an open-source machine learning library, where we classify the fNIRS signal based on previous examples [66]. This classification is sent to our Java application that holds the movie and rating database and serves as the backbone of our recommendation model. The application updates the database and recommendation model with new user information, and searches for the top recommended movie given all other previous data about the user. Finally, the selected recommendation is sent to a browser that navigates to the movie's corresponding IMDB page.

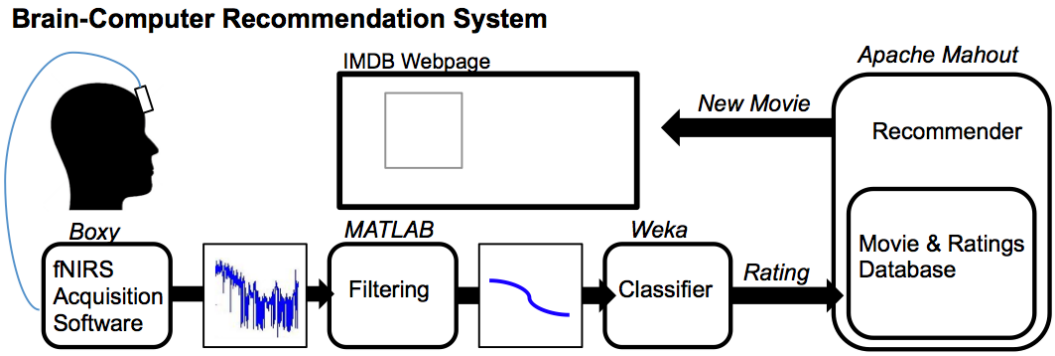


Figure 4.2: Basic architecture of the real-time classification system: Raw fNIRS data is filtered before being classified by a support vector machine (SVM). This preference classification updates our movie database about the user, and after refreshing our recommendation model, we show the top recommended movie.

4.4.1 Measuring and Filtering the fNIRS Signal

We used a multichannel frequency domain OxiplexTX from ISS Inc. (Champaign, IL) for data acquisition. Similar to the previous chapter, two fNIRS probes were placed on the forehead in order to measure the two hemispheres of the anterior prefrontal cortex. The source-detector distances were 1.5, 2, 2.5, and 3 cm. 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 6.25 Hz.

4.4.1.1 Filtering

In order to remove noise that might be the result of user movement, respiration, or heart beats, we apply filtering techniques described by Solovey et al. in their adaptive system that also used fNIRS input [150]. First, we used an elliptic low pass filter with a cutoff frequency of 0.025 Hz, stoppage frequency of 0.03 Hz, max ripple of 3 dB and a stop band attenuation of 50 dB. Next, we used a z-score to normalize the data in each information channel. Finally, for each training example, we calculated the signal change of each time point from the first time point in the example.

4.4.2 Building the fNIRS Classifier

Once each of the filtering steps was completed, we built a new preference model for each participant based on the training protocol we describe in the experiment section of this paper. We constructed a classifier that differentiated between low and high preference for each of the 16 information channels on our fNIRS device (2 probes x 4 distances x 2 wavelengths), using the filtered light readings at each time point of a trial as individual features to the classifer. Since we sampled data at 6.25 Hz, a 25 second trial would consist of approximately 156 features. Finally, we used a built-in support vector machine (SVM) algorithm from Weka's sequential minimal optimization (SMO) package.

4.4.3 Mapping Preference to 5-Point Rating Scale

While previous fNIRS work [98] suggested that we could discriminate between periods of low and high preference, the movie dataset we used to ground our recommendations was based on a 1 to 5 star rating. This left us with a mapping problem. Recall that we built a separate classifier for each information channel of the fNIRS device. To map classifications of low v. high preference to a 5-point rating scale,

we took a percentage vote from the classifiers. For example, if 80-100% of our information channels classified the incoming data as a period of high preference, we mapped this classification to a 5 star rating. If 60-80% of our information channels classified the data as a period of high preference, we mapped this classification to a 4 star rating.

This mapping is not ideal in a real-world scenario, as classification uncertainty is not equivalent to preference intensity. However, we use this approach to accommodate for the necessary time constraints of a normal experimental session. We suggest a more robust approach in the discussion section.

4.4.4 Dataset and Recommendation Engine

To build the movie recommendation engine, we used the HetRec 2011 MovieLens Data Set, an extension of MovieLens10M dataset [22]. This dataset contains 2113 users, 10197 movies, and 855598 ratings from Rotten Tomatoes and The Internet Movie Database (IMDB), two major movie rating websites [84, 137]. Thus, there is an average of 404.921 ratings per user and 84.637 ratings per movie. Our movie recommendation engine was constructed using Apache Mahout, an open source machine learning library for Java that includes built-in collaborative filtering algorithms. Our user recommendation system was based on a nearest neighbor algorithm using euclidean distance as a similarity metric.

4.5 Experiment

In order to evaluate the brain-computer recommender, we describe the experiment protocol in two sections: *training* and *testing*. For each participant, the *training* section consists of sending fNIRS examples to a machine learning model on known values, or in this case, movies that we already know the participant likes or dislikes. Instead of using preference ratings entered by people, the *testing* section uses machine learning classifiers to predict preference in real-time, which is used to provide updated movie recommendations to the user.

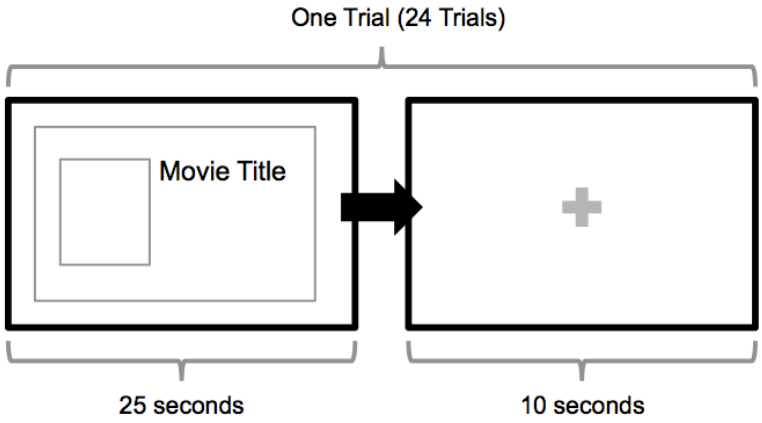


Figure 4.3: During training, participants viewed screenshots of their most favorite and least favorite movies for 25 seconds, followed by a 10 second rest period.

4.5.1 Training

At the start of the experiment, we provided participants with a list of movies picked from IMDB's list of 250 best movies and 100 worst movies and asked them to select their top three and bottom three. Participants viewed a timed slide show of selected movie webpages during which we recorded their brain activity with fNIRS. We showed each movie webpage for 25 seconds, followed by a rest period of 10 seconds (Figure 4.3). Participants viewed 12 slides of their top 3 movie titles and 12 slides of their bottom 3 movie titles. The brain activity recorded during these slides were used as training examples to our preference model. At the completion of training, the model was not altered for the remainder of the experiment.

4.5.2 Testing

In the testing section, participants viewed two trials, each of which consisted of a string of twenty movie websites, viewed sequentially. For each movie, participants viewed an IMDB page for 25 seconds, followed by an 8 second explicit rating period, and an 18 second rest period that enabled us to refresh the recommendation model (Figure 4.4).

4.5.2.1 Brain Recommender v. Control Condition

Our motivation for exploring fNIRS is that it provides an implicit, unique signal from the user that is not accessible by other physiological sensors. Because of this, we believe that the first step in assessing its value is to compare it against a no-input environment. Thus, users interacted with two trials — one involving the brain recommendation system and one of a no input control condition. If our system can deliver suitable recommendations running exclusively on brain measures, then fNIRS can be used as an augmentative implicit input to information filtering systems that is complementary to other physiological sensors.

We tested the following conditions:

- **Control Condition:** a series of pre-defined movies with average ratings are used for all participants. This serves as a baseline for a no-input recommender.
- **Brain Recommender Condition:** implicit preference ratings, as predicted by our fNIRS data classifier, are fed into a movie recommendation engine. We show the same start movie as the control condition, but new movies are selected based on previous preference values. For example, the 3rd movie is based on recorded preferences for the 1st and 2nd movies.

fNIRS sensors remained attached to the participants during the course of the entire experiment, giving no indication of the condition. Following each movie in both conditions, participants were asked to provide an explicit preference rating of the movie (1-5 stars). We used this rating to evaluate the performance of our system. Unlike the implicit fNIRS readings, the ratings did not influence future recommendations in any way.

4.6 Results

We ran this study with 6 male and 8 female volunteers *(N=14)*, aged 19-28 with a mean age of 22. The order of conditions was counterbalanced across all participants.

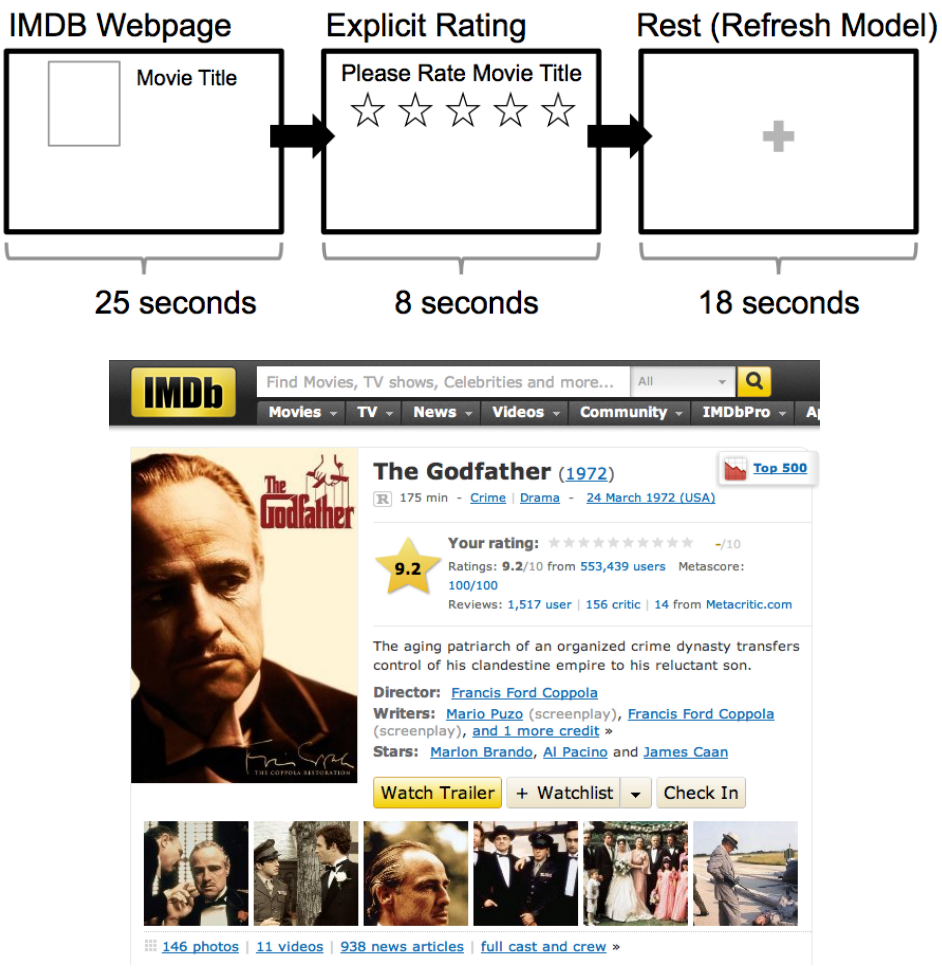


Figure 4.4: (Top) For each condition, participants viewed 20 movies. Each movie consisted of 25 seconds of viewing an IMDB page, an 8 second rating period, and an 18 second rest period. (Bottom) An example of an IMDB webpage participants viewed

Given our hypothesis, we identify three measures to explore the efficacy of fNIRS in driving our recommendation system.

1. *Recommendation ratings by condition*: How did participants rate movies in the brain recommender condition in comparison to the control condition?

2. *Recommendations over time*: A good recommender should improve over time as it constructs a more accurate picture of the user's likes and dislikes. Does the brain-driven recommender give better recommendations over time?

3. *Classification accuracy*: How well did our system guess the user's preference

for a given movie?

4.6.1 Recommendation Ratings by Condition

The key finding is that the brain recommender provided higher-rated movies than the control condition as the experiment session progressed. This difference becomes statistically significant after the 13th movie in each session. We display the distribution of all ratings in each condition in Figure 4.5.

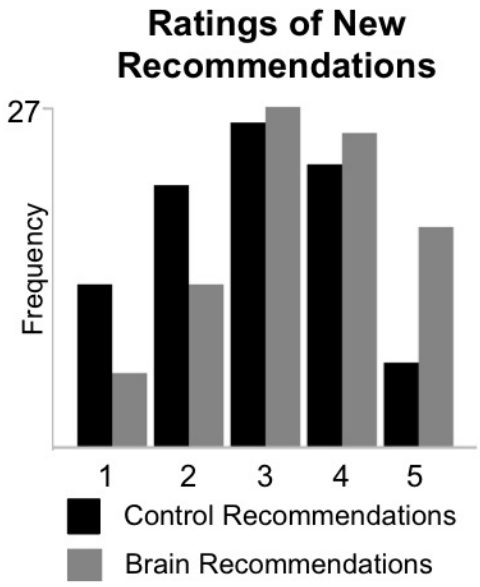


Figure 4.5: Histogram of ratings in the two conditions. We see that brain recommendations tend to be rated higher — mostly 3s, 4s, and 5s.

We would not expect a recommendation system to perform well until it had seen enough examples to provide suitable recommendations. In our system, we saw this switch occur typically after the 13th movie. We therefore analyzed the median rating in movies 14-20 for each participant across both conditions (see Table 1 for a summary). Running Mann-Whitney's U test on movies 14-20 revealed a significant effect of condition (the mean ranks of the control condition and brain recommendation condition were 10.46 and 18.54, respectively; $U = 41.5, Z = 2.88, p < 0.01, r = 0.54$).

We also analyzed the median rating for the entire 20 movie session. As

Condition	Median	Mean	Std. dev.
Control	3	2.9	1.17
Brain	4	3.6	1.15

Table 4.1: Ratings across movies 14 to 20

expected, we did not find a significant effect in condition (the mean ranks of the control condition and the brain condition were 11.75 and 17.25, respectively; $U = 59.5, Z = 1.86, p = .07, r = 0.29$).

Condition	Median	Mean	Std. dev.
Control	3	2.9	1.21
Brain	3	3.3	1.17

Table 4.2: Ratings across all movies

4.6.1.1 Unique Recommendations

To ensure the validity of these observations, we investigated whether our brain-computer recommender was aligning itself to individual preferences or simply gravitating to a small set of generally highly-rated movies. We found that 125 out of 280 (45%) movie recommendations in the brain condition were unique selections, meaning that each participant saw an average of 9 movies no other participant viewed. These results support our primary hypothesis that **the brain-driven recommendation system recommended movies that catered to the participant's individual preferences**.

4.6.2 Recommendations Over Time

Independent of the control condition, we find that **recommendations from our system improved over time, suggesting that the preference model was gradually learning about the user.** Across all participants, we analyzed the median rating given to movies at each time point (1-20) for each condition. For the brain recommender, we ran a linear regression and found that the total number of movies seen was a predictor of rating ($b = 0.046, t(20) = 2.541, p = 0.021$). This

means that over the course of 20 movies, the median recommendation improved by roughly one rating point (from 3 to 4 out of 5). The overall model fit was $R^2 = 0.223$. By comparison, applying a regression to the control condition determined that the number of movies seen did not predict movie rating ($b = 0.004, t(20) = 0.154, p = 0.898$).

In fig. 4.6, we show this trend by plotting the mean rating at each time point in the control condition with the brain recommender condition, and apply a linear fit line to the data.

