Using fNIRS to Detect Mental Workload and Emotional Valence in web form filling task

Kristiyan Lukanov

2015-09-18

Acknowledgements

lorem ipsum

Abstract

In this dissertation we evaluated a desktop interface using functional near infrared spectroscopy(fNIRS), and verified the practicality of this brain imaging modality. More specifically, we tested the web page layout of online insurance claim process...

Contents

Chapter 1

Introduction

Users often has to fill web pages containing more than 10 forms for example, when registering for a web site, posting classified ad, or sending online insurance claim. Sometimes this is really important in Human computer interaction(HCI) viewpoint, like filling insurance claim forms, and online banking to be intuitive and aiding the user through the process. To achieve that web forms should support the users working memory[68, 89] by minimizing the effort to perceive, process and respond to the web form. That is why we are interested in measuring the mental demands imposed by the web form filling task. Furthermore, it has been suggested that attractive interfaces increase creativity[69] of the user. Hence, it can be of high value for the researchers to know what workload and emotional state the users are experiencing during interaction with a certain interface.

Recently, functional near infrared spectroscopy(fNIRS) has been suggested as a suitable brain imaging method for HCI studies[60, 92, 76] because participants can wear it during normal interaction with a computer interface. In addition, the brain scanning device is non-intrusive and relatively resistant to motion artefacts which will not affect task performance and data collected, in contrast to other brain imaging modalities. Moreover, as it has been suggested by cognitive neuroscience studies that the prefrontal cortex(PFC) area of the brain is involved with higher order cognition[14] and emotion processing[28]. Thus, by placing the fNIRS device on the forehead of individuals we can infer about their level of demand and emotional state.

However, according to our knowledge only one study was found[74] that uses hemodynamic data from fNIRS to compare and evaluate different variations of an interface. Other fNIRS studies experiment with simple tasks, like mental arithmetic, and n-back tasks. Accordingly, we want to implement the fNIRS device in a user trial evaluation study of an web interface because it is often encountered task in our daily lives.

1.1 Purpose of study

We aim to find a way to improve web interfaces that has more than 10 forms, and are considered long forms, as this process is often encountered during daily web surfing, for example, when user registers to a new web site, or enter information

for financial institutions, like insurance companies and banks. We strive to find more generalizable results that can produce certain web form design guidelines for interacting with long forms. Accordingly, we decided to test the layout of the web forms, and examine how it influences user performance. We also, aim to assess the practicality of fNIRS brain imaging technique in HCI evaluation studies.

1.2 Research questions

In this master thesis we aim to answer the following questions:

1. Which of the three layouts elicit the least mental workload and which is more preferred by the users?
2. Is fNIRS sensitive method in measuring mental workload changes in web form filling task?
3. Can we detect emotional valence with fNIRS, from web interface that has no emotional cues.
4. Is fNIRS brain imaging modality practical to use in HCI evaluation studies?

1.3 Industry partner

This work has been motivated by the need of entity partner funding my masters course. The industry partner operates an insurance customer relationship management(CRM) software, and it was requested to provide insights in the web form filling process and provide design guidelines.

1.4 Structure of the thesis

In the next chapter we will first review the background literature behind usability and web form filling, the concept of mental workload and working memory, emotion processing, and finally, relevant brain sensing techniques. In chapter 3 we will describe the User study, including description of the method we used, and the results obtained. Finally, we will discuss the finding from the experiment and then propose implications for design.

Chapter 2

Literature review

In the HCI field evaluation approaches ca be divided in three general categories: analytical, field study and lab study[80]. First, analytical methods are designed to predict user behaviour such as, heuristic evaluation or expert reviews, so no experiment has to be conducted. Second, field studies are conducted in context in order to collect relevant and valid data, like observations. Lastly, lab studies use artificial settings but the experiment variables can be controlled easier, and also, comparative tests can be conducted. Our aim of the study is to evaluate a web form filling interface for the insurance domain, and more precisely, online auto insurance claim process. Therefore we are interested in conducting lab study because we cannot simulate road accident and we are able to compare variations of insurance claim web form. Furthermore, there are variety of evaluation methods, like interviews which will give us information about what users think about the interface, or observations which will let us recognize typical behaviour of users and obstacles they encounter while using the interface.

It has to be mentioned that lab studies give us the chance to prepare the environment and record more performance measures which provide valuable objective information. Typical performance measures in usability experiments are time to complete, errors encountered, and number of events. In general, mental workload and emotional valence are of high interest in HCI evaluations because measuring workload gives us important information about the task demands and also, knowing whether a user is feeling positive or negative towards an interface or task may give us valuable information about their general preferences. Furthermore, Nielsen[67] suggested that user preferences correlate with user performance, thus we can rely solely on user preferences, however this is not always the case. Hence, Nielsen and Levy [67] advise researchers to use combination of subjective and objective data in usability studies, in order to identify bias and provide richer information about the process. Accordingly, we have decided to employ user trials(subjective data) combined with psychophysiological measurements(objective data). In addition, functional near infrared spectroscopy (fNIRS), has been recently suggested as a promising method for HCI evaluations[60, 76] because it was suggested to measure mental workload[61]. Based on this, we decided to test the usefulness of fNIRS in HCI evaluation studies.

Based on the information above this master thesis considers measuring mental workload and emotional valence to inform the user interface design of the insurance claim web form. The literature review will proceed in the following way: first, we will review relevant literature on web form filling and usability. Second, we discuss working memory models and the mental workload concept. Third, emotional processing literature will be revised. Fourth, current brain sensing techniques used in cognitive experiments will be reviewed and, fifht, fNIRS studies examining mental workload and emotional valence will be examined. Finally, we will summarize the reviewed literature.