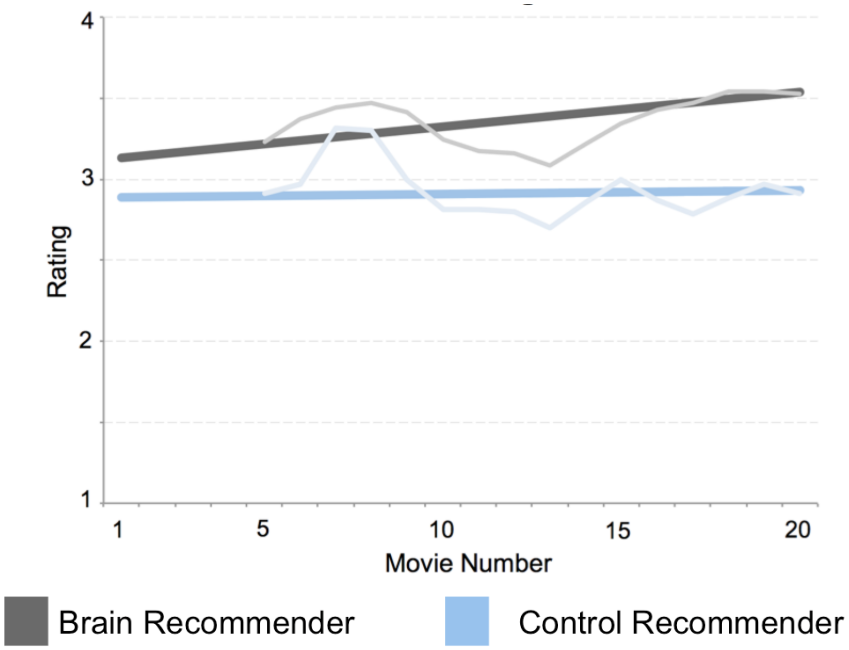


Figure 4.6: As participants viewed more movies, the system improved its recommendations. We show a moving average of movie rating (window size = 5) during the course of the 20 movie trial as well as a linear fit line. While the signal is noisy, we found a statistically significant improvement in the brain recommender's ratings over time.

4.6.3 Classification Accuracy

Recall that we used a percentage vote from our classifiers to translate classifications of high and low preference into a 5-point rating scale. To describe the accuracy of our system, we will use *low preference* to refer to ratings of 1 or 2 (out of 5) and

high preference to refer to ratings of 4 or 5.

In general, we found that our model skewed towards classifying movies as low preference (141 out of 280), while users tended to gravitate towards higher ratings. Figure 4.7 shows that when our model classified a movie as low preference, users were just as likely to have highly preferred the movie as they were to dislike it. However, when the model classified a movie as high preference (57 out of 280), users were five times more likely to give the movie a rating of 4 or 5 (out of 5) than 1 or 2. This result is likely what drove the results from our system.

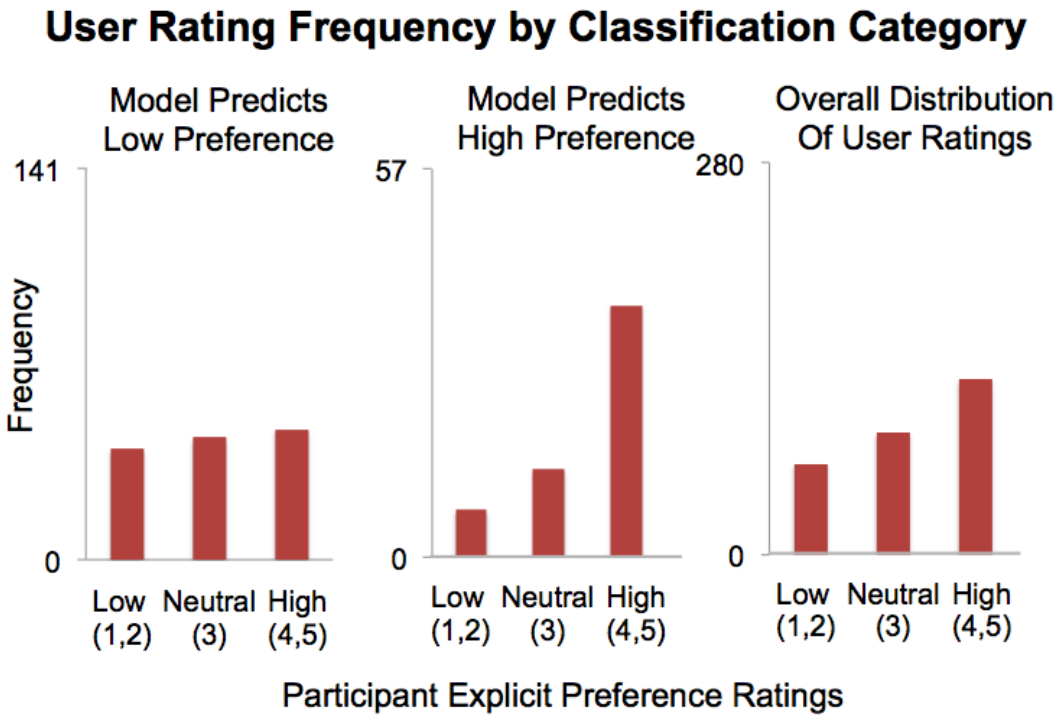


Figure 4.7: The preference model skewed towards classifying movies as low preference. However, when it classified the user's state as high preference, the user's explicit preference often agreed.

Taking a more fine-grained view of accuracy, the system precisely predicted the user's explicit preference rating in 27% of movies shown to the participant, and predicted within a single rating point for 72% of movies. Although the overall classification accuracy of the system indicates that improvements need to be made in signal processing or machine learning, we found that the mean prediction for each user rating (14 participants x 20 preference predictions) was accurate relative

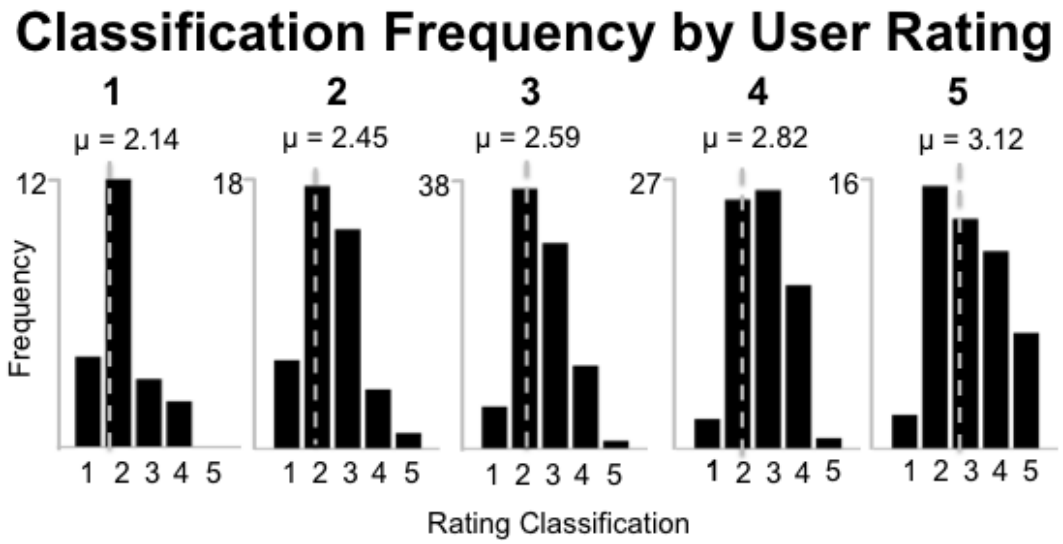


Figure 4.8: Frequency of predicted ratings for each user rating category (1-5). Although there was a wide variance of classifications for each rating category, relative to each other, the distributions of fNIRS classifications accurately mirror ratings by participants after each movie title.

to each other rating category (see Figure 4.8).

4.6.4 Anecdotal Evidence

After completing the experiment, we asked participants which of the two batches of movies they preferred. Despite there being no detectable difference between conditions, 12 out of 14 participants immediately identified the group of movies recommended by the brain condition. In addition, several participants expressed regret for not recording movie titles that were recommended to them during the brain condition.

4.7 fNIRS as Input and Future Work

In this paper, we presented a movie recommendation system that was driven by fNIRS input and performed better than a no-input recommender. We found that it

provided unique recommendations across participants and that the preference model improved with increased interaction. **These results suggest that fNIRS can be used as input to information filtering system**.

Nonetheless, as we found in our experiment, misclassifications of user state are unavoidable. This is particularly true in early systems such as our own. Due to the necessary constraints imposed by user studies, we updated our model of user preference with each movie interaction, regardless of our confidence in the input. This lead to a reliance on classifications we *knew* were probably incorrect.

In a real system, we might ignore all classifications that fall beneath some confidence threshold, ensuring that the information we integrate into the user model is more likely to be reliable. For example, the performance of our preference classifier significantly increased when over 80% of our information channels classified the user as having high preference. If we build a system that exclusively relies on that information as input, we would expect to see more personalized recommendations and sharper increases in user satisfaction.

Moving forward, there are a number of active research areas that will serve to improve these classification rates: upgrading and increasing fNIRS sensors for better coverage of the prefrontal cortex, identifying information-rich channels on the probe, establishing features of the signal that best represent preference values, and improving training periods to discern optimal examples of user preference. With advances in machine learning and brain sensing technology we expect classifications of user state to increase in accuracy. In the meantime, designers must minimize the impact of misclassifications on the user.

4.8 Implications

Despite the current challenges of translating fNIRS input in real-time applications, we believe that brain sensing has the potential to positively influence the delivery of information to the user. We examine three ways in which fNIRS input may be employed in information filtering systems.

4.8.1 fNIRS as an Augmentative Input

Because measurements of brain activity are largely an untapped source of information, performing better than a no-input control condition demonstrates that **fNIRS can be used as an augmentative input in current recommendation systems.** For example, one can envision Amazon or Netflix combining brain ratings with other implicit signals, such as purchase history and viewing history, to improve the overall accuracy of their model. Additionally, the user may be engaged in a high performance task where avoiding disruptions is critical. These implicit measures help preserve user attention because they do not force an externalization of subjective feelings onto a rating scale.

4.8.2 fNIRS as an Alternative Input

Moving away from movies and consumer products, we suggest that the true potential of neural measures lies outside current recommendation systems. For example, Baeza-Yates, Broder, and Maarek identify implicit search - or addressing user needs without a search query - as a primary challenge for the future of web search technology [8]. Thus, recommendations are advantageous in *any* information-saturated environment. In our study, participants viewed each movie website for a preset amount of time, and they were directed on a path of movies without an explicit option to diverge from it. While these measures were put in place to increase experimental control, they suggest that **brain input may improve preference measures in domains where other implicit measures are difficult to obtain.** For example, we can imagine a car radio station that naturally adapts its music to individual preferences without any intervention from the user.

4.8.3 Recommending Across Domains

Finally, we suggest that **brain sensing may improve the comparison of information from disparate sources**. Our work generalized Luu and Chau's measure of drink preference to movie preference. Thus, we do not expect dramatic changes in

the physiological response to preference across disparate information sources. Using this general measure to redirect people towards relevant information both within and across websites could prevent disruption and save cognitive resources for primary working tasks.

General measures of preference can positively impact current approaches to information filtering. Unifying and prioritizing social network status messages, for example, is a nontrivial task. People are constantly interrupted with information from Facebook, Twitter, Google Plus, Foursquare, etc. Because this information takes different forms across networks, it is unclear how to compare a Twitter message with a Facebook message without explicit responses from the user.

4.9 Conclusion

In this work, we have shown that fNIRS brain sensing can be classified in real-time and applied to information filtering systems. Although there is still significant work before fNIRS can be translated to real-world environments, we suggest that brain-computer interfaces have the potential to aid users in everyday decisions and judgments as they continue to wrestle with an increasing quantity of information. In the past, researchers have identified periods of high workload in a number of scenarios using fNIRS. While that research highlighted the potential of BCIs recognizing *when* users may need information to be filtered, our work measuring preference begins to offer a solution for *which* information should be filtered or prioritized. Given these results, we believe that BCIs may one day provide user performance and satisfaction gains in an information saturated environment. In the next chapter, we extend this work to demonstrate the potential impact of combining multiple neural measures.

Chapter 5

When and Which: Using Passive Brain Input for Intelligent Interruption

Until this point, I have shown how fNIRS can be used a measure for *how* information is presented, and as input to a system which selects *which* information should be prioritized. Previous work using fNIRS has successfully created adaptive systems that hinge on *when* information should be delivered [3, 150]. However, to fully optimize the delivery of information to a user, an system will need to integrate all of these measures simultaneously. That is, the system would deliver the relevant information at the best possible moment and in the best possible format for the user.

In this chapter, I demonstrate a notification management system that builds towards combining measures of both measures of message relevance and workload of the user. While this study focuses on the measurement of message relevance (and simulates workload), we ground the integration of relevance and workload on previous work in HCI and demonstrate how physiological signals can be used to deliver the right information at the right moment.

5.1 Background

The study of understanding and minimizing the impact of interruptions has been an active area of research in Human-Computer Interaction for more than 15 years. However, the problem of disruptions has not diminished. A 2012 Pew Internet Survey found that 25% of smartphone owners believe their device makes it more difficult to focus on a task without being distracted, and 67% of cell owners find themselves checking their phone for notifications even when they do not notice their phone ringing or vibrating [149]. While the disruption of a single notification can seem trivial, interruptions have been shown to increase levels of stress, annoyance, and anxiety [10, 24]. As a result, researchers have continued to promote the need for attention-aware systems or attentive user interfaces [9, 29].

To choose more intelligent moments to interrupt users, Horvitz and Apacible [75] devised a mathematical method for estimating the cost of interruption. Their equation takes into account both the attentional state of the user and the utility of the interruption, or the cost of a user in a particular attentional state being disrupted by a task or communication event. Following this work is a long history of research that uses various behavioral indicators to automatically model user state. However, these attempts have met with mixed success for at least two reasons: First, the utility of the interruption is extremely difficult to calculate without explicit input from the user. Second, collecting behavioral indices often requires instrumentation of both applications and the operating system.

Physiological sensing has been a reasonable solution proposed in several papers in order to gain information about user state without recording behavior throughout a variety of programs. For example, pupil size, electroencephalography (EEG), heart rate variability (HRV), and electromyogram (EMG) have all been used to detect task boundaries by correlating their signals with changes in a users workload [87, 29]. While all of these methods have shown promise in detecting opportune moments to deliver notifications, they largely have not been translated to systems that operate in real-time. In addition, these metrics do not address the utility of the

interruption, which involves factors such as the relevance or perceived value of the incoming information. This final piece is critical because there is a well-documented tension between the deferral of notifications and the awareness of the user [86, 168]. Critical messages should generally be delivered to users immediately, regardless of their current level of engagement.

In this chapter, we demonstrate the use of fNIRS brain sensing to model the relevance of incoming disruptions in a given task. Combined with workload estimations, we believe that this work moves towards a system that passively and automatically fulfills each of the terms in Horvitzs Estimated Cost of Interruption. To demonstrate this concept, we build the **CARSON system (Cognitive-based Automatic Real-Time Sending of Notifications)** that uses fNIRS data to predict and deliver notifications at optimal moments. We make the following contributions:

- We run a controlled experiment and **find that fNIRS classifies the relevance of emails at above-chance levels for 12 out of 14 participants**.
- We **demonstrate a physiological deferral policy** which is mediated by system's confidence in classifications. We find that this graded adaptation strategy maintains the user's sense of control in the system despite misclassifications.
- We use **fNIRS measures of relevance to estimate the cost of interruption in an information monitoring task.** We apply this measurement in an adaptive system to mediate the delivery of email notifications, and find that participants performa significantly better than in a maladaptive condition.

5.2 Interrupting the User

Experiencing interruptions during everyday tasks is increasingly becoming a common experience for workers. Whether it is a notice from a text message, instant message, email, or simply a colleague stopping by the office, interruptions have em-

bedded themselves as fixtures in the modern officeplace. Although they can take many forms and derive from many sources, more formally, an interruption refers to disrupting a user's attentional focus while performing a task.

While users tend try to compensate for disruptions by working more quickly, this change has been correlated with more effort, frustration, stress, annoyance, and anxiety [10, 24, 100]. In an analysis of 414 programmers, Parnin and Rugaber found that only 10% of programming sessions resume activity in less than 1 minute after an interruption [121]. Thus, the challenge is whether systems can determine more opportune moments of interruption.

Periods of low workload often correlated with task breakpoints [1, 10, 11, 85]. These breakpoints, which can be thought of as moments between tasks, create opportunity for interruptions in which users can handle incoming information without severe disruption. By contrast, periods of high workload signal that the user may be highly engaged with a task and that an interruption may be more disruptive.

When interruptions can be deferred, users often delay acknowledging notifications during periods of high workload, opting to wait until a moment of low workload [140]. However, knowing whether an interruption is important or not severely limits the value of this approach, as users seek to balance awareness and interruptions [76, 158].

While interruptions have clearly been shown to be harmful to user performance, users tend to view notifications as a mechanism for passive awareness of information [86]. Curiousity drives users' attention to the inbox, where they infer email utility based on top-level cues such as subject name [168]. Users can be so intent on maintaining this level of awareness that they are actively willing to be disrupted in order to be kept aware of the state of their inbox. In a study of Microsoft developers and managers, turning off notifications resulted in some users checking their email account *more often* [86].

Given this tension, it is clear that the context of an interruption is important [103]. For example, Cutrell et al. found that messages that are relevant to the user are less disruptive than irrelevant messages [38]. It is with this consideration

of context and engagement that researchers have attempted to build interruption management systems.

5.2.1 Interruption Management Systems

For more than two decades, there has been an effort to build systems that are aware of how users manage their attention [75, 77, 101]. Bailey and Konstan write that "Attention-aware systems could mitigate effects of interruption by deferring presentaiton of peripheral information until coarse boundaries are reached during task execution" [9]. However, the task of constructing models of interruption is not trivial, as the system must have an understanding of both the user as well as the user's context.

Horvitz et al. devised a mathematical approach to estimating the *expected cost of interruption (ECI)* [75]:

$$ECI = \sum_{j} p(A_j|E)u(D_i, A_j) \tag{5.1}$$

where $u(D_i, A_j)$ is the cost of a user in an attentional state A_j being disrupted by a task or communication event D_i, and $p(A_j|E)$ is the probability of the attentional state, conditioned on evidence stream E. Computing the expected cost of all interruptions requires the summing of all utilities, weighted by the likelihood of attentional state.

While Horvitz computed $u(D_i, A_j)$ by having users assess the cost that they would be willing to pay in order to avoid the outcome tuple (using dollars), ideally, a system could infer the utility of a message without explicit user action. However, most work on mitigating disruption focuses exclusively on modeling the users' attention in order to estimate optimal moments for interruption [75]. In addition, modeling engagement often requires special instrumentation at the application or operating system level.

5.2.2 Physiological Computing Systems for Interruption

Physiological computing has been suggested as an alternative input to models of interruption in order to improve the quality of input to the system [87, 30, 29]. In addition, the potential generalizability of physiological would reduce the need for instrumentation. To date, there have been at least two investigations of brain or body signals as input to a physiological computing system for interruption.

Chen and Vertegaal found that both Heart Rate Variability (HRV) ($r = 0.96$) and Electromyogram (EMG) ($r = 0.85$) correlated with participants' self reports of interruptability [29]. Using these measures, they constructed *Plog: A Physiological Weblog* which broadcast a simple visualization of a user's interruptability. While Chen proposes the use of physiological metrics to build a 'Physiologically Attentive Interface', no evaluation of such a system was ever conducted [30, 29].

Similarly, Iqbal [87] constructed the *MeWS-IT (Mental Workload Based System for Interruption Timing)* system, which leveraged pupil size to estimate the user's workload and provide more optimal interruptions. Iqbal proposed that workload measures could be combined with other external cues to determine opportune moments, modifying Horvitz' original equation:

$$COI_{combined} = W_{wl} * COI_{wl} + W_{ec} * COI_{ec} \tag{5.2}$$

where W is a weight that can be manipulated based on the quality of the data source, wl is the workload, and ec is the external cues. We will employ a similar approach in this study. Similar to Chen [30], the proposed system was not evaluated.

Finally, while each of these studies focus on capturing the user's workload or attentional state with physiological measures, they largely neglect the utility. In this chapter, we attempt to extend this work by constructing and evaluating an interruption management system that uses physiological metrics of message relevance.

5.3 Using fNIRS for Relevance

Although very little work has explicitly looked at fNIRS as it relates to relevant information, there is a body of literature that suggests this kind of detection may be possible. For example, fNIRS has been used on a single trial basis to detect and classify signals that correlate with preference [80, 98, 126](or positive experience [94]). Referring to the previous chapter in this dissertation, we demonstrated that this signal could be classified in real-time and used in an adaptive scenario.

In this work, we focus on work by Solovey et al. [153, 150], which demonstrated using fNIRS to detect the impact of multitasking. Their critical finding was that there are signal differences between branching, or performing a secondary task while holding in mind the goals from the primary task, and delay task, when a secondary task is largely ignored. We hypothesize that irrelevant notifications will force a user to engage with content long enough to induce branching/dual-task signals. Furthermore, relevant information actually aids a users primary task, potentially reducing the resources required for it. In our study, we attempt to capture this signal as users engage with relevant and irrelevant information.

5.4 Improving Real-Time Classification of fNIRS Signal

Starting with work described in Chapter 3, the biocybernetic loop used in this study employs many of the same filtering and analysis techniques. However, two significant additions were made to the our classification system in order to improve its use in a real time environment: feature definition and classification probability estimates.

5.4.1 Feature Definition

In the previous chapter, examples were constructed using the raw fNIRS data at each time point as an individual feature. However, combining such a large feature space with the relatively small training set acquired during calibration can result in what is commonly known as the curse of dimensionality. As data becomes increasingly high dimensional, every example may appear to be sparse in the context of the enormous

feature space. This is particularly problematic in scenarios where it is difficult to acquire a large training set due to timing constraints. The practical consequences of the curse of dimensionality is the construction of models that overfit the training data and do not generalize to other tasks.

As we surveyed in the related work section, a number of more descriptive, high-level features have been shown to be successful in offline analysis. For this study we focus on the mean change in signal during a trial as well as a best-fit slope of the signal for each trial. Thus, we generate 32 features (4 sources x 2 detectors x 2 wavelengths x 2 features) for each labelled example to our model.

5.4.2 Using Probability Estimates for Physiological Computing Systems

Because of the inherent complexity of the brain and general noisiness of physiological data, we primarily rely on binary classification schemes (low v high workload, low v high relevance) in order to achieve the highest possible classification accuracy. However, directly mapping adaptive mechanisms that are triggered by one class or another can result in jarring responses by the system, which may constantly be reacting to predicted (and possibly misclassified) user state.

To construct a more graded approach to physiological computing systems, we turn to machine learning algorithms that not only provide a classification, but also a probability estimate of that classification. For example, the system may classify two moments during interaction as high workload, however, the first may be assigned a probability estimate of 60% while the second may have a probability estimate of 98%. Given these varying levels of confidence, the system should respond to the user differently. We employed this approach in a previous study by Afergan et al. discussed in Chapter 2 [3]. In this study, we use the probability estimates in order to manipulate deferral policies, discussed further in section 5.5.2.

5.5 CARSON: System Overview

CARSON (Cognitive automatic real-time selection of notifications) is a backend system that is designed to select opportune moments to interrupt the user based on physiological input. It is able to compute the cost of interruption for a user by predicting the relevance of an incoming notification as well as working memory load of a user. In this section, we discuss high-level details of the CARSON system.

5.5.1 Calculating the Cost of Interruption

To calculate the cost of interruption of a single message, we return to Iqbal's COI calculation:

$$COI_{combined} = W_{wl} * COI_{wl} + W_{ec} * COI_{ec} \tag{5.3}$$

where W is a weight that can be manipulated based on the quality of the data source, wl is the workload, and ec is the external cues. Given previous research by Iqbal and Bailey [85], we use workload as an indicator of the attentional state of the user, and we use message relevance as an indicator of the utility of interruption (or external cues). CARSON listens to incoming classifications from OFAC (online fNIRS analysis and classification system), described initially in Chapter 4.

Considering both Horvitz' *ECI* and Iqbal's *COI*, as well as the workload and relevance metrics that CARSON incoporates, we construct the following *COI* equation for use with physiological input.

$$COI = (W_{wl} * COI_{wl})(W_{util} * COI_{util}) \tag{5.4}$$

Workload (wl) varies between low and high, where the COI during periods of high workload is 1 and the COI during periods of low workload is 0. Similarly, communication events can be either highly relevant or irrelevant (low relevance), where COI during periods of high relevance is 0 and COI of low relevance emails is 1. Similar to Iqbal's equation, W represents a weight based on the quality of the data source. In this case, we use the probability estimate of fNIRS classifications

to estimate the quality of the data source. For example, a classification of high workload with 75% confidence will result in $(W_{wl} * COI_{wl}) = 0.75$.

The practical implication of this equation is that the COI is computed to be high when a message of low relevance (or high utility that is approaching 1) is delivered during a period high user engagement. A message that is highly relevant to the primary task (or has a utility approaching 0) is always delivered, regardless of the user's engagement.

For a brief overview of the system, a relevance model of incoming notifications is constructed using only physiological data. We use this model to estimate the utility of the message, or $(W_{util} * COI_{util})$ in the previous equation. Next, we apply this model to a working environment in which workload levels are either simulated or estimated using fNIRS. These workload levels are used as a proxy for $(W_{wl} * COI_{wl})$. Combined, we estimate the cost of interruption of an incoming notification to the user and attempt to defer the notification until a more opportune moment.

5.5.2 Physiological Deferral Policy

When the estimated cost of interruption exceeds a predefined COI threshold, CARSON defers the interruption until a more opportune moment, placing the incoming notification into a deferral queue. Every second, as the users physiological state changes, the system polls all of the notifications in the deferral queue and recalculates their cost of interruption. If the new cost is below the COI threshold, the message is immediately delivered to the user.

However, strictly adhering to this policy ignores the tension between the cost of interruption and the level of awareness that users prefer to have about their information. It is not acceptable to defer a notification indefinitely simply because the system has not determined a good moment for interruption, especially since model-based systems may misclassify either the users attentional state or the utility of the message. As a result, CARSON decreases the utility of the message (or increases the perceived relevance of the message) over time, thereby decreasing the COI of the message, and making an interruption more and more likely the longer it

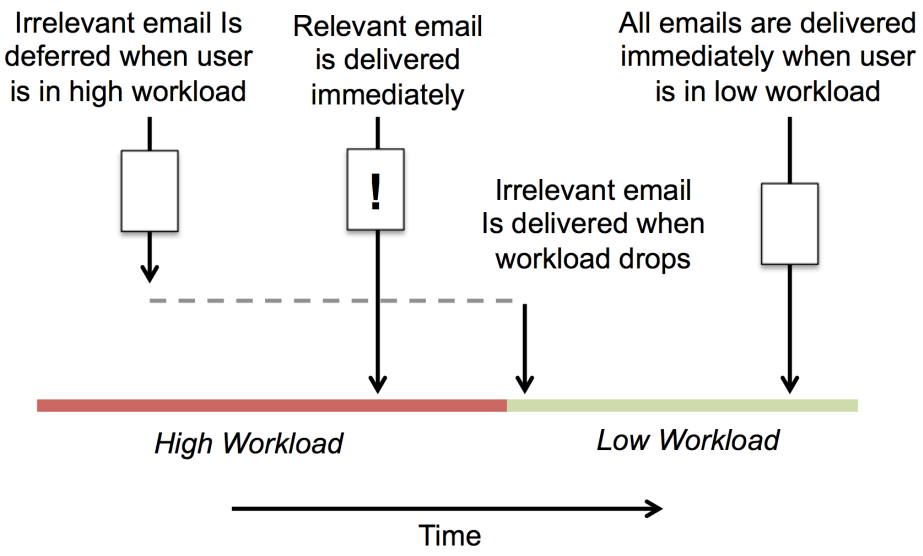


Figure 5.1: If the system had perfect confidence in the physiological input, irrelevant interruptions would be deferred until the user is in a state of low workload. Relevant interruptions would always be delivered immediately.

sits in the deferral queue.

This approach has an added benefit to physiological systems. Since the utility is calculated as a function of the confidence (probability) in the classification, the more confident that CARSON is that an incoming message is low-priority, the higher its estimated cost of interruption will be. As a result, CARSON will allow the message to sit in the deferral queue longer before being delivered to the user. Likewise, if the system is not very confident in its prediction, the message has a lower maximum deferral time. This strategy transforms an interaction that is largely binary and discrete (interrupt or do not interrupt) into one that is continuous.

5.5.3 Optimizing Adaptation Parameters

One of the design challenges in constructing an interruption notification system is selecting a cost of interruption (COI) threshold for the user. If the threshold is set too high, all messages will be deferred, regardless of their importance. If the threshold is set too low, all messages will immediately be delivered, regardless of how disruptive they may be.

This problem is worsened in the context of physiological metrics, where the challenges in calibration and translation to real-time measurements results in variations in classification accuracy (or confidence). For some participants, the system may never receive high-confidence values from the classifier, regardless of how critical the message is. As a result, assigning a neutral COI threshold would naturally skew interaction with the system towards immediate delivery. Over increased use with the system, this problem may naturally fix itself. However, in the context of an experiment, this skewing can have a strong negative impact on interaction with the system.

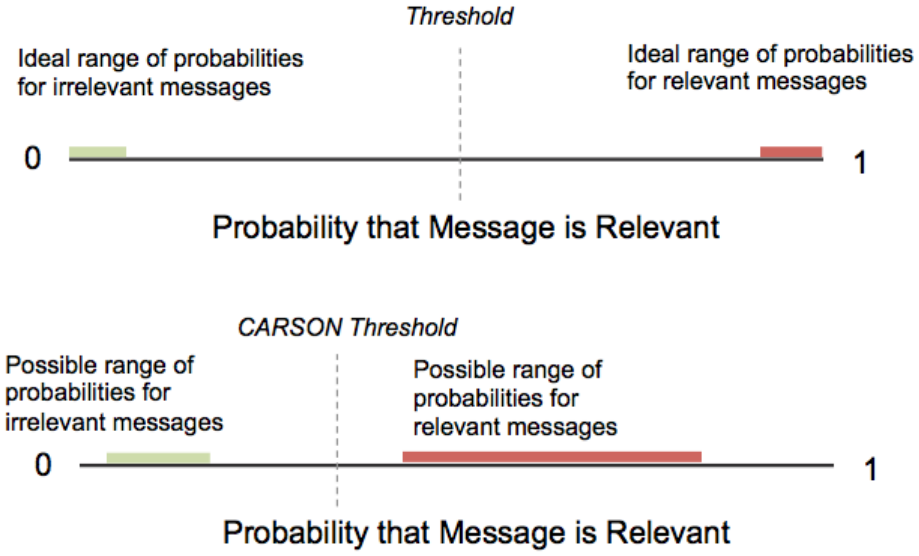


Figure 5.2: Rather than using a single cost of interruption theshold for all users, CARSON attempts to optimize a threshold based on the distribution of classification probabilities for relevant and irrelevant messages

As a result, CARSON is designed to manage the COI threshold by analyzing the distribution of classification probabilities for both high utility messages and low utility message and removing outliers (two standard deviations above/below the mean). It then assigns a COI threshold by determining the midpoint of the resulting mean classification proabilites for relevant and irrelevant messages (fig. 5.2). This approach creates personalized adaptation parameters that may improve interaction

even during short experimental sessions.

5.5.4 Extending COI: Integrating Cost of Delivery

In this chapter, we will primarily investigate the integration of fNIRS metrics of relevance with estimations of the user's workload. However, we began this thesis by also considering *how* information is presented. The design behind CARSON is to eventually optimize for three different metrics: engagement of the user, utility of the message, and delivery mechanism of the message. To accomodate for this, we can make a simple addition to the previous equation.

$$COI = (W_{wl} * COI_{wl})(W_{util} * COI_{util})(W_{del} * COI_{del}) \tag{5.5}$$

where COI_{del} now represents the cost of the delivery mechanism or presentation style given the user's context.