2.1 Usability and Web form filling

Web form filling is often encountered activity in daily surfing of web users, however, according to our knowledge, there is scarce of empirical research in the Human Computer Interaction(HCI) literature for this topic. First, a study by Wstlund[103] compared two web page layouts - one that all the text is in the same page, and one where the text is separated in four pages. Authors concluded that users experienced less workload with the divided web form(4 pages), compared to the single page web form. Second, two books specially written for web form filling design[49, 102] suggest splitting long web forms into several pages, in order to improve the process. Lastly, most of the research on web form filling and design is focused in optimizing the experience and accessibility for elderly population[86, 22, 57, 85].

In reviewing positive and negative affect and their implications for design Norman[69] proposes that "Positively valenced affect broadens the thought processes hence, enhanced creativity". Therefore, increase in task performance should be observed when a user has positively valenced affect towards certain interface. Suggesting that emotional valence plays pivotal role in task performance. Also, A couple of studies suggest that the longer it takes for a task(short or long term) to be completed the more the perceived frustration the users experience increases[62, 9]. Therefore, we should aim to minimize frustration by reducing the time to complete a task.

Because of insufficiency of relevant literature on the web form filling and design we are going to examine general usability guidelines and recommendations as they are widely accepted by the Human Computer Interaction researchers. The two most popular usability heuristics are those of Nielsen[68], and Shneidermann[89]. They express similar suggestions, like, maintain consistency, provide feedback, support expert users, prevent and optimize error messages, provide help documentation, permit easy reversal of information, and minimize working memory load. Consequently, the design of the tested variants in the usability study in chapter 3 is informed by them. Also, because both heuristics advocate minimizing the load on working memory we consider that reducing it will provide better user experience. In addition, because usability of certain interface depends on the context, user differences, and that there is not perfect solution to a interface problem, and designers often have to make tradeoffs[71], we will rely on cognitive science in order, to predict which layout is more appropriate.

2.2 Working Memory and Mental Workload

The concept of mental workload (MW) is intuitive in nature and it represents how busy an operator is when performing a certain task. The concept has been referred in the literature with many terms, like cognitive load, stress, strain, and arousal. Many definitions has been proposed by many authors, however researchers are still unable to find a consensus on the term[Linton et al 1989]. Wickens[98] defines it as "The demand imposed by tasks on the human's limited resources, whether considered single or multiple". Depending on the studied task at hand, knowing workload experienced by different design variations will help choose the one that generates desired operator performance. Also, in terms of operator experience of MW, Rouse et al classifies different factors like, fatigue, mood, individual differences, as person-specific workload[83]. Similarly, Norman and Bobrow classified operator performance on data-limited and resource-limited[70]. They hypothesize that even if operator spends high amount of attentional resources, the task can have a bad representation that will degrade the performance. In contrast, resource-limited performance depends on how much attentional resources the task demands, and it can be considered that every real life task consists of combination of both.

2.2.1 Working memory models

Rather than searching for definition researchers in cognitive science use models of working memory in order to understand cognition, predict and explain workload and performance. Furthermore, theories of working memory try to define the processes going into human mind, and explain concepts such as, attention, perception, long term memory, decision making, action selection, and execution[99, 4, 63]. Most of those models are based on human as information processor approach [15, 16, 66, 99], which relates the processes of human mind with those of a computer processor. Also, the framework is based on the assumption that the human operator has a limited resource capacity[50, 45] and if the task demands more resources than the capacity of the operator, workload overload is observed. Moreover, the information from the environment or the task is processed by series of processing systems, like perception, attention, short-term memory, long-term memory.

In attempt to describe the web form filling task we can use the working memory model from Baddeley and Hitch [4] which processes information in verbal and spatial form . It consists of a central executive, which is acts as an administration system which controls the information input and output of its slave systems. The visuo-spatial sketch pad is involved in holding visual information in spatial form like, objects and colours. The phonological loop stores verbal information, such as words and names. And the later proposed [3] episodic buffer is responsible for the storage and retrieval of memories or events. Because the task of web form filling involves multiple cognitive processes like, visual search, speech synthesis, planning, memory retrieval, decision making, thus utilizing all slave systems of the model, we can label the web form filling process as one that involves complex cognition. We can also consider the multiple resources model by Wickens[98, 97] which is suited for predicting the workload of an operator performing multiple tasks at one time. The 4-D multiple resources model is visualised as 4 dimensional cube as illustrated in Figure 2.2.The approach is

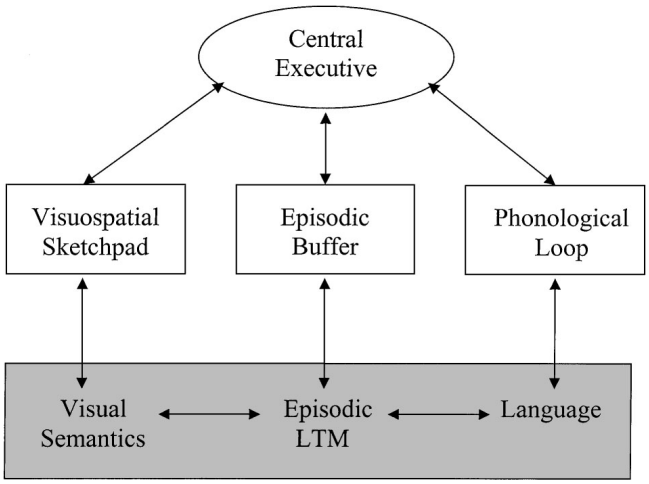

Figure 2.1: Working memory model by Baddeley and Hitch, displaying the 'slave systems' visuo-spatial sketch pad, episodic buffer and phonological loop, controlled by the central executive.

based on four basic assumptions:
1) in the stages of processing dimension, perceptual and cognitive tasks use different resources than response selection and execution;
2) spatial activity uses different resources than verbal or linguistic activity;
3) the modalities dimension, different resources are used for auditory and visual perception
4) visual channels are divided on focal and ambient vision
And the main argument of the theory is "to the extent that two tasks use different levels along each of the three dimensions, time-sharing will be better" [98]. The model provides an account on how different elements of the human information processor, like attention, perception, working memory, response selection and execution interact between each other. This theory is also based on research from cognitive neuroscience that suggests different modalities have different locations in the human brain, like the auditory cortex is involved with auditory perception and the visual perception is processed mainly by the visual cortex(occupational lobe).

However, mental workload can be influenced by the initial perception of the task at hand or the 'appraisal' of it. Similarly to MW appraisal is complex and multidimensional concept[37, 73] that is not well defined.

2.2.2 Measuring Mental Workload

The concept has been explained differently by different authors, and inferences made from various empirical measures which can be divided on primary, secondary, subjective and psychophysiological measures. There are also analytical techniques but we are not concerned with them in this dissertation.

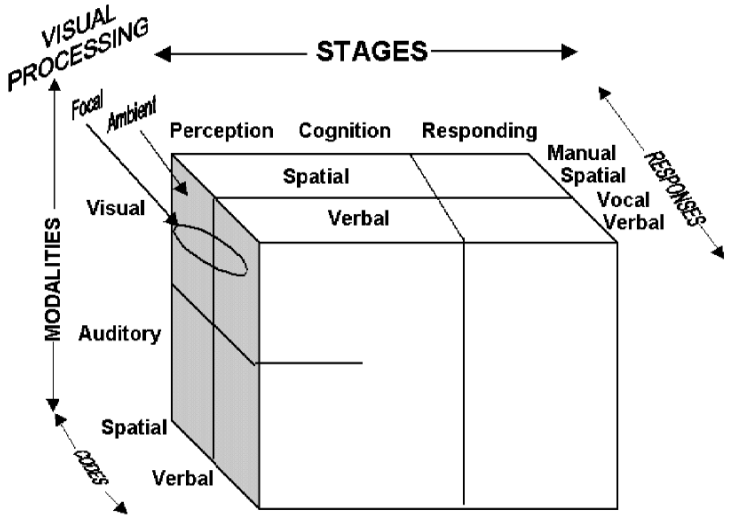

Figure 2.2: Wickens 4-D multiple resources model consisting of two codes(spatial and verbal), 2 modalities(auditory and visual), 3 stages(perception, cognition and responding), and 2 types of responses(manual and vocal).

Primary and secondary task measures

Primary measures rely on operator performance to predict workload. However, a limitation of using primary measures alone is that an operator can spend high amount of effort but this may not be apparent from the performance[101]. Consequently, primary task measures should be combined with other workload measures. An example performance measures are task completion time, number of errors, and response time. In our case we will use a combination of all but secondary task measures. Secondary task involves inclusion of a additional simple task to the primary one, which is done concurrently, if the primary task has low or moderate demand and the level of workload cannot be inferred only from the primary measure. It is used to detect when operators performance deteriorates and this is due to workload overload. However, as we are not going to use secondary measures, more explanation will provided for the other types of measurements.

Subjective measures

Subjective measures use rating scale and are based on operator opinion of their perceived workload during or after completion of task. They are preferred method for WL estimation because they are easy to administer, cheap, and with high face validity. They are classified as uni dimensional and multidimensional. They consist of single subjective scale of workload and multiple scales of types of workload, accordingly. From the unidimensional subjective measures the Cooper-Harper[23] is the most popular among ergonomists and cognitive researchers. However, it is designed for the aircraft domain, and therefore a modified cooper-Harper scale[100] was created for use in other domains. However, as mental workload is influenced by different environmental

and personal factors[83], therefore the concept of MW should be considered as multidimensional concept, in order to improve diagnosticity. Accordingly, multidimensional subjective scales should be used to better understand the aspects of MW. The most used scales are NASA-TLX[43], SWAT[79] and Workload profile[94]. NASA-TLX is based on rigorous laboratory research, and it includes 6 scales (mental, physical, temporal demand, experienced effort, frustration and performance) consisting of 20 intervals each and it is relatively easy to administer. In addition, a single measure of workload can be weighted, although it is not necessarily required because there is high correlation between weighted and unweighed results[19]. The SWAT subjective scale is also widely used, however the process of implementation is laborious and more complex than the other subjective scales. The other popular scale is Workload profile(WP)[94] scale which is basen on the Wickens multiple resources model, and asks questions about each of the four dimensions proposed by the theory. Hence, it is very useful when combined with multiple resource theory interpretation of the results. Finally, Longo et al. [58] compared the three measures mentioned above in a web browsing/searching task, and observed correlations in the results of the three measures claiming that they measure the same concept of mental workload.

Psychophysical measures

Psychophysical measures are used to give objective data about mental workload by not relying on subjective scales or performance measures . They can be obtained by recording cardiac activity, electrodermal activity, eye function or imaging the brain. These techniques detect the change in the arousal from the autonomic nervous system level which can be inferred to as mental workload. However, different psychophysical measures capture different aspects mental workload[20], therefore consideration should be put in choosing the most appropriate measure for the given task.
Mean heart rate(HR) and heart rate variability(HRV) are one of the most used techniques to infer arousal because it is relatively cheap and easy to administer. However, HR not always correlates to subjective measures of MW[39] and because of this HRV can be considered as more valid measure. Moreover, the beat to beat interval of the heart can be measured using different statistical approaches[10], like, standard deviation from heart beat intervals. Measurements of eye activity, like blink rate, pupil diameter are also being found to correlate to MW. Furthermore, increase in pupil diameter is correlated to rise in arousal [50], and Beatty claimed that it has high sensitivity [8] and it can be used to distinguish between data-limited and resource limited processing, which can make it very useful in the HCI field. However, incoming light at the eye can change the pupil diameter, which is a process unrelated to the task, thus influencing the measurements, therefore it is suitable for experiments in controlled environment.
Finally, a number of brain imaging techniques are used to obtain measurements from the brain activity, including electroencephalography(EEG), functional near-infrared spectroscopy(fNIRS), functional magnetic resonance imaging(fMRI), however these will be discussed later in section "Brain sensing".