In this model, we can envision a classifier that attempts to measure the impact of visual design, for example, when engagement and utility are held constant (much like the fNIRS measurements in Chapter 3). COI_{del} may also measure the impact of delivery for a given device. For example, we could assign a cost to notifications presented on a smart phone vs. notifications presented on a Google Glass. While we will not explore this third term any further in the scope of this thesis, a system that truly optimizes the delivery of information will likely incorporate an understanding of the impact of the delivery mechanism.

5.6 User Scenario: Information Specialist

To test the ability to detect message relevance and apply it to a working environment, we constructed a hypothetical scenario for participants to act as information specialists for a news station. Their objective was to monitor a Twitter feed about the days event and periodically retweet messages to keep their followers informed. To do this, they clicked on every 3rd tweet of each topic they were assigned to follow (Figure 5.3). These topics were assigned by the system, and could vary in both

content and number (with more topics being more difficult). A new tweet entered the participant's stream at a random interval between 1500 and 3500 ms.

Figure 5.3: Monitoring a Twitter feed. Participants were asked to retweet (or click) every 3rd tweet of each topic

While tracking these topics, they were also instructed to respond to incoming emails from their bosses. These bosses either sent them emails that were relevant or irrelevant to the monitoring task. Relevant emails specified that the user should prioritize looking for a tweet from a specific user, and until that tweet was identified, they no longer had to retweet other statuses on the topic. Irrelevant emails simply acted as distractors, mentioning tweets from topics that other information specialists were monitoring. In both cases, participants indicated the relevance of these emails in their response (Figure 5.4).

5.6.1 Interface Details

As new tweets entered the system at the top of the simulated Twitter client, they pushed down old tweets. Tweets were eventually pushed off the screen without any mechanism for the user to scroll and see them again. This mechanism enforced the

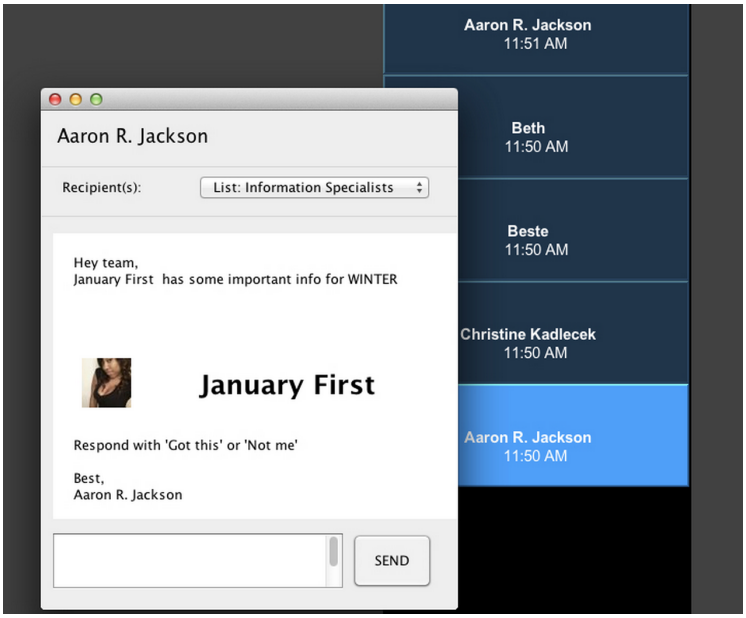


Figure 5.4: Participants also received emails that were either relevant or irrelevant to their information monitoring task

temporal urgency of the task. When participants clicked on a tweet (retweeting it), it was highlighted in red. When topics changed, the bar where topic labels were displayed (Figure 5.3) lit up for a moment and a notification sound played. Emails told participants what the response options were (Figure 5.4). When an email was received, a notification sound played (different than the topic switch notification).

5.7 Experiment: Email Relevance

We adopt a methodology similar to the one employed by Googles Priority Inbox as well as spam filters. In these systems, users tag information that they deem to be important (or in the case of spam filters, spam). Later, as a new email enters the system, GMail uses a number of features about the email to predict and flag important priority messages. The goal of our experiment is to demonstrate a similar system that is completely independent of explicit user input, instead relying on physiological data to estimate the relevance of new information. In the following table, we compare these two approaches:

Current Practice	CARSON
	1. *Calibration:* Using messages that are known to have high or low relevance, listen to fNIRS data immediately following interaction and build a model to classify emails based exclusively on email data
1. *Assigning Utility:* Email utility is based on emails that are manually flagged by users	2. *Assigning Utility:* All new emails with unknown utility are passively assigned a utility value based on fNIRS classifications
2. *Creating an Email Model:* An email relevance model is constructed (and updated) based on features of flagged emails.	3. *Creating an Email Model:* Instead of using email features to cluster emails, CARSON uses a wizard-of-oz approach and clusters them based on their true label (relevant or irrelevant). However, the system assigns each clusters utility using only fNIRS classifications.
3. *Predicting Utility:* New emails are compared against the email model to predict their relevance.	4. *Predicting Utility:* Since CARSON is aware of the true label of incoming emails, that label is used to look up the avg. fNIRS utility calculated in the previous step. This value is inserted into the COI equation.

5.7.1 Calibration: Training an fNIRS Relevance Classifier

The first step to detecting email relevance in real time is to build a model that uses fNIRS signals as input and outputs a prediction of an email's relevance. In order to compile a set of labelled examples to build such a model, we used a controlled experiment in which the CARSON system was aware of the true label of each incoming email (e.g. whether a message is relevant to a user or not).

Returning to our information specialist scenario, participants performed a 15 minute calibration session in which they were interrupted by 14 emails that were relevant to their task and 14 emails that were irrelevant. Interruptions occurred, on average, every 23 seconds, and the order of relevant vs. irrelevant emails was randomized. During this time, the difficulty of the users monitoring remained constant (i.e., the number of monitored topics stayed the same) to prevent confounds from the primary task.

The online fNIRS classification system recorded 20 seconds of fNIRS data following the opening of each email and extracted high-level features of the fNIRS data described in the previous section. These features were then used to construct

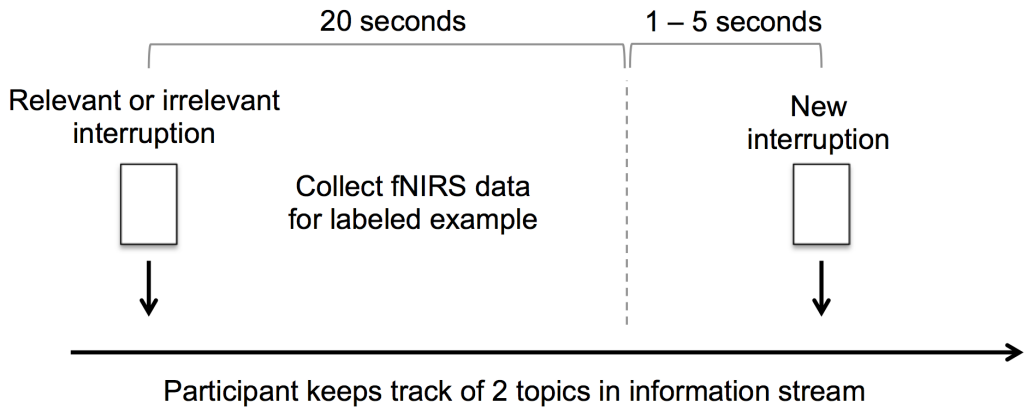


Figure 5.5: In the calibration task, participants were interrupted by 14 relevant emails and 14 irrelevant emails

labelled examples (each email was either labelled as relevant or irrelevant), and following the calibration period, a model was constructed to differentiate between relevant or irrelevant emails.

5.7.2 Assigning Utility to Emails using fNIRS

Although cross-validation is a suitable approach to testing the success of a model, it often does not capture the difficult task of classifying new data in real time - a hurdle that any brain-computer interface must overcome. As users interact a with new emails, a successful implementation of our relevance model should be able to identify email relevance at above-chance levels using only fNIRS input.

After the calibration session, participants performed a similarly designed session in which participants interacted with 20 emails (10 relevant, 10 irrelevant) over a ten minute period. However, in this iteration we treat the utility of incoming emails as unknowns that can only be assigned using fNIRS.

During this time, a sliding window of the past 20 seconds of fNIRS data was continually fed to our model, retaining classifications approximately 3 times per second. Based on the results of a series of pilots, we found that the classifications that most heavily correlated with ground-truth occurred between 15 and 25 seconds following interaction with an email. Therefore, we recorded and averaged

the classifications during this period, and assigned this value as the emails utility.

5.7.3 Building an Email Relevance Model

At this stage, a deployed system would likely cluster emails that demonstrated similar utility, then search for common features between those emails. This would enable the system to predict a new emails utility before interaction.

Since this step was not a focus of ours (and would introduce an extra layer of experimental complexity), we use a "wizard of oz" approach, simulating the construction of an email relevance model. Instead of using message features to cluster emails, CARSON clusters them based on their true label (relevant or irrelevant). However, the system assigns each clusters utility using only fNIRS classifications, acting as if it is unaware of true relevance of the message. As a consequence, the high relevance cluster has the potential to be incorrectly assigned high utility because of poor fNIRS classifications. For the following sections, we use this model to predict the utility of incoming emails.

5.7.4 Applying Relevance to Interruption Deferral

In order to apply our relevance model to the information specialist example, we examine its impact in three different conditions (the effect of which is shown 5.6). For each of these three conditions, participants interacted with seven 45-second trials of interruptions during high workload (monitoring two topics) and seven during low workload (monitoring one topic). These trials were presented in random order, which meant that the number of task switches from high workload to low workload (or vice versa) could vary. On average, participants viewed 20 relevant notifications and 20 irrelevant notifications during this testing period.

- **Adaptive:** The cost of interruption is calculated using predictions from the email relevance model. To isolate the impact of these predictions, we estimate workload based on the number of topics being presented to the user (2 topics = high workload, 1 topic = low workload).

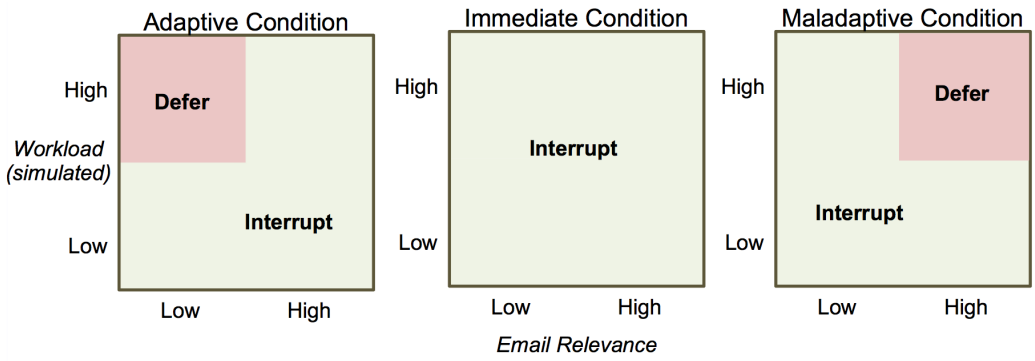


Figure 5.6: The effect of different conditions on interruption and deferral policies.

- **Immediate:** Notifications are delivered immediately, regardless of the COI.
- **Maladaptive:** Similar to adaptive, however, we inverse the predictions from the email relevance model. For example, if the model predicted there was a 0.7 probability of an email being relevant, we use 0.3 as input to the COI equation. This condition gives us an indication if the classifications we are receiving are meaningful. If they are, there should be a significant difference between performance in the maladaptive and adaptive conditions.

5.7.5 Measures

We investigated the following dependent behavioral measures:

- **Twitter Task Accuracy:** We use Levenshtein distance to determine the accuracy of users in the Twitter task [96]. Once user input is transformed to a string and compared to a string of correct system input, Levenshtein distance calculates the minimum number of edits (single-character) to change one string into another. This approach allow us to gauge user accuracy in such a way that a single counting error does not cascade through the entire trial.
- **Relevance Miscues:** When the participant received a relevant email, they were instructed to stop counting tweets in the specified category until they saw their high-priority tweet. Thus, if the participant did not respond to relevant emails quickly enough, they may erroneously continue to click tweets

in that particular category. This metric counts the number of tweets that were erroneously clicked following the delivery of a relevant email.

- **Response Time:** The amount of time from when a tweet entered the information stream until the tweet was selected (only relevant for selected tweets).

In addition, we analyzed survey data that was given to participants following each condition:

- **System Survey:** Based loosely off of a survey used by Gajos et al. [55], asks questions that are more specific to the user's experience with an adaptive interfaces: How useful did you find the system? How confusing did you find the system? How in control of the system did you feel? How efficient did you feel the system was?

- **NASA-TLX:** Shown in the Appendix, asks task-related questions about mental load, physical load, temporal load, performance, effort, and frustration.

5.8 Results

We ran this study with 14 participants, aged 18-23 with a mean age of 21. The order of conditions was counterbalanced across all participants.

5.8.1 Model Building: Different People, Different Success

During each query to the relevance model, the system returns a probability estimate from 0 to 100% of whether the user had interacted with a high-relevance email. To determine the success of our model, we calculated the mean probability estimate of each email that user encountered. Finally, to prevent outliers from disrupting our analysis, emails that received probability estimates more than two standard deviations from the mean were thrown out from the analysis. We hypothesized that the average probability estimate for relevant emails would be higher than those for irrelevant emails. These results can be seen in fig 5.7.

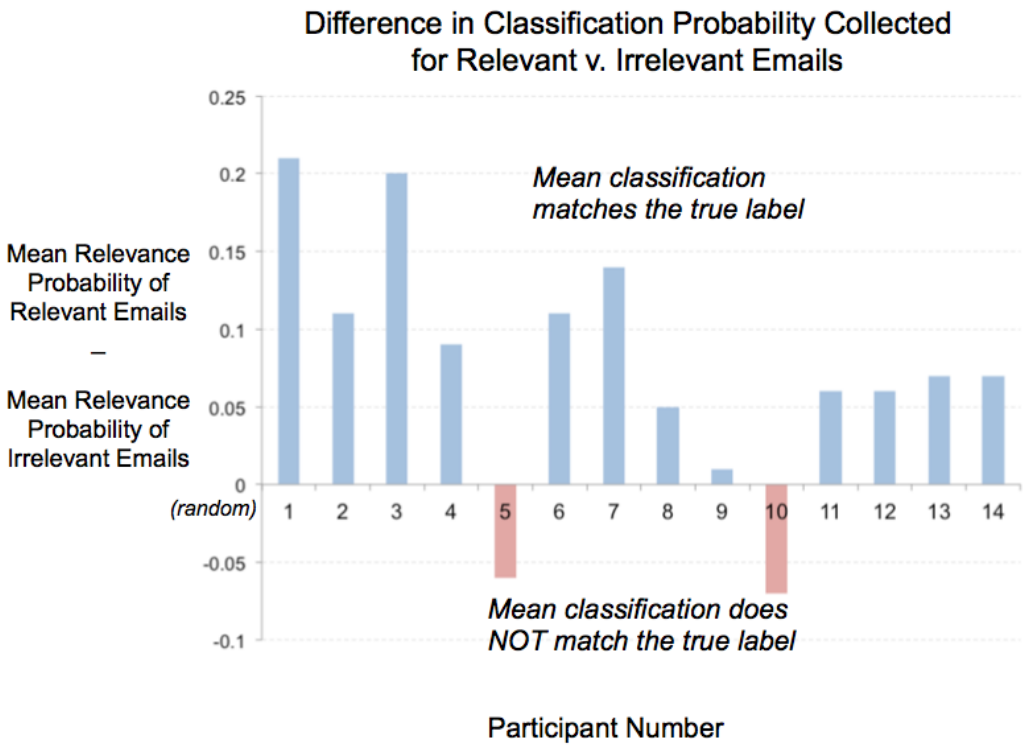


Figure 5.7: We calculated the mean probability estimate that an email was relevant for each message that participants interacted with. For each participant, we show the average probability estimate for relevant emails (true label) - the mean probability estimate for irrelevant emails. A difference of 0 would indicate completely random classifications. A perfect classification rate would result in a difference of 1, as all relevant emails would receive a 100% probability of being relevant and all irrelevant emails would receive a 0% probability of being relevant.