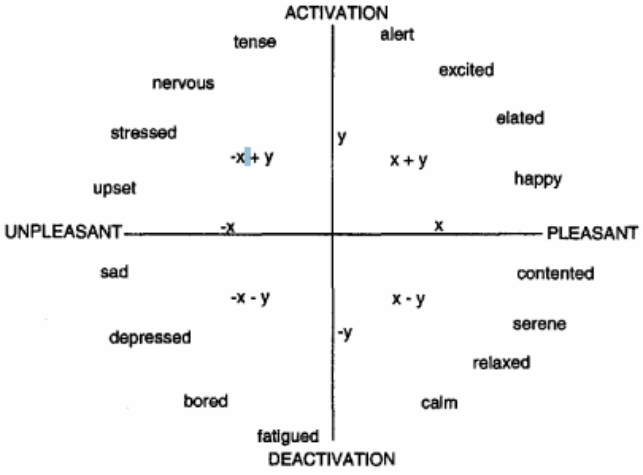

Figure 2.3: The two dimensional approach of emotion. The x and y axes represent semantic components: x=pleasant and unpleasant emotions, and y=level of activation. The image is taken from Russell and Feldman "Independence and bipolarity in the structure of current affect"[36]

2.3 Emotion processing

To begin with, mood and emotion are frequently referred as separate concepts. For example, emotions are thought to last for shorter time like, several minutes, in comparison to moods which can last for a whole day or more. Also, emotions are generally caused by certain events, like winning a game, in contrast to moods where often there is no reason as to why an individual is in certain mood. However, there is no clear distinction between them because moods can cause certain emotions and emotions can cause moods. To solve that problem researchers often use the term "affect" to encompass both emotions and moods. Furthermore, we can define positive moods or emotions as positive affect, and negative emotion and moods as negative affect.

There are two major approaches in describing emotions - the categorical[47] and the dimensional approach[36]. The first one, categorises several different emotions like, happiness, sadness, anger, fear, and disgust, and often matches the subjective experience of individuals. The second one, the dimensional approach considers emotions to have two distinct dimensions of pleasure-misery (emotional valence) and arousal-sleep. For example, the happy emotion can be pointed to have high positive valence and moderate activation or arousal. In contrast, the emotion of sadness has high negative valence and again moderate arousal. Furthermore, there is a considerable debate on which approach should be adopted by researchers[38], however we decided to use the dimensional approach because we wanted to have a numerical measure of emotional valence, which than can be compared to the objective data.

Also, emotion processing depends on top-down(appraising a situation based on similar previous knowledge) and bottom-up(processing external stimuli) processes. The first one(top-down) is more cognitive process because it uses attention and memory in order to assign a valence to a given stimuli. The bottom

up processing is influenced by external stimuli, so it uses more visual perception. Generally, there are considerable differences between them, and many theorists considers which one of them is more involved with emotion generation. However, there are many "appraisal theories" that suggest cognition strongly influences when we experience emotional states and what particular states we experience in a certain moment. For instance, one can appraise a non-threatening situation as threatening and therefore she will experience negative valence. Moreover, many theorists argue that appraisal is both conscious and automatic. Smith and Kirby[90] suggested two types of appraisal: one that is based on reasoning(deliberate thinking), and one that is based on activation of memories(automatic processes). The first one is claimed to be slower and more flexible.

It has been assumed that emotional states influence cognition[11], like attention, perception, decision making, and others. Also, what we remember from a situation is strongly affected by what we attended during that time[35]. In addition, an influential theory by Easterbrook[34] suggest the number of attentional cues processed declines as arousal or anxiety increases. It can be interpreted as strong negative valence causes "tunnel vision", and positive valence produces breadth of attention. For example, a individual in highly stressful situation, like auto mobile incident, will remember only the low amount of details related to the moment of the accident. In contrast, when one is experiencing positive situation, it is likely that she will remember more unrelated details about the event. However, Harmon-Jones et al.[40] argued that the above mentioned theory considering only emotional valence is over simplified and it lacks the addition of *motivational intensity*. Generally, there are two types of motivation: approach and avoidance motivation. According to the authors the approach motivation can be low, for example listening to music, or high, for instance recognizing a attractive subject from the opposite sex. On the other hand, low avoidance motivation can be exposure to unfair situation, and high avoidance motivation can be dealing with life-threatening situation. Harmon-Jones et al.[40] postulated that high motivational intensity leads to narrowing of the attention for both positive and negative experiences because it helps individuals to successfully accomplish tasks. On the contrary, low motivational intensity for both positive and negative experiences leads to attentional broadening because individuals leave spare attentional resources, in order to be able to encounter and react accordingly to a new and maybe more valuable situation.

Russell[84] proposed a valence model of emotion which states that significant higher activation of the left hemisphere compared to the right was associated with positive emotions, whereas significantly higher activation in the right hemisphere should be associated with negative emotion. Therefore, we can infer whether an affect was positive or negative if we measure the brain activation during an experiment.

Generally, there are two methods that are widely used for measuring affect: objective and subjective. One subjective measurement technique is the self assessment manikin(SAM)[12], which we used because it provides two scales - emotional valence and arousal. Other widely used measure is the Positive and Negative Affect Schedule(Panas)[96]. However, it is more complex to use, for example, explain to participants how to use it, and also, the procedure takes more time. However, we want to avoid that because of the increasing uncomfortability of wearing fNIRS device with the time. Also, the scales had visual

icons that help participants recognise what each point of the scale means. The objective techniques that try to infer emotional valence include psychophysiological measures of galvanic skin response, heart rate, respiration[53], and brain activity[6]. We decided to combine subjective and objective measure of emotional valence, in order to compare them and later make inferences about. We chose SAM because it is widely accepted, and easy to implement, and brain sensing technique for the objective measure, as we try to prove the left vs right hemisphere hypothesis of Davidson[29] which is explained below. Next, we will review brain sensing techniques and pick one suitable for this master thesis experiment.

2.4 Brain sensing

Initially, brain sensing was used in for medical purposes, however in recent years it has been used in other fields, like cognitive psychology, and lately, for Brain-Computer interface(BCI). We review only brain sensing techniques that are suitable for measuring complex cognition and are non-invasive so it can be used in HCI experiment. First of all, the functional magnetic resonance imaging (fMRI) is measuring the haemoglobin oxygenation and deoxygenation in the brain which is referred to as blood oxygen level-dependent contrast(BOLD) signal. It uses a large magnet which causes a strong magnetic field and a short radio-frequency pulse is emitted, in order to detect areas of activation in the brain. Furthermore, MRI has high spatial resolution which detects activation in brain regions with up to 1mm precision and is suitable for distinguishing particular brain regions of activation for the studied task. Also, its temporal resolution has 2-3 seconds delay, as it takes some time for the neural activity to occur and be detected. Typical fMRI studies involve emotion induction[75, 13] and mental arhytmetic[51]. However, during experiments participants should stand still and even the slightest movements can cause artefacts and distort the signal. Therefore, this technique is not suitable for HCI evaluations because subjects cannot physically move.

Another widely used brain imaging technique is the EEG which measures electrical changes at the surface of the scalp. It is non-invasive technique, which makes it suitable for cognitive and HCI research. Waveforms with different bands are calculated from the electrical signal that can later be analysed. The strength of EEG is its temporal resolution as it can detect changes in brain activity with accuracy of a few milliseconds. However, it is highly susceptible to motion artefacts, and even the slightest movement of fingers can cause deformation of the signal. Consequently, it is not suitable for web interface evaluations because when users type on the keyboard or use the mouse will cause considerable distortions of the brain signal.

A more recent brain imaging technique is the functional near infrared spectroscopy (fNIRS) which unlike EEG is an optical-imaging modality. Furthermore, fNIRS can detect cerebral hemodynamics by calculating the oxygenated haemoglobin(Hbo) and deoxygenated haemoglobin(Hbr). It uses infrared light which is emitted to the participants skull using emitter-detector pairs consisting of infrared LED emitter, and infrared sensors. They usually operate with two or more wavelengths (650-1000nm)[87]. Furthermore, because Hbo and Hbr have different light absorption coefficients and the infrared sensors can detect

the reflected light from them, the concentration of Hbo and Hbr is calculated using modified Beer-Lamberts law[31]. The fNIRS has low temporal resolution with a 2-8 sec delay[46, 92] depending on the task. However, it has good spatial resolution and can detect signals 1cm inside the cortex depending on the emitter-detector configuration, and has shown good correlation with the fMRI data[26] for cognitive tasks. Another advantage is that it provides a continuous data and different periods of the task can be defined and later analysed. In addition, fNIRS devices are becoming more and more portable, have high spatial resolution, and most importantly they have low sensitivity to motion artefacts which makes it particularly suitable for HCI studies[60, 92]. Hence, we decided to use fNIRS because of its applicability for usability experiments. For more comprehensive review of the fNIRS brain imaging instrumentation and methodology Scholkmann et al[87] article provides detailed information. Next, relevant studies that try to infer mental workload and emotional valence using fNIRS will be reviewed in the subsections below.

2.4.1 PFC, cognition and emotional processing

Generally, the prefrontal cortex(PFC) has been associated with higher cognitive functions by studies examining brain damaged individuals[88, 91]. Also, experiments on healthy subjects using the n-back task[14] have supported this claim. More specifically, activation was observed in the dorsolateral prefrontal cortex (BA 9/46), inferior frontal (BA 6/44) and parietal (BA 7/44) when the task demanded more working memory resources. However, it is difficult to point which brain region is involved with which processes because one brain area is usually involved in multiple cognitive tasks[17]. Moreover, Yarkoni[104] supported that claim by reviewing 3489 studies, which considered areas of human brain. Interestingly, activation in the same brain regions (dorsolateral prefrontal cortex, anterior insula and anterior cingulate cortex) was observed in one fifth of the studies. He used a machine learning algorithm and classified different studies into ones that assessed working memory, emotion and pain. It also, can be seen from the graphs that both working memory and emotion processing studies evoked activation in the PFC. Furthermore, individual differences in brain structure exist, especially in the PFC area[93]. In addition, gender differences in the brain structure are also observed[25].

There are also evidence that the PFC is involved with emotion processing and emotion regulation[30, 28, 5]. More specifically, the ventromedial prefrontal cortex is suggested to be involved in the representation of positive and negative emotional states, and the dorsolateral prefrontal cortex in the representation of the goal states towards these affective states are directed. Also, it is suggested that amygdala processes threat stimuli and processes negative affect of fear[30]. Furthermore, the orbitofrontal cortex in the PFC has been linked to reward processing and reinforcement learning[81], therefore playing a role in the assignment in emotional valence and intrinsic motivation.

Davidson proposed the "valence asymmetry hypothesis"[29] which states that positive affect which is linked to approach motivation is experienced when the left frontal cortical region has higher activation than the right. In contrast, negative affect which is linked to avoidance motivation is experienced when more activation is observed in the right frontal cortical region compared to the left. However, the valence model of emotions by Russell[84] does not connect

emotional valence with approach-avoidance motivation because the emotional state of anger. Consequently, we are interested in using fNIRS in order to measure the activations in left and right frontal hemispheres and interpret the results using the *valence asymmetry hypothesis*. We will review more relative studies that used fNIRS to investigate the relationship between left and right hemispheres in the "fNIRS and Emotional Valence" subsection below.