We found that 12 out of 14 participants exhibited a higher probability estimate with high relevance emails than low relevance. However, for those 12 participants, the average difference between high relevance and low relevance emails was only 8%, and varied from 1% to 21%. This finding suggests that while there are signal differences between low relevance and high relevance emails, both in offline and online environments, these differences may be subtle. The resulting challenge for physiological systems is to drive adaptations with even subtle differences in fNIRS signals without disrupting the user's interaction.

It's also worth mentioning that the training period and testing period were truncated in order to adhere to the normal timing constraints of a user study. It

is possible that providing the system with more examples would have significantly improved classification. In addition, as participants familiarized themselves with the task, it is possible that they adapted different strategies that fundamentally changed the neural response to relevant and irrelevant emails. Despite these relatively low accuracy levels, we will discuss how to apply such a model to an adaptive system in the following sections.

5.8.2 Performance

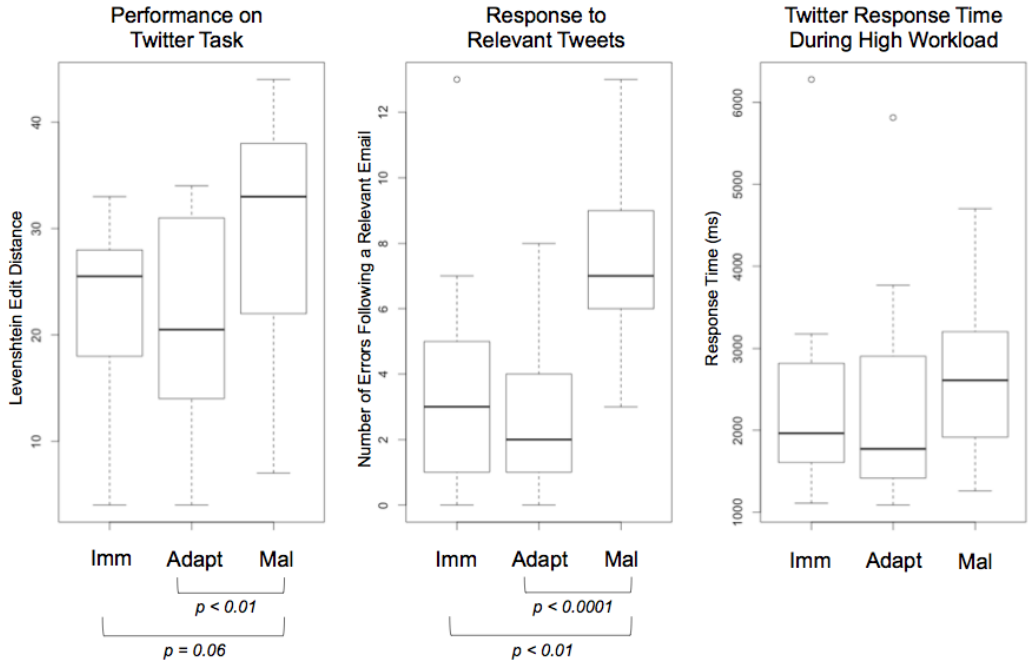


Figure 5.8: Boxplots of major behavioral metrics that describe the user's interaction in the Adaptive, Maladaptive, and Immediate conditions. The whiskers represent the max/min values, excluding outliers. Outliers are assigned by being more/less than 1.5 times the value of the upper/lower quartiles.

We found that Twitter task accuracy (Levenshtein edit distance) in the **Adaptive** condition ($M = 21.57, SD = 9.6$) outperformed the **Maladaptive** condition ($M = 29.36, SD = 11.76$) ($t(13) = -3.1049, p = .008$). In addition, the number of relevance miscues was significantly lower in the **Adaptive** condition ($M = 2.71, SD = 2.55$) than the **Maladaptive** condition ($M = 7.64, SD = 2.65$)

$(t(13) = -7.7859, p < .0001)$.

While the **Immediate** condition ($M = 3.64, SD = 3.52$) also outperformed the **Maladaptive** condition in relevance miscues ($t(13) = -3.6056, p = .003$), there was not a significant difference in task accuracy ($t(13) = -2.0624, p = .06$). Finally, despite an improvement in the overall performance means (fig 5.8), we found no significant difference between the **Adaptive** condition and the **Immediate** condition. In the next section, we discuss the impact of classification accuracy on behavioral results.

5.8.3 Attitudes Towards Adaptation

Turning to the surveys, a couple of participants noted differences in the responses of the different conditions. For example, one participant responded that the **Adaptive** condition was *"Easier to handle/manage than the last one"* (referring to the Immediate condition). When the participant interacted with the **Maladaptive** condition, they wrote *"Much less efficient with this system. Got lost"*.

However, these observations were rare. Overall, we found no differences between conditions in any question from the NASA-TLX or system survey. This attitude is reflected best in one participant, who wrote *"It seemed like it was the same as before?"*, referring to a comparison between the **Adaptive** and **Maladaptive** conditions. Despite significant differences in performance between the adaptive and maladaptive conditions, the responses across all participants were indistinguishable. We discuss the implications of this finding in the next section.

5.9 Discussion

5.9.1 Increased Accuracy, Increased Impact

In CARSON, we employed a deferral policy that was heavily dependent on the system's perceived confidence in its classifications (calculated using probability estimates). While we observed trends across all participants that indicated a differentation between relevant and irrelevant emails, at an individual level, the differentation

was small. As a result, the system constrained its maximum deferral time as short as 5-7 seconds in order to preserve user awareness.

While we believe that this is a suitable approach to translate physiological computing to realistic environments, within the context of the experimental design, these deferral periods did not extend long enough to reach natural breakpoints, or optimal moments of interruption. This translated to small (or negligible) differences between the *Immediate* condition and the *Adaptive* condition for many participants.

Given stronger classification results, the system would naturally extends its maximum differal period and reach natural breakpoints. If the signal classification cannot be improved, another option is to change the adaptation parameters in order to extend the maximum deferral period. In the next chapter, we discuss the potential of optimizing adaptation parameters at an individual level.

5.9.2 Indistiguishable Adaptation

We found that participant attitudes to all versions of our system were virtually indistinguishable. These results differ from studies that either discover participant attitudes to align with performance gains or that the presence of system interventions disrupts the user's perception of control.

This finding may be an indication that physiological metrics are well suited for use in interruption deferral systems. Users are generally not aware of the timing or content of incoming notifications until they arrive. Despite the presence of system manipulations, many participants could not differentiate between systems.

In addition, the lack of differentiation between conditions also suggests that the physiological deferral policy successfully provided a subtle and gentle adaptation. While users performed significantly better in the adaptive and immediate condition when compared to the maladaptive condition, the maladaptive condition was not disruptive enough to register on any dimension of the NASA-TLX or system survey.

5.9.3 Multiple Measures: Workload and Relevance

In this experiment, we explored the use of fNIRS to detect the level of support an incoming notification provided the user, distinguishing between relevant and irrelevant messages. However, we estimated user's workload based exclusively on task (following 1 topics or 2 topics). In a real system, the user's workload is likely would not conform perfectly to task designation and would fluctuate based on other environmental factors (such notifications, other working tasks).

Previous work has used brain-based metrics of workload as input to adaptive systems [3], and CARSON was constructed to be able to integreate this second measure. Similar to relevance, CARSON keeps a running tally of probability estimates that the user is in a state of high workload. When the system sees an incoming message, it evaluates the average workload over the past 15 seconds and inserts this value into the COI equation:

$$COI = (W_{wl} * COI_{wl})(W_{util} * COI_{util}) \tag{5.6}$$

We plan to investigate the integration of these two brain-driven metrics in a future experiment.

One of the primary challenges in considering a system that uses multiple user states is that the physiological response to these states may overlap or interfere with each other. CARSON attempts to circumvent this problem by distinguishing between continuous measurements of workload and relevance measurements that are triggered by discrete events. The system is designed to constantly monitor and store the user workload while they are engaged with their primary task. However, as soon as a notification arrives, the system assigns message utility (or relevance) based on how this secondary task impacts the user's workload.

In general, the exploration of this topic will be critical in the future, as it's unlikely that BCI will be able to successfully capture the entirety of the user's context with a single user state. In the next chapter, we motivate the investigation of fNIRS as a complementary source of input as a fruitful direction for future research.

5.10 Conclusion

In this chapter, I presented the CARSON (cognitive automatic real-time selection of notifications) system, an intelligent interruption system that uses physiological metrics as input. Although previous work on notification systems use metrics of workload to estimate suitable breakpoints for users, we focus on message relevance. We found that fNIRS is capable of detecting small, but significant differences between relevant and irrelevant messages. We also designed a interruption deferral policy for physiological computing systems that modulates its intensity based on its confidence in the user state model. We found that this mechanism allowed for subtle manipulations of the system without negatively impacting the user's perception of the system. Finally, we suggested that message relevance may be integrated with the user's workload to create a system that delivers the right information at the right moment.

Chapter 6

Conclusions

6.1 Summary of Work and Contributions

The use of brain data as passive input to intelligent systems has the promise of improving the bandwidth between users and their computing devices. However, the application and use of these passive signals is still largely unexplored. In particular, there have been few examples of successfully using fNIRS to drive adaptive applications that improve experience for everyday users. In this thesis, I explored the potential of fNIRS in information delivery systems by investigating the suitability of using fNIRS measurements in three application areas:

- **How information is presented to the user**. We found that the fNIRS signal can offer insight into the use of information visualization during complex tasks. We observed signals that reflected participants' subjective experience of workload and differed based on the individual.

- **Which information is presented to the user**. We found that we can detect fNIRS indicators of preference and use them to drive an information filtering system, improving user satisfaction with new information (in this case, movie recommendations).

- **Combining which and when information is presented to the user**. Grounded in previous literature, we found that the combination of measures

has the potential improve user interaction. We used fNIRS measures of message relevance to construct an intelligent interruption system and proposed an interruption deferral policy based on physiological measures. We also outlined a future system that incorporates *how* information is presented.

In addition to demonstrating the feasibility of using fNIRS in these systems, I implement designs to facilitate the use of physiologically-driven systems, which frequently suffer from misclassification:

- In Chapter 4, I implemented **a physiological approach to tagging information**. We polled multiple classifiers (based on source-detector pairings) to map low v. high preference classifications onto a more granular, 5-star rating system. Movies in which the preference classifiers did not produce consistent classifications were assigned neutral (or uncertain) ratings in the system. This mitigated the impact of misclassifications and allowed the system to improve user experience despite relatively low classification rates. Motivated by the system's classification rates, we also suggested that future systems opt not to tag information with a preference prediction if confidence levels were low.

- In Chapter 5, I implemented **a physiological approach to deferring incoming notifications** in which the maximum deferral time was modified depending on the confidence of user state classifications. The result of this design was that participants felt no loss of control in the system despite misclassifications.

Finally, based on previous literature, we built on top of the work of Girouard [59] and Solovey [151] real-time fNIRS system in order to enhance the detection of user state with fNIRS signals. We made the following significant changes to the system:

- **Probability estimates of classifications:** rather than simply receiving a nominal classification, the system provides a probability estimate that the reported classification is correct. In our case, this probability estimate is facil-

itated by LibSVM, however, many machine learning libraries provide similar functionality.

- **Feature extraction and selection mechanisms:** in Chapter 2, I discussed the large feature space that researcers are exploring to classify fNIRS input. We designed a module that allows researchers to shift sets of features and apply them to real-time classification.

- **Standardized input and output:** Each time real-time classificaiton system is used, it outputs all of the raw fNIRS data, as well as the labels and timing for markers being sent to the system. We redesigned the system to be run on its own output, allowing researchers to replay data with the precise timing of any experimental session. While this modification was not discussed in this thesis, it was critical for formulating effective experimental designs.

Taken together, this work provides evidence for the continued development and application of adaptive systems that are driven by fNIRS input. In addition, it provides designs guildelines and areas of application for systems that utilize passive user input to optimize interaction. While the usage of passive brain-computer interfaces in everyday scenarios can still be considered in its infancy, these methods will aid the use of BCI as it wrestles with noisy data and translation to real world environments.

6.2 Future Work: Improving fNIRS BCIs

There has been significant process in the state of fNIRS BCIs over the past few years. However, the field is still young with few strong examples to point to. As there becomes increasing interest in moving fNIRS to more real world scenarios, we point out four potential areas of future research that would make fNIRS-driven BCIs more robust: improving the user of fNIRS as a complementary input, reducing the calibration (or training) time, improving the calibration task, and personalizing adaptation parameters.

6.2.1 Improved Understanding of fNIRS as Complementary Input

As we have motivated throughout this thesis, fNIRS has many desirable features for researchers and practitioners - easy setup, relatively resistant to movement artifacts, and detects physiological parameters that are similar to fMRI. However, it is unlikely that future of physiological interfaces will be ruled by a single sensor. As we motivated with the comparison between EEG and fNIRS in Chapter 2, different sensors have distinct advantages and disadvantages. In the future, intelligent information delivery systems will likely integrate and leverage information from many sources of passive input from the user [47]. As a result, it should be an objective for future work to differentiate between physiological sensors, and working contexts that offer unique or redundant information.

This approach, or *sensor fusion*, has been explored often in the context of other physiological sensors. Fairclough writes that the mapping between a physiological measures and the pyschological construct as having one of three mappings [47]:

- **One-to-One:** a physiological signal maps to one and only only user state. Similarly, a user state maps to one and only one physiological signal. This kind of mapping is extremely rare.
- **One-to-Many:** a physiological signal maps to multiple user states. For example, an increased heart rate may map to either emotional arousal or increasing workload.
- **Many-to-One:** perhaps the most common occurrence, many physiological signals work as indicators of a single user state.

Turning to fNIRS, existing work that collects fNIRS measuresments alongside physiological measures (for example, EEG, heart-rate, etc.) rarely translate the results to clear design guidelines. Understanding many-to-one mappings across sensors may help with the improved accuracy of real time classification, offering validation checks to a predicted user state. Similarly, comparing and combining multiple sensors may help solve one-to-many mapping problems that can plague a

single sensor. Combining passive input may also aid the sluggish temporal response of the BOLD signal, allowing for faster user state detection. Falk et al. believes that this triangulation of physiological data will be critcal in order for fNIRS to remain viable outside of the laboratory [48].

6.2.2 Reduced Calibration Time

Reducing the calibration time (or training time) will also be important for adoption of NIRS-based interfaces, and BCI in general. Currently, models based on fNIRS input are typically constructed with a limited number of labelled examples that are far from optimal. However, increasing the number of training trials necessarily increases the calibration period, reducing the usability of the device. For fNIRS to be adopted in everyday settings, it would ideal for it to be 'plug and play'. I outline three methods for future research which would likely decrease the calibration time and potentially increase the accuracy of the model.

- **Cross-subject models:** Rather than creating personalized models, it may be possible to construct a model of a particular user state that is generalizable across many people.

- **Cross-session models**: Typically, participants engage in a single experimental session in which they interact with a training period. However, whether this model can be used across multiple sessions that span hours, or days, or months is largely unknown.

- **Increased examples through data sampling:** Most of the studies in this thesis employ a similar calibration strategy - one timed trial of a particular task is used as a single labelled example in the model. However, it may be possible to use other data sampling techniques to increase the number of examples that can be extracted from each trial. While this would likely lead to more robust models, researchers must take care to maintain the integrity of the model.

6.2.3 Improved Calibration Tasks

While most work in BCI focuses on the signal analysis and modeling algorithms to classify user state, the selection of an appropriate calibration task is non-trivial, and equally critical. Machine learning algorithms typically work under the assumption that the labels of training examples are correct. For example, given the fNIRS data during a 3-back task and the associated label of 'high workload', it is assumed that the data is a good representation of high workload. However, this is may not always be the case. There are at least two potential areas where gains in classification accuracy may be made.