2.4.2 Fnirs and mental workload

In this subsection we review studies that used fNIRS in cognitive tasks. To begin with, fNIRS has been placed in different brain regions like, prefrontal cortex[2], motor cortex[44] and auditory cortex[77]. However, because a positive correlation has been observed between the hemodynamic data from the prefrontal cortex and mental workload[72] we review only studies that are measuring PFC hemodynamic activation. Generally, fNIRS has been used in various tasks, including remotely operating vehicles[33, 2], mental arithmetic[76], n-back tasks[33, 2], and other complex cognition tasks like video games[48, 18, 2]. However, those studies measure whether there is difference in the hemodynamics between conditions when changing the difficulty of the tasks. According to our knowledge there is only one study using fNIRS for evaluation of desktop visualisation interfaces, and more specifically comparing bar graphs and pie charts[74]. They tested 16 participants, however, they could not find statistically significant difference in Hbr activation between the two graphs. Furthermore, it has been suggested that Hbo activation tends to be more sensitive compared to Hbr and Hbt during mental task[65]. Another study by Plichta et al.[78] studied the reliability over time of the hemodynamic data from fNIRS and concluded that Hbo data was more reliable and stable than Hbr data. Therefore, we will use primarily Hbo for inferring mental workload and emotional valence, although Hbr and Hbt data will still be examined and results reported.

2.4.3 fNIRS and emotional valence

In this section we review studies that use fNIRS data to infer about affective states measuring the hemodynamic activity of the PFC. It was not found a usability study incorporating fNIRS for measuring emotional valence. Most of the research papers use a set of predefined pictures, videos, face expressions, and words to induce positive or negative emotions.

Most of studies use a set of affective pictures[54] to elicit emotion. Balconi et al.[7] used multiple measures from fNIRS, EEG, skin conductance, and heart rate showed affective pictures to test the valence asymmetry hypothesis. They also used the SAM subjective questionnaire to relate the psychophysiological to the self reported one. The measured Hbo from the fNIRS in the frontal right hemisphere was significantly higher when negative pictures were presented to participants. Moreover, the EEG and skin conductance data yield equivalent results. Also, it was found out that negative stimuli were more arousing according to the data from skin conductance measure. Generally, the results from this study were in accordance to the valence asymmetry hypothesis. Another study found that positive happy face expressions were identified faster and the reaction times were lower compared to negative or angry expressions[82]. Also, right handers reacted faster to positive face expressions with their right hand, and to

negative expressions with their left hand. The opposite pattern was observed for the left handers which reacted faster to positive stimuli with their left hand and to negative with their right hand. A similar study[52] used words and facial expression to induce affect, and yield identical results as the previous one. This supports the body-specificity hypothesis[21] which states that right handers associate positive emotions with their right hand and negative emotions with their left hand. The and opposite pattern should be observed for the left handers. Consequently, it left-handers and right-handers will activate different areas of the brain viewing the same affective stimuli. Taking in consideration the limited time we have for this master thesis, and the need for significant statistical difference, we should study either right-handed or left-handed individuals.

Furthermore, a study showing a emotional video clips[55] to participants deducted that there was a gender difference in the delay period to initial PFC activation. Later, the same authors made a similar experiment [56] using video clips to elicit affect found out that there was activation even after finishing the video clip and the off periods should be adjusted so that post-stimuli activation do not overlap with the next condition.

However, the above mentioned studies are designed to induce emotions, but want to study the opposite effect: can we obtain emotional valence rating from neutral web interface using fNIRS, and then infer which of the variations was objectively more preferred?

2.5 Summary

To sum up, many researchers do not consider the positive or negative affective state during task execution, and how it influences operator performance and perceived workload, we have combined multidimensional subjective workload scale (NASA-TLX) with emotional valence scale (SAM). We expect to gain better understanding of the operator performance and motivation during the web form filling task. Also, we expect to find correlations between these subjective measures and the objective measure of the mean Hbo from the fNIRS device because they are both suggested to measure mental workload and emotional valence.

Chapter 3

User study

3.1 Hypothesis and expectations

Based on the literature review we state the following hypothesis:
1) There is significant statistical difference between the three web forms
2) Subjective data from NASA-TLX correlates to objective data from the fNIRS
3) The difference in left and right hemisphere activations correlate to subjective SAM scale of emotional valence

We are aiming to reduce to imposed load by the task by minimizing the visual search(index3) and aiding the episodic memory by placing a description in the beginning of the web form(index2) compared to the control condition(index1). Thus, we expect to find differences between the three conditions. Because the number of attended cues affects the operator performance, we decided to split the web form into 3 pages, in order to minimize visual search and clutter.

Hypothesis 2 was drawn from the fact that both NASA-TLX and the mean Hbo data from fNIRS have been successfully used to measure the concept of mental workload, therefore, we expect to find correlations between the two measures.

In addition, because fNIRS can measure the left and right hemisphere activation, which is correlated to positive and negative affect based on the valence asymmetry hypothesis, thus we can infer about the affective and motivational state of the participant. And also, hypothesis 3 was stated because the subjective SAM scale is valid measure for obtaining emotional valence(the affective state), we expect to find correlations between the SAM and left/right hemisphere Hbo activation.

3.2 Method

We have use fNIRS because it is non invasive, motion artefact resistant, and with high temporal resolution and, thus suitable for using in user trials. We combine the objective measure of fNIRS with the subjective NASA-TLX and SAM measures, as they are validated and reliable methods. This way we gather comprehensive information about how participants perceive the interfaces and relate it to the psychophysiological data. Also, using fNIRS is novel method for

conducting usability studies, and we also, aim to assess its practicality in this kind of HCI evaluation studies.

3.2.1 Participants

A total of 20 right handed participants (9 female) with mean age of 26 (SD = 4.04) took part in the study. All of the participant were healthy, however only one pointed that it suffers "Von Willebraud disease" which impairs blood's ability to clot, and his data from the psychophysiological measures was excluded. All participants had normal or corrected to normal vision, and report no history of brain damage. Also, 14 of them reported they have advanced computer literacy, 5 of them stated average computer literacy and 1 did not answer this question. All of the participants were current or graduated students. The ethics committee of the University of Nottingham approved this study. Informed consent was obtained from the participants and they were compensated with 10 Amazon voucher.

3.2.2 Apparatus

Laptop computer

The experiment was executed on 15" laptop, HP probook 450 with screen resolution 1366x768. An external mouse was attached, which participants used during the process. The participant was presented with a screen with links to the three different videos and web forms. They were instructed by the researcher to manually start certain condition or video.

fNIRS

(Picture of fNIRS and emitter-detector pairs)-can I use the one from the TAP paper?
Hemodynamic data was recorded using the fNIRS300 device along with the COBI studio recording software developed by Biopac Systems inc. The device consists of a headband with 4 infrared LED emitters and 10 infrared detectors. They operated on 730nm and 850nm wavelengths. The combination between them was used to calculate 16 channels which can measure the associated Hbo and Hbr concentration in the PFC. The fNIRS device was placed on the forehead of the participants targeting the dorsolateral(BA 9/46) and orbitofrontal(BA 10) cortices.

3.2.3 Materials

NASA-TLX

The NASA-TLX[43] is multidimensional subjective scale and it is used as a tool for assessing operator workload based weighted average of its six scales. The individual scales are presented in the following order: Mental demand, Physical demand, Temporal demand, Performance, Effort, and Frustration. We used the paper version of the questionnaire. The NASA-TLX measures were obtained from participants each time after completing one of the web form conditions. We did not follow the weighting procedure because it was time consuming,

and instead we calculated the mean values of all individual subscales, which is suggested to be also, a valid measure[42], and we refer to this variable as "total tlx". Furthermore, each of the individual subscales was analysed independently.

Self assessment manikin scale (SAM)

Self assessment manikin[12] is a two dimensional scale for measuring the perceived emotional valence and arousal. We implemented the 5 point version of it. The participants were asked to fill the questionnaire after each video watched and each web form filled. First they state their emotional valence(negative or positive) by choosing between 1 and 5, where 5 is strongly perceived positive emotion, and 1 is considered strong negative emotion. Then, they fill the arousal level scale, where 1 indicates low perceived arousal or boredom, and 5 signifies high perceived arousal or high level of excitement. Each of the subscales is supported by image visualisations illustrating the affective state.

Web forms

A total of 3 HTML/CSS variations of a standard web form for insurance claiming were produced. They were created to resemble an actual online insurance claim form. All of the conditions were divided on 3 main divisions: Personal information, Accident information and Summary of accident. The personal information division consisted of 5 individual forms, namely: First name, Last name, Date of birth, Email, You are(choose type of stakeholder, which was option field). The accident information consisted of one text input field(Today's date), four drop down lists(Number of passengers, Number of cars involved in the accident, Was anyone injured in this incident, and Was the accident caused by your fault), and a checkbox form for selecting which areas of the vehicle were damaged. Lastly, the third division consisted of a text-area, where participants had to write a description of the accident. The first version or the control version, referred as index1, consisted of the 3 division areas laid out in the order that was described here, namely, at the top of the page is the personal information, followed by accident information and lastly summary of accident. In the second web form, which we refer to as index2, the summary of accident area was placed on the top of the page, followed by personal and accident information, accordingly. The third condition, referred as index 3 had the same order as index1, however each division area was situated on separate page, and consisted of total 3 pages. Users navigated between the three pages using a submit button with the label"Next". Also, on the top of the form below the heading there was a progress feedback text indicating of how many steps the web form consisted. For more detailed information, the three web form conditions can be viewed in Appendix number

Video capture

A video capture of the computer screen was recorded during the experiment using Fraps[59]. The participants voice was recorded too. The timestamp of the beginning of the video recording was obtained, in order to be able to calculate duration of tasks, and their actual start and end time.

Video clips of auto accidents

Three video clips of automotive accidents were selected, in order to simulate the conditions before the filling of insurance claim. The video clips of the accidents were chosen to be lightweight avoiding any scenes of gore, injured bodies, or fatalities. All of the accidents happened in low speed. One of the clips has a duration of minute and a half, while the others were half a minute long. The three videos can be seen in...

3.2.4 Design

The study used repeated measures within subjects design. The depended variable was the task completion time(or user preference), and the independent variable was the layout of the three web forms. The control condition was index1. Before the start of each web form, a video of road accident was played. The three variations of videos and the web forms were counterbalanced using latin square rotation, in order to avoid learning effects from the order of presentation of the video clips.

3.2.5 Procedure

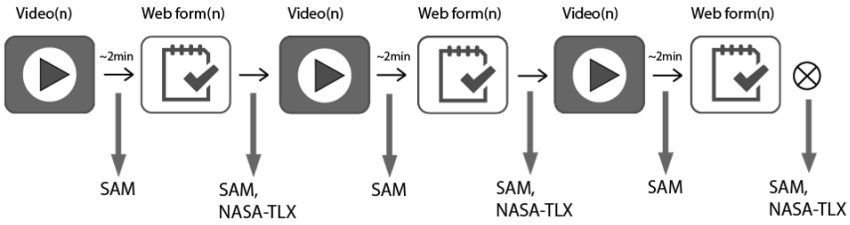

Figure 3.1: The image illustrates, the study procedure followed in this experiment. First participants, watched video, then fill the SAM measure. After that, they filled the a web form, and then filled both the NASA-TLX and SAM scales. The process was repeated 3 times.