- Selecting calibration tasks that resemble real-world environments while maintaining the generalizability of the model.
- Use engagement checks by the system and discard training trials where the user is not engaged

In Chapter 4, we suggested that one possible reason for underwhelming preference classifications was that there was no reasonable method for checking the integrity of a training trial's label. The design of the calibration task, modeled after previous work on preference elicitation using fNIRS, used static images with no interaction. Thus, although the user was presented with stimuli that the system assumed *should* induce periods of high and low preference, there was no guarantee. The user may simply have been bored or not engaged with the task.

Many of the calibration tasks that exist for fNIRS (and BCIs in general) suffer the same fate. They are largely grounded in repetitive psychology tasks that can be divorced from real life scenarios (e.g. the n-back task or the TABLET task). As a result, participants lose interest, resulting in the construction of models built on incorrect information. While performance measures lend insight to a person's interaction, distinguishing between an engaged user and a unengaged user is not often trivial. In addition, these performance metrics are often nonexistent in the calibration periods of some user states, such as emotion.

One potential remedy for this problem is the investigation of calibration tasks that move closer to real-life tasks, while remaining generalizable enough to translate to multiple contexts. Girouard et al. [59], for example, used carefully selected videos as a replacement for rest-period stimuli (a static screen, often with a grey fixation cross) in order to create a non-activation task for the prefrontal cortext that remained engaging. Similarly, Flatla et al. [50] used motivating game elements during calibration task to maintain engagement.

Improving calibration methods will almost certainly improve the classification accuracy of models built on fNIRS data, however, researchers must remain cautious about comprimising the generalizability of the model.

6.2.4 Personalizing Parameters for Adaptive Mechanisms

Each of the adaptive mechanisms used (or proposed) in this thesis were designed to maximize the benefit of the user while minimizing the potential cost of misclassifications. However, the parameters of these adaptations may also significantly impact interaction. For example, in Afergan et al. [3], confidence thresholds were used to determine whether the user's workload should be increased or decreased. It is very likely that significantly altering those thresholds would also significantly alter a user's experience with the system. In this particular example, Afergan used extensive piloting to determine a set of thresholds that were suitable across most participants. This approach was successful, however, it may very well be the case that personalizing the threshold parameters to each individual would have yielded even stronger results.

This personalization could be done in one of two ways. First, users could manually modify their adaptation parameters during interaction with the system. While this has the benefit of potentially maximizing a user's perceived experience with the system, subtle adaptation mechanisms may make the task of optimizing these parameters difficult for a user.

It may also be possible to automatically personalize parameters. In the CARSON system, each participant's cost-of-interruption threshold was calculated

based on the average confidence value collected for relevant and irrelevant messages. Creating thresholds using a sampling of existing classifications enabled the system to personalize the adaptation parameters. It's also possible that this approach could aid the translation of models constructed during calibration periods to a real-time environment. However, it should be noted that in order for this optimization to occur, the system must receive a sampling of classifications from desired user state, potentially increasing the training period.

6.3 Closing Remarks

The use of brain and body metrics has become increasingly prevalent with the popular emergence of wearable computing. At the same time, the amount of information that is generated each year has increased at an exponential rate. While our computing devices have remarkable power and speed, they remain insensitive to our mental states, frequently breaching the norms that we maintain in social interactions with each other.

In this thesis, I explored the use of fNIRS as input to information delivery systems, and showed that physiological input has the potential to improve the way that people engage with information. As the number of sensors that our computers have access to increases, our computing devices may leverage information from brain and body sensors to respond in a manner that is more appropriate or more familar to us. The field of brain-computer interfaces and physiological computing is still young, but the potential is large. It is possible that the addition of physiological sensors will transform our computers from tools to collaborators.

Appendix A

fNIRS Plots During Interaction with Bar Graphs and Pie Charts

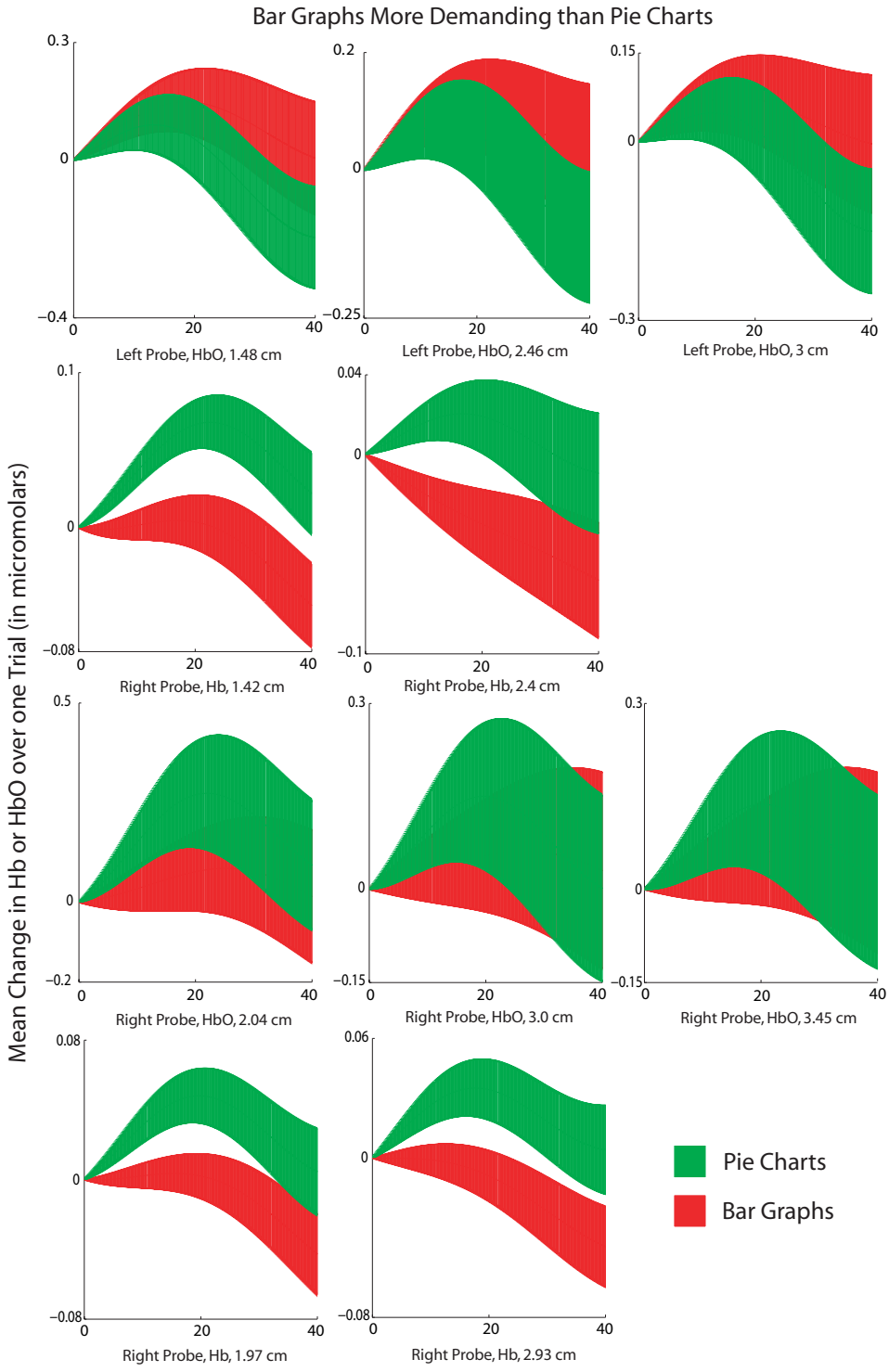


Figure A.1: Supplementing the deoxy-Hb plots in Chapter 3, the mean change in deoxy-Hb and oxy-Hb hemoglobin across all trials for participants who believed that bar graphs were more demanding than pie charts. This shows data from all source-detector pairs not presented in the chapter. At the time of the experiment, the 2 cm distance on the left probe and 2.5 cm distance on the right probe were not functioning correctly. Therefore, they are omitted.

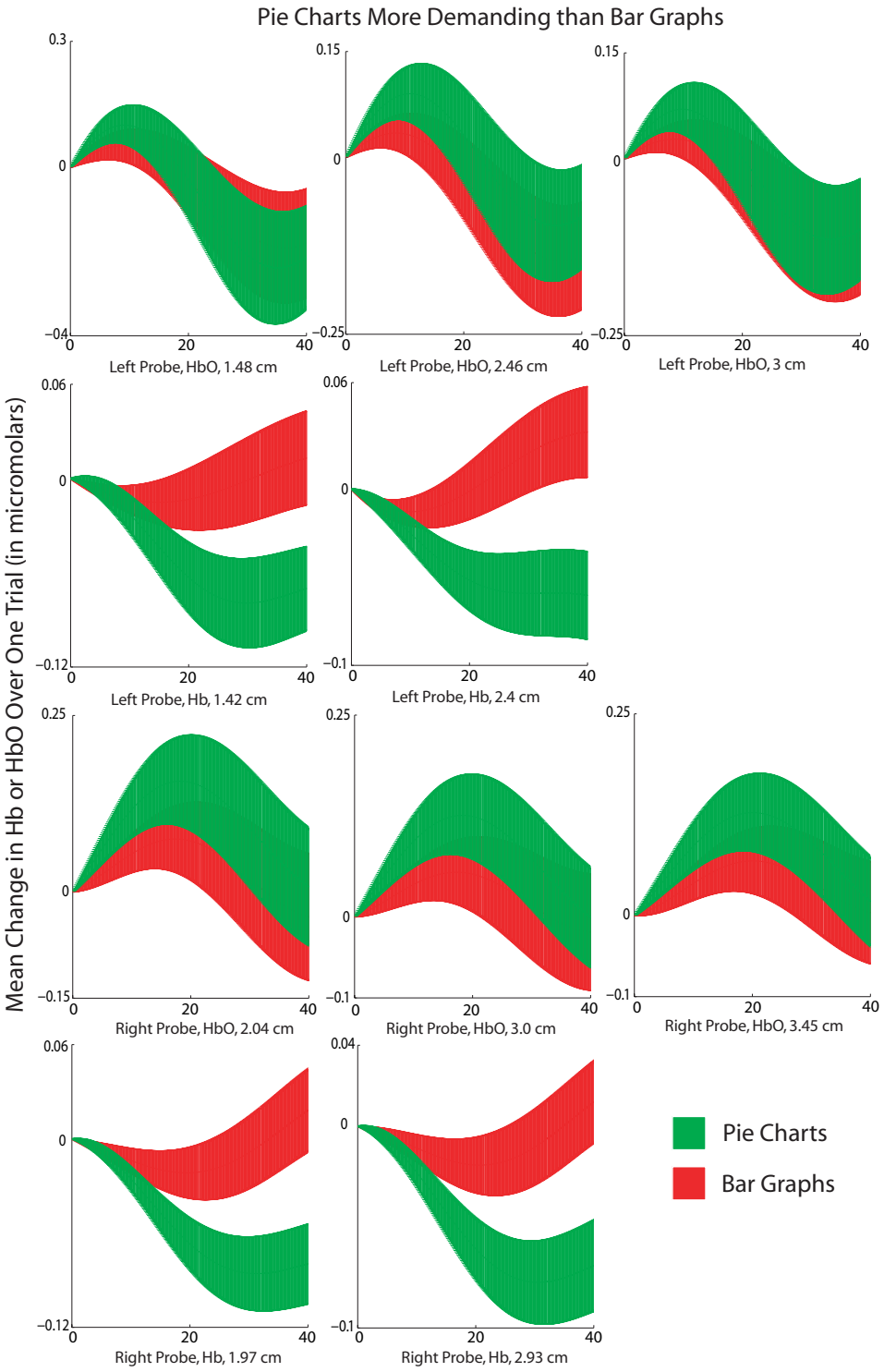


Figure A.2: Supplementing the deoxy-Hb plots in Chapter 3, the mean change in deoxy-Hb and oxy-Hb hemoglobin across all trials for participants who believed that pie charts were more demanding than bar graph. This shows data from all source-detector pairs not presented in the chapter. At the time of the experiment, the 2 cm distance on the left probe and 2.5 cm distance on the right probe were not functioning correctly. Therefore, they are omitted.

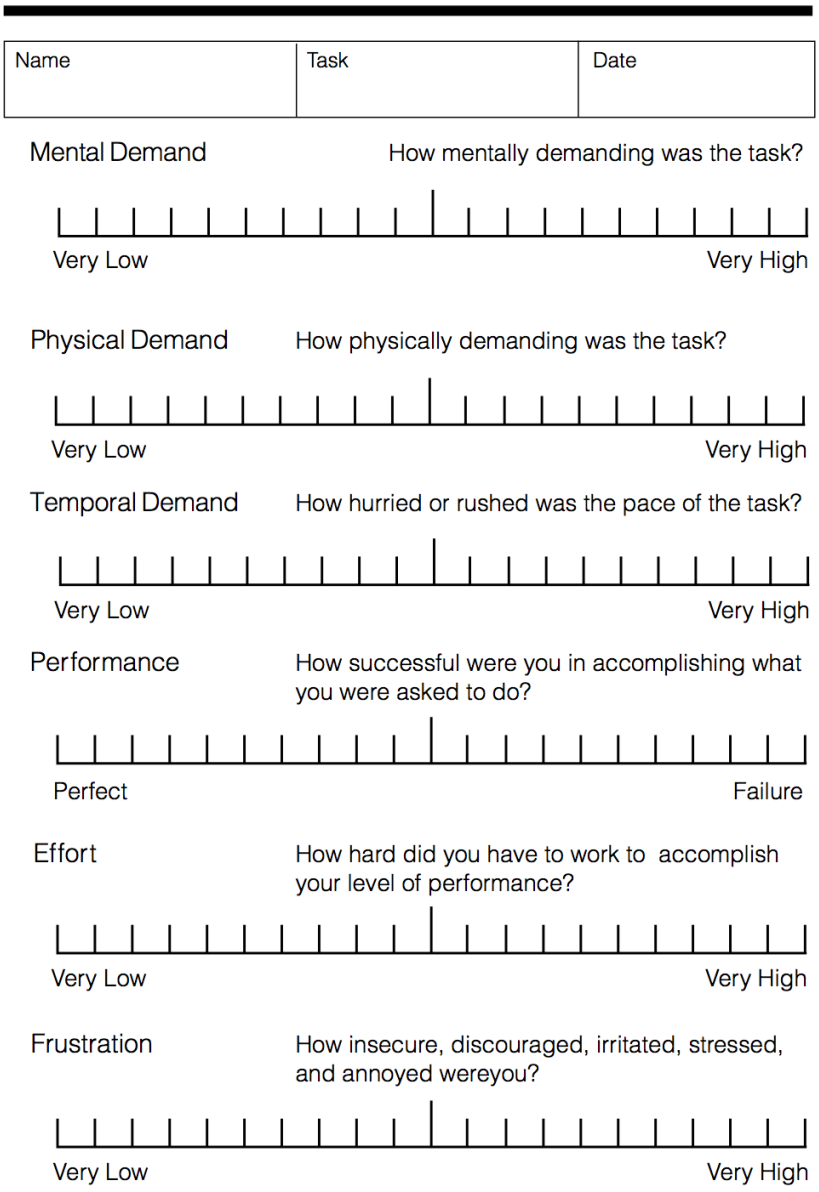
Appendix B

NASA-TLX

Figure 8.6

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.



Appendix C

CARSON System Survey

System Survey

Please answer each of the following questions about how

* Required

ID Number
Please ask the research if you are not sure

How useful did you find the system? *

	1	2	3	4	5	6	7	8	9	
Not at all	○	○	○	○	○	○	○	○	○	Very

How confusing did you find the system? *

	1	2	3	4	5	6	7	8	9	
Not at all	○	○	○	○	○	○	○	○	○	Very

How in control of the system did you feel?

	1	2	3	4	5	6	7	8	9	
Not at all	○	○	○	○	○	○	○	○	○	Very

How efficient did you feel the system was?

	1	2	3	4	5	6	7	8	9	
Not at all	○	○	○	○	○	○	○	○	○	Very

Any other comments? (Optional)

Bibliography

[1] Adamczyk, P. D., and Bailey, B. P. If Not Now, When?: The Effects of Interruption at Different Moments Within Task Execution. In *Proc. of ACM CHI 2004* (2004), pp. 271–278.

[2] Afergan, D., Peck, E. M., Chang, R., and Jacob, R. J. Using Passive Input to Adapt Visualization Systems to the Individual. In *ACM CHI Workshop: Many People, Many Eyes* (2013).

[3] Afergan, D., Peck, E. M., Solovey, E. T., Jenkins, A., Hincks, S. W., Brown, E. T., Chang, R., and Jacob, R. J. Dynamic difficulty using brain metrics of workload. In *Proc. of ACM CHI 2014* (New York, New York, USA, 2014), ACM Press, pp. 3797–3806.

[4] 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* (2011), pp. 791–800.

[5] Arias-Hernandez, R., Kaastra, L. T., Green, T., and Fisher, B. Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics. *Proc. of HICSS 2011* (2011), 1–10.

[6] Ashby, F. G., Valentin, V., and Turken, U. The effects of positive affect and arousal on working memory and executive attention. *Book: Advances in Consciousness Research* (2002), 245–288.

[7] 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 (Jan. 2012), 36–47.

[8] Baeza-Yates, R., Broder, A., and Maarek, Y. The new frontier of web search technology: seven challenges. *Search Computing*, 2 (2011), 3–9.

[9] Bailey, B., and Konstan, J. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior 22*, 4 (July 2006), 685–708.

[10] Bailey, B., Konstan, J., and Carlis, J. The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *INTERACT* (2001), vol. 1, pp. 593–601.

[11] Bailey, B. P., and Iqbal, S. T. Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction 14*, 4 (Jan. 2008), 1–28.

[12] Barnum, G. J., and Mattson, C. a. A Computationally Assisted Methodology for Preference-Guided Conceptual Design. *Journal of Mechanical Design 132*, 12 (2010), 121003.

[13] Bawden, D., and Robinson, L. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of Information Science 35*, 2 (Nov. 2008), 180–191.

[14] Bechara, A., Damasio, H., Damasio, A. R., and Lee, G. P. Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *The Journal of Neuroscience 19*, 13 (July 1999), 5473–81.

[15] Bell, C. J., Shenoy, P., Chalodhorn, R., and Rao, R. P. N. Control of a humanoid robot by a noninvasive brain-computer interface in humans. *Journal of Neural Engineering 5*, 2 (June 2008), 214–20.

[16] Bertini, E., Perer, A., Plaisant, C., and Santucci, G. Beyond time and errors: novel evaluation methods for Information Visualization. *A Workshop of CHI 2008*.

[17] Blood, A. J., and Zatorre, R. J. Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the United States of America 98*, 20 (Oct. 2001), 11818–23.

[18] Brouwer, A.-m., Erp, J. V., Heylen, D., Jensen, O., and Poel, M. Effortless Passive BCIs for Healthy Users Passive BCIs for Healthy Users. In *Universal Access in Human-Computer Interaction. Design Methods, Tools, and Interaction Techniques.* Springer Berlin Hedelberg, 2013, pp. 615–622.

[19] Byrne, E. a., and Parasuraman, R. Psychophysiology and adaptive automation. *Biological psychology 42*, 3 (Feb. 1996), 249–68.

[20] Cahill, L., Uncapher, M., Kilpatrick, L., Alkire, M. T., and Turner, J. Sex-related hemispheric lateralization of amygdala function in emotionally influenced memory: an FMRI investigation. *Learning & Memory 11*, 3 (2004), 261–266.

[21] Calcanis, C., Callaghan, V., Gardner, M., and Walker, M. Towards end-user physiological profiling for video recommendation engines. In *Proc. of Intelligent Environments* (2008), Iee, pp. 1–5.

[22] Cantador, I., Brusilovsky, P., and Kuflik, T. Second workshop on information heterogeneity and fusion in recommender systems (HetRec2011). In *Proc. of RecSys 2011* (2011), pp. 387–388.

[23] Carpendale, S. Evaluating Information Visualization. In *Information Visualization*. Springer Berlin Heidelberg, 2008, pp. 19–45.

[24] Carton, A. M., and Aiello, J. R. Control and Anticipation of Social Interruptions : Reduced Stress and Improved Task Performance. *Journal of Applied Social Psychology 39*, 1 (2009), 169–185.

[25] Causse, M., and Hurter, C. The physiological user's response as a clue to assess visual variables effectiveness. In *Human Centered Design*. Springer Berlin Heidelberg, 2009, pp. 167–176.

[26] 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 (May 1998), 411–23.

[27] Chen, C. Individual differences in a spatial-semantic virtual environment. *Journal of the American Society for Information Science 51*, 6 (2000), 529–542.

[28] Chen, C., and Czerwinski, M. Spatial ability and visual navigation: an empirical study. *New Review of Hypermedia and Multimedia 3*, 1 (1997), 67–89.

[29] Chen, D., Hart, J., and Vertegaal, R. Towards a Physiological Model of User Interruptability. In *Proc. of INTERACT 2007* (2007), pp. 439–451.

[30] Chen, D., and Vertegaal, R. Using mental load for managing interruptions in physiologically attentive user interfaces. In *Proc. of ACM CHI 2004 Extended Abstracts* (2004), ACM Press, pp. 1513–1516.

[31] Chesebro, J. L., and McCroskey, J. C. The relationship of teacher clarity and immediacy with student state receiver apprehension, affect, and cognitive learning. *Communication Education 50*, 1 (Jan. 2001), 59–68.

[32] Chul Kwon, B., Fisher, B., and Yi, J. Visual analytic roadblocks for novice investigators. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on* (2011), IEEE, pp. 3–11.

[33] Claypool, M., Brown, D., and Le, P. Inferring user interest. *IEEE Internet Computing 5*, 6 (2001), 32–39.

[34] 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 (Sept. 1984), 531–554.

[35] Conati, C., and Maclaren, H. Exploring the role of individual differences in information visualization. *In Proc. of Working Conference On Advanced Visual Interfaces* (2009), 199–206.

[36] Coyle, S., Ward, T., and Markham, C. Brain-computer interface using a simplified functional near-infrared spectroscopy system. *Journal of Neural Engineering 3* (2007), 219.

[37] Cui, X., Bray, S., and Reiss, A. L. Speeded Near Infrared Spectroscopy (NIRS) Response Detection. *PLoS ONE 5*, 11 (Nov. 2010), e15474.

[38] Cutrell, E., Czerwinski, M., and Horvitz, E. Effects of instant messaging interruptions on computing tasks. In *Proc. of ACM CHI 2000 Extended Abstracts* (2000), pp. 99–100.

[39] Cutrell, E., and Tan, D. BCI for passive input in HCI. In *Proc. of ACM CHI 2007* (2007), pp. 1–3.

[40] Davis, M. H., Meunier, F., and Marslen-Wilson, W. D. Neural responses to morphological, syntactic, and semantic properties of single words: an fMRI study. *Brain and language 89*, 3 (June 2004), 439–49.

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

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

[43] Deppe, M., Schwindt, W., Kugel, H., and H. Nonlinear responses within the medial prefrontal cortex reveal when specific implicit information influences economic decision making. *Journal of Neuroimaging 15* (2005), 171–182.

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

[45] Dou, W., Jeong, D. H., Stukes, F., Ribarsky, W., Lipford, H., and Chang, R. Recovering Reasoning Processes from User Interactions. *Computer Graphics and Applications, IEEE 29*, 3 (2009), 109–130.

[46] Fairclough, S. H. Psychophysiological Inference and Physiological Computer Games. In *BRAINPLAY 07 Brain-Computer Interfaces and Games Workshop at ACE* (2007), p. 19.

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

[48] Falk, T. H., Guirgis, M., Power, S., and Chau, T. T. Taking NIRS-BCIs outside the lab: towards achieving robustness against environment noise. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 19*, 2 (Apr. 2011), 136–46.

[49] Findlater, L., and McGrenere, J. A comparison of static, adaptive, and adaptable menus. *Proc. of ACM CHI 2004 6*, 1 (2004), 89–96.

[50] Flatla, D., Gutwin, C., Nacke, L., Bateman, S., and Mandryk, R. Calibration games: making calibration tasks enjoyable by adding motivating game elements. In *Proc. of ACM UIST 2011* (2011), pp. 403–412.

[51] Franceschini, M. A., Joseph, D. K., Huppert, T. J., Diamond, S. G., and Boas, D. a. Diffuse optical imaging of the whole head. *Journal of Biomedical Optics 11*, 5 (2006), 054007.

[52] Fredrickson, B. What good are positive emotions? *Review of General Psychology 2*, 3 (1998), 300.

[53] Fu, X., Budzik, J., and Hammond, K. Mining navigation history for recommendation. In *Proc. of Intelligent User Interfaces 2000* (2000), pp. 106–112.

[54] Gajos, K., and Weld, D. S. Preference elicitation for interface optimization. In *Proc. of ACM UIST 2005* (2005), pp. 173–182.

[55] Gajos, K. Z., Everitt, K., Tan, D. S., Czerwinski, M., and Weld, D. S. Predictability and accuracy in adaptive user interfaces. In *Proc. of ACM CHI 2008* (2008), p. 1271.

[56] Galán, F., Nuttin, M., Lew, E., Ferrez, P. W., Vanacker, G., Philips, J., and Millán, J. D. R. A brain-actuated wheelchair: asynchronous and non-invasive Brain-computer interfaces for continuous control of robots. *Clinical Neurophysiology 119*, 9 (Sept. 2008), 2159–2169.

[57] George, L., and Lécuyer, A. An overview of research on passive brain-computer interfaces for implicit human-computer interaction. In *Proc. of International Conference on Applied Bionics and Biomechanics: Workshop 'Brain-Computer Interfacing and Virtual Reality'* (2010).

[58] Gilleade, K., Dix, A., and Allanson, J. Affective videogames and modes of affective gaming: assist me, challenge me, emote me. In *DiGRA 2005* (2005), pp. 1–7.

[59] Girouard, A., Solovey, E., and Jacob, R. Designing a passive brain computer interface using real time classification of functional nearinfrared spectroscopy. *International Journal of Autonomous and Adaptive Communications Systems 6*, 1 (2013), 26–44.

[60] Girouard, A., Solovey, E. T., Hirshfield, L. M., Chauncey, K., Sassaroli, A., Fantini, S., and Jacob, R. J. K. Distinguishing difficulty levels with non-invasive brain activity measurements. In *INTERACT* (2009), pp. 440–452.

[61] 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. In *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction*. Springer, 2010, pp. 221–237.

[62] Gluck, J., Bunt, A., and Mcgrenere, J. Matching Attentional Draw with Utility in Interruption. In *Proc. of ACM CHI 2007* (2007), pp. 41–50.

[63] Gore, J. C. Principles and practice of functional MRI of the human brain. *The Journal of Clinical Investigation 112*, 1 (2003), 4–9.

[64] Green, T., and Fisher, B. The impact of personality factors on visual analytics interface interaction. *Proc. of IEEE Visual Analytics Science and Technology (VAST) 2010* (2010), 203–210.

[65] Gürkök, H., and Bos, D. Towards multiplayer BCI games. *BioSPlay: Workshop on Multiuser and Social Biosignal Adaptive Games and Playful Applications. Workshop at Fun and Games* (2010), 1–4.

[66] Hall, M., Frank, E., Holmes, G., and Pfahringer, B. The WEKA data mining software: an update. In *ACM SIGKDD* (2009), vol. 11, pp. 10–18.

[67] Harrivel, A., Hylton, A., and Hearn, T. Best Practices for the Application of Functional Near Infrared Spectroscopy to Operator State Sensing. Tech. Rep. July, NASA Glenn Research Center. Cleveland, Ohio, 2012.

[68] Hart, S. G., and Staveland, L. E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology 52* (1988), 139–183.

[69] Healey, J., Picard, R., and Dabek, F. A New Affect-Perceiving Interface and Its Application to Personalized Music Selection. In *Workshop on Perceptual User Interfaces* (1998), pp. 2–5.

[70] Heer, J., and Bostock, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proc. of ACM CHI 2010* (2010), pp. 203–212.

[71] Herff, C., Heger, D., Putze, F., Guan, C., and Schultz, T. Cross-subject classification of speaking modes using fNIRS. In *Neural Information Processing*. Springer Berlin Heidelberg, 2012, pp. 417–424.

[72] Heuer, R. *Psychology of intelligence analysis*. United States Govt Printing Office, 1999.

[73] 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 *Proc. of ACM CHI 2011* (2011), pp. 373–382.

[74] 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 *Proc. of ACM CHI 2009* (2009), pp. 2185–2194.

[75] Horvitz, E., and Apacible, J. Learning and reasoning about interruption. In *Proc. of Multimodal Interfaces 2003* (2003), pp. 20–27.

[76] Horvitz, E., Jacobs, A., and Hovel, D. Attention-Sensitive Alerting. In *Proc. of UAI 1999* (1999), vol. 98025, pp. 305–313.

[77] Horvitz, E., Kadie, C., Paek, T., and Hovel, D. Models of attention in computing and communication: from principles to applications. *Communications of the ACM 46*, 3 (2003), 52–59.

[78] Hoshi, Y. Near-Infrared Spectroscopy for Studying Higher Cognition. In *Neural Correlates of Thinking*. Springer Berlin Hedelberg, 2009, pp. 88–93.

[79] Hoshi, Y. Towards the next generation of near-infrared spectroscopy. *Philosophical transactions of the Royal Society A: Mathematical, Physical, and Engineering Sciences 369*, 1955 (Nov. 2011), 4425–4439.

[80] Hosseini, S. H., Mano, Y., Rostami, M., Takahashi, M., Sugiura, M., and Kawashima, R. Decoding what one likes or dislikes from single-trial fNIRS measurements. *Neuroreport 22*, 6 (2011), 269–273.

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

[82] Huber, J., Ariely, D., and Fischer, G. Expressing Preferences in a Principal-Agent Task: A Comparison of Choice, Rating and Matching. *Organizational Behavior and Human Decision Processes 87*, 1 (2002), 66–90.

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

[84] IMDB. http://www.imdb.com.

[85] Iqbal, S., and Adamczyk, P. Towards an index of opportunity: understanding changes in mental workload during task execution. In *Proc. of ACM CHI 2005* (2005), pp. 311–320.

[86] Iqbal, S., and Horvitz, E. Notifications and awareness: a field study of alert usage and preferences. In *Proc. of ACM CSCW 2010* (2010), pp. 1–4.

[87] Iqbal, S. T. MeWS-IT : A Mental Workload Based System for Interruption Timing. In *Proc. of ACM UIST 2005, Doctoral Symposium* (2005).

[88] Jacob, R. J. K., Girouard, A., Horn, M. S., and Solovey, E. T. Reality-Based Interaction: A Framework for Post-WIMP Interfaces. *Proc. of ACM CHI 2008* (2008), 201–210.

[89] Kapoor, A., and Shenoy, P. Combining brain computer interfaces with vision for object categorization. *Proc. of IEEE CVPR 2008* (June 2008), 1–8.

[90] Kellen, V. J. *The Effects of Diagrams and Relational Complexity on User Performance in Conditional Probability Problems in a Non-Learning Context.* College of computing and digital media dissertations, DePaul University, 2012.

[91] Klein, K., and Boals, A. Expressive writing can increase working memory capacity. *Journal of Experimental Psychology 130*, 3 (2001), 520.

[92] Koelstra, S., Yazdani, A., Soleymani, M., Muhl, C., Lee, J., Nijholt, A., Pun, T., Ebrahimi, T., and Patras, I. Single trial classification of EEG and peripheral physiological

signals for recognition of emotions induced by music videos. In *Brain Informatics*. Springer Berlin Heidelberg, 2010, pp. 89–100.

[93] Kohlmorgen, J., Dornhege, G., Braun, M., Blankertz, B., Muller, K.-R., Curio, G., Hagemann, K., Bruns, A., Schrauf, M., and Kincses, W. E. Improving human performance in a real operating environment through real-time mental workload detection. In *Toward Brain-Computer Interfacing*. MIT Press, 2007, pp. 409–422.

[94] Kreplin, U., and Fairclough, S. H. Activation of the rostromedial prefrontal cortex during the experience of positive emotion in the context of esthetic experience. An fNIRS study. *Frontiers in Human Neuroscience 7*, 879 (Jan. 2013), 1–9.

[95] Langley, P., and Fehling, M. The experimental study of adaptive user interfaces (No. 98-3). Tech. rep., 1996.

[96] Levenshtein, V. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. *Soviet Physics Doklady 10*, 8 (1966), 707–710.

[97] Liu, C., Agrawal, P., Sarkar, N., and Chen, S. Dynamic Difficulty Adjustment in Computer Games Through Real-Time Anxiety-Based Affective Feedback. *International Journal of Human-Computer Interaction 25*, 6 (Aug. 2009), 506–529.

[98] Luu, S., and Chau, T. Decoding subjective preference from single-trial near-infrared spectroscopy signals. *Journal of Neural Engineering 6* (2008).

[99] Makeig, S., Leslie, G., and Mullen, T. First demonstration of a musical emotion BCI. In *Affective Computing and Intelligent Interaction*. Springer Berlin Heidelberg, 2011, pp. 487–496.

[100] Mark, G., and Gudith, D. The cost of interrupted work: more speed and stress. In *Proc. of ACM CHI 2008* (2008), pp. 8–11.

[101] Marx, M., and Schmandt, C. CLUES: dynamic personalized message filtering. In *Proc. of CSCW 1996* (1996), ACM, pp. 113–121.

[102] McClure, S. M., Li, J., Tomlin, D., Cypert, K. S., Montague, L. M., and Montague, P. R. Neural correlates of behavioral preference for culturally familiar drinks. *Neuron 44*, 2 (Oct. 2004), 379–87.

[103] McFarlane, D., and Latorella, K. The scope and importance of human interruption in human-computer interaction design. *Human-Computer Interaction 17*, 1 (2002), 1–61.

[104] Millán, J., and Ferrez, P. Simultaneous real-time detection of motor imagery and error-related potentials for improved BCI accuracy. In *Proc. of 4th Brain-Computer Interface Workshop and Training Course* (2008), pp. 197–202.

[105] Miller, S., Kirlik, A., Kosorukoff, A., and Tsai, J. Supporting joint human-computer judgment under uncertainty. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 52*, 4 (2008), 408–412.

[106] Moghimi, S., Kushki, A., Power, S., Guerguerian, A. M., and Chau, T. Automatic detection of a prefrontal cortical response to emotionally rated music using multi-channel near-infrared spectroscopy. *Journal of Neural Engineering 9*, 2 (Apr. 2012), 026022.

[107] Molina, G. G., Nijholt, A., and Twente, U. Emotional Brain-Computer Interfaces. *International Journal of Autonomous and Adaptive Communications Systems 6*, 1 (2009), 9–25.

[108] Morita, M., and Shinoda, Y. Information filtering based on user behavior analysis and best match text retrieval. In *Proc. of SIGIR '94* (1994).

[109] 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. In *Aerospace and Electronics Conference* (1992), pp. 734–740.

[110] Nacke, L., Kalyn, M., Lough, C., and Mandryk, R. Biofeedback Game Design: Using Direct and Indirect Physiological Control to Enhance Game Interaction. In *Proc. of ACM CHI 2011* (2011), vol. 11, pp. 103–112.

[111] Nacke, L. E., and Mandryk, R. L. Designing Affective Games with Physiological Input. In *Workshop on Multiuser and Social Biosignal Adaptive Games and Playful Applications in Fun and Games Conference (BioS-Play)* (2011).

[112] Nass, C., and Moon, Y. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues 56*, 1 (Jan. 2000), 81–103.

[113] Nass, C., Steuer, J., and Tauber, E. R. Computers are social actors. In *Proc. of ACM CHI 2004* (1994), pp. 72–78.

[114] Nichols, D. Implicit rating and filtering. In *Proc. of DELOS Workshop on Filtering and Collaborative Filtering* (1997).

[115] Nomura, T., and Saeki, K. Effects of Polite Behaviors Expressed by Robots: A Case Study in Japan. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology* (2009), Ieee, pp. 108–114.

[116] Ottley, A., Peck, E. M., Harrison, L., and Chang, R. The Adaptive User : Priming to Improve Interaction. In *ACM CHI Workshop: Many People, Many Eyes* (2013).

[117] 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 (May 2005), 46–59.

[118] 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.

[119] 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*, 4 (1993), 737–743.

[120] Parasuraman, R., and Miller, C. A. Trust and Etiquette in High-Criticality Automated Systems. *Communications of the ACM 47*, 4 (2004), 51–55.

[121] Parnin, C., and Rugaber, S. Resumption strategies for interrupted programming tasks. *Software Quality Journal 19*, 1 (Aug. 2010), 5–34.

[122] Parsons, J., Ralph, P., and Gallagher, K. Using viewing time to infer user preference in recommender systems. In *Proc. of AAAI Workshop on Semantic Web Personalization* (2004).

[123] Paulus, M. P., and Frank, L. R. Ventromedial prefrontal cortex activation is critical for preference judgments. *Neuroreport 14*, 10 (July 2003), 1311–1315.

[124] Paymans, T. F., Lindenberg, J., and Neerincx, M. Usability trade-offs for adaptive user interfaces : ease of use and learnability. In *Proc. of IUI 2004* (2004), no. c, pp. 301–303.

[125] Peck, E., Yuksel, B., Ottley, A., Jacob, R. J., and Chang, R. Using fNIRS Brain Sensing to Evaluate Information Visualization Interfaces. In *Proc. of ACM CHI 2013* (2013), pp. 473–482.

[126] Peck, E. M., Afergan, D., and Jacob, R. J. K. Investigation of fNIRS brain sensing as input to information filtering systems. *Proc. of Augmented Human 2013* (2013), 142–149.

[127] Peck, E. M., Afergan, D., Yuksel, B. F., Lalooses, F., and Jacob, R. J. K. Using fNIRS to Measure Mental Workload in the Real World. In *Advances in Physiological Computing*, S. H. Fairclough and K. Gilleade, Eds., HumanComputer Interaction Series. Springer London, London, 2014, pp. 117–139.

[128] Peck, E. M., Lalooses, F., and Chauncey, K. Framing Meaningful Adaptation in a Social Context. In *ACM CHI Workshop: Brain and Body Interfaces* (2011), pp. 1–4.

[129] Peck, E. M., Solovey, E. T., Sassaroli, A., Fantini, S., Jacob, R. J. K., Hirshfield, L. M., and College, H. Your Brain, Your Computer, and You. *IEEE Computer* (2010), 86–90.

[130] Peck, E. M., Solovey, E. T., Su, S., Jacob, R. J. K., and Chang, R. Near to the Brain : Functional Near-Infrared Spectroscopy as a Lightweight Brain Imaging Technique for Visualization. In *Proc. of IEEE Infovis 2011 (Poster)* (2011).

[131] Peck, E. M., Yuksel, B. F., Harrison, L., Ottley, A., and Chang, R. Position Paper : Towards a 3-Dimensional Model of Individual Cognitive Differences. In *Proc. of BELIV 2012* (2012).

[132] Phelps, E., and Carrasco, M. Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological Science 17*, 4 (2006), 292–299.

[133] Pope, A. T., Bogart, E. H., and Bartolome, D. S. Biocybernetic system evaluates indices of operator engagement in automated task. *Biological Psychology 40*, 1-2 (May 1995), 187–95.

[134] Presacco, A., Goodman, R., Forrester, L., and Contreras-Vidal, J. L. Neural decoding of treadmill walking from noninvasive electroencephalographic signals. *Journal of Neurophysiology 106*, 4 (Oct. 2011), 1875–1887.

[135] Ramnani, N., and Owen, A. M. Anterior Prefrontal Cortex: Insights into Function from Anatomy and Neuroimaging. *Nature Reviews. Neuroscience 5*, 3 (Mar. 2004), 184–194.

[136] Riche, N. Beyond system logging: human logging for evaluating information visualization. In *Proc. of BELIV 2010 Workshop* (2010).

[137] Rotten Tomatoes. http://www.rottentomatoes.com.

[138] Rowe, G., Hirsh, J., and Anderson, A. Positive affect increases the breadth of attentional selection. *Proceedings of the National Academy of Sciences 104*, 1 (2007), 383–388.

[139] 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 (June 2000), 1656–1660.

[140] Salvucci, D., and Bogunovich, P. Multitasking and monotasking: the effects of mental workload on deferred task interruptions. *Proc. of CHI 2010* (2010), 85–88.

[141] Sase, I., Takatsuki, A., Seki, J., Yanagida, T., and Seiyama, A. Noncontact backscatter-mode near-infrared time-resolved imaging system: Preliminary study for functional brain mapping. *Journal of Biomedical Optics 11*, 5 (2012), 054006.

[142] Sassaroli, A., Zheng, F., Coutts, M., Hirshfield, L. H., Girouard, A., Solovey, E. T., Jacob, R. J. K., Tong, Y., DeB Frederick, B., and Fantini, S. Application of near-infrared spectroscopy for discrimination of mental workloads. *Proc. of SPIE BiOS: Biomedical Optics* (2009), 71741H–71741H.

[143] Sassaroli, A., Zheng, F., Hirshfield, L. M., Girouard, A., Solovey, E. T., Jacob, R. J. K., and Fantini, S. Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy. *Journal of Innovative Optical Health Sciences 1*, 2 (2008), 227–237.

[144] Seno, B. D., Matteuccia, M., and Mainardib, L. Online detection of P300 and Error Potentials in a BCI speller. *Computational Intelligence and Neuroscience 2010*, 11 (2010).

[145] Shackman, A., Sarinopoulos, I., Maxwell, J., Pizzagalli, D., Lavric, A., and Davidson, R. Anxiety selectively disrupts visuospatial working memory. *Emotion 6*, 1 (2006), 40.

[146] Shneiderman, B. Direct Manipulation for Comprehensible , User Interfaces Predictable and Controllable. In *Proc. of IUI 1997* (1997), pp. 33–39.

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

[148] Simon, H. A. Designing Organizations for an Information-Rich World. *Computers, Communications, and the Public Interest* (1971), 37–72.

[149] Smith, A. The Best (and Worst) of Mobile Connectivity. Tech. rep., Pew Research Center, Washington, DC, 2012.

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

[151] Solovey, E. T. *Real-Time fNIRS Brain Input For Enhancing Interactive Systems*. PhD thesis, 2012.

[152] 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 *Proc. of UIST 2009* (2009), pp. 157–166.

[153] 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 *Proc. of ACM CHI 2011* (2011), pp. 383–392.

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

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

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

[157] 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 (Oct. 2002), 719–731.

[158] Streefkerk, J. W., McCrickard, D. S., van Esch-Bussemakers, M. P., and Neerincx, M. a. Balancing Awareness and Interruption in Mobile Patrol using Context-Aware

Notification. *International Journal of Mobile Human Computer Interaction 4*, 3 (Jan. 2012), 1–27.

[159] Szafir, D., and Mutlu, B. Pay Attention! Designing Adaptive Agents that Monitor and Improve User Engagement. *Proc. of ACM CHI 2012* (2012), 11–20.

[160] Teevan, J., Dumais, S., and Horvitz, E. Characterizing the value of personalizing search. In *Proc. of ACM Research and Development in Information Retrieval 2007* (2007), pp. 757–758.

[161] Thomas, J., and Cook, K. *Illuminating the path: The research and development agenda for visual analytics*, vol. 54. IEEE, 2005.

[162] Toker, D., Conati, C., Carenini, G., and Haraty, M. Towards adaptive information visualization: on the influence of user characteristics. *User Modeling, Adaptation, and Personalization* (2012), 274–285.

[163] Van Velsen, L., Van Der Geest, T., Klaassen, R., and Steehouder, M. User-centered evaluation of adaptive and adaptable systems: a literature review. *The Knowledge Engineering Review 23*, 03 (Sept. 2008), 261–281.

[164] Velez, M., Silver, D., and Tremaine, M. Understanding visualization through spatial ability differences. In *Proc. of IEEE Visualization 2005* (2005), IEEE, pp. 511–518.

[165] Vi, C., and Subramanian, S. Detecting error-related negativity for interaction design. In *Proc. of ACM CHI 2012* (2012), ACM Press, pp. 493–502.

[166] Vidaurre, C., Sannelli, C., Muller, K.-R., and Blankertz, B. Machine-Learning-Based Coadaptive Calibration for Brain-Computer Interfaces. *Neural computation 816* (Dec. 2010), 791–816.

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

[168] Wainer, J., Dabbish, L., and Kraut, R. Should I open this email?: Inbox-level cues, curiosity and attention to email. In *Proc. of ACM CHI 2011* (2011), pp. 3439–3448.

[169] Whitworth, B., and Ahmad, A. Polite computing. *Behaviour & Information Technology 24*, 5 (Sept. 2005), 353–363.

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

[171] Wigdor, D., Shen, C., Forlines, C., and Balakrishnan, R. Perception of elementary graphical elements in tabletop and multi-surface environments. In *Proc. of ACM CHI 2007* (New York, New York, USA, 2007), ACM Press, pp. 473–482.

[172] Wigdor, D., and Wixon, D. *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*. Elsevier, 2011.

[173] Wildey, C., MacFarlane, D., Khan, B., Tian, F., Liu, H., and Alexandrakis, G. Improved fNIRS Using a Novel Brush Optrode. *Frontiers in Optics 2010/Laser Science JTuA23* (2010).

[174] 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 (2007), 1–13.

[175] Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., and Vaughan, T. M. Brain-computer interfaces for communication and control. *Clinical Neurophysiology 113*, 6 (2002), 767–791.

[176] Wu, P., and Miller, C. Can polite computers produce better human performance? In *Proc. of AFFINE '10* (2010), pp. 87–92.

[177] Wu, P., Ott, T., and Miller, C. Evaluating the Effects of Culture and Etiquette on Human-Computer Interaction and Human Performance. In *Proc. of the AAAI Spring Symposium* (2009), pp. 23–25.

[178] Wu, T.-L., Hung, Y.-P., Chuang, Y.-Y., Chen, H.-H., Chen, H. H., Chen, J.-H., Jeng, S.-K., Wang, H.-K., Ho, C.-C., Lin, Y.-P., Hu, T.-T., Weng, M.-F., Chan, L.-W., Yang, C.-H., and Yang, Y.-H. Interactive content presentation based on expressed emotion and physiological feedback. In *ACM Multimedia* (2008), ACM Press, pp. 1009–1010.

[179] Yeh, Y.-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.

[180] Yi, J. S. Implications of individual differences on evaluating information visualization techniques. In *Proc. of ACM CHI 2010 Conference on Human Factors in Computing Systems: Workshop on BELIV* (2010).

[181] Yuksel, B. F., Donnerer, M., Tompkin, J., and Steed, A. A novel brain-computer interface using a multi-touch surface. In *Proc. of ACM CHI 2010* (2010), pp. 855–858.

[182] Zander, T., and Jatzev, S. Detecting affective covert user states with passive brain-computer interfaces. In *Affective Computing and Intelligent Interaction and Workshops* (2009), IEEE, pp. 1–9.

[183] Zander, T., Kothe, C., Jatzev, S., and Gaertner, M. Enhancing human-computer interaction with input from active and passive brain-computer interfaces. *Brain-Computer Interfaces* (2010), 181–199.

[184] Ziemkiewicz, C., Crouser, R., Yauilla, A., Su, S., Ribarsky, W., and Chang, R. How locus of control influences compatibility with visualization style. *IEEE Visual Analytics Science and Technology* (2011), 81–90.

[185] Ziemkiewicz, C., and Kosara, R. Laws of Attraction: From Perceptual Forces to Conceptual Similarity. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 1009–1016.

[186] Zuk, T., and Carpendale, S. Visualization of uncertainty and reasoning. In *Smart Graphics*. 2007, pp. 164–177.