Participants followed the procedure illustrated by Figure3.1. First, participants were asked to read and sign information sheet and consent forms. Second, the fNIRS device was cleaned then equipped and started. Third, participants were briefed about the procedure of the experiment, and it was explained how to fill the subjective scales. Also, because of ethical considerations that the participant should not enter personal data in the web form, a artificial personal credentials were provided, that she should fill in the web forms. Fourth, after the video capture and the fNIRS device were started, participants were asked to relax and try not to think about anything, in order to record a baseline of the hemodynamic activity in the PFC while participants are at rest. Next, they are asked to open one of the three videos, depending on the counterbalancing table. After the video was finished, participants fill SAM subjective scale. Fifth, there was approximately 2 minute waiting period between the video and the web form filling task, so that participant's memory is not fresh. Finally, after participant has completed the web form, the SAM and then the NASA-TLX scales are given

to be completed, accordingly. This process was repeated three times, following the within subjects experimental design with counter balancing between the videos and the web forms. Because fNIRS device used one computer and the experiment was conducted on different computer, before each experiment, the clocks between the two computers were synchronized. Also, timestamps using the Cobi Studio software manual markers were created in the beginning and end of each condition and video.

3.2.6 Data Analysis

The fNIRS data was analyzed with fnirsoft[1]. Data from N=1 participant was excluded from the analysis because of "von willebraud disease" which is known to alter the signal. Also, data from another N=8 participants was excluded due to the fNIRS apparatus not able to detect signal from the channels for those participants. When calculating the correlation between fNIRS data and other measurements we excluded the data from these 9 problematic participants from the subjective or performance measures accordingly.

Signal acquisition

The fNIRS headband was placed on participants forehead, targeting the prefrontal cortex. The emitter-detector separation was 2.5cm and the sampling rate was 2Hz.

Preprocessing

Instrument noise was reduced by placing a hat over the fNIRS headband, in order to block external light. First, low-pass filter with cut off frequencies of 0.1 Hz, was used in order to remove physiological noise, like heartbeat and blood flow movement that is not associated with brain activity or Mayer waves. Then, the NIRS signal was processed with modified Beer-Lambert law[24], in order to calculate oxygenated, and deoxygenated hemoglobin values. Finally, to remove motion artefacts, the correlation based signal improvement(CBSI)[27] method was applied to the data.

Feature Extraction/selection

fNIRS mean Hbo, Hbr, and Hbt data
After data preprocessing the mean, and standard deviation for Hbo, Hbr and Hbt data was calculated from all channels, in order to infer about activation in the participants PFC. Hbo has been suggested to positively correlate to mental workload, in contrast to Hbr which is proposed to have negative correlation to mental workload.

fNIRS mean differences
To calculate the left versus right hemisphere activation from the fNIRS we subtracted the mean data from channels 1-8, which are situated at the left side of the participants PFC, from the mean data of channels 9-16 which are on the right side of the PFC. That gave us the difference value between the left and right hemisphere. It indicates that the higher the mean difference value

the more positive affect or affordance motivation the web form elicited. And the opposite pattern: the lower the mean difference the more negative affect or avoidance motivation the participant experience according to the valence asymmetry hypothesis.

3.3 Results

3.3.1 Mental Workload

fNIRS data

A one-way repeated measures ANOVA was conducted to determine whether there was a statistically significant difference in mean Hbo values between the three web forms. The assumption of sphericity was met, as assessed by Mauchly's test of sphericity, $X^2(2) = 0.195, p = 0.907$. There was no significant statistical difference in the mean Hbo between the 3 web forms $F(2, 20) = 3.400, p < .054$, partial $\eta^2 = .254$ with mean Hbo decreasing from 0.2377 (SD = 1.19) in index3 to -0.1166 (SD = 0.82) and -0.117 (SD = 1) for index2 and index1 respectfully. Which means that index3 indicates the highest workload, compared to the rest of the conditions. After conducting t-tests between index1

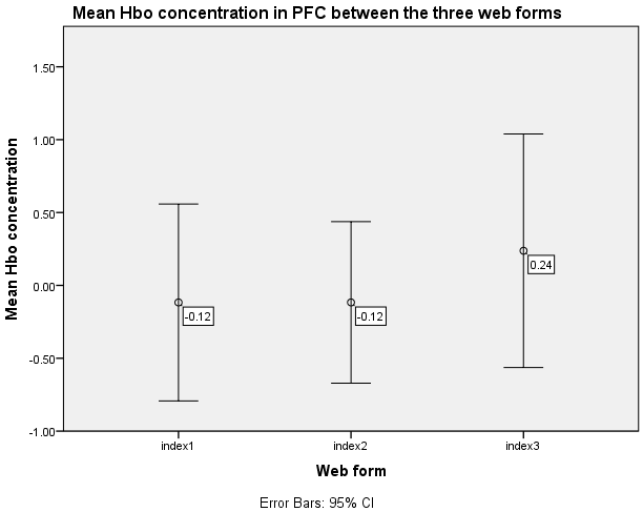

Figure 3.2: Mean Hbo activation between the three web form conditions as measured by fNIRS. Higher Hbo values indicates higher mental workload experienced by the operator.

and index3 there was found a marginal statistical difference p=0.054. Which means that the measured Hbo between the two versions has a 94.6% statistical probability that the difference are not caused by random sampling error. However, we fail to reject the first null hypothesis. This means that we could not find a statistically significant interaction, however we were very close to statistical significance, as p=0.054 and we can assume we have marginal statistical significance.

Also, no statistical significance was found when comparing the means of Hbr between the three conditions $F(2, 20) = 2.044, p < .156$, partial $\eta^2 = .170$ where index2 had the highest Hbr mean 0.05 (SD = 0.85), index1 with -0.07 (SD = 0.96) and index3 with the lowest Hbr mean -0.36 (SD = 1.43). Which, accordingly negatively correlated to Hbo data, and again indicated that index3 evoke the highest workload than the other conditions. Furthermore, a repeated measures ANOVA test was conducted to elicit significant statistical differences between mean Hbt between the three web forms, however no statistical significance was found $F(2, 20) = 0.685, p < .516$, partial $\eta^2 = .064$ where index2 had the highest Hbt mean -0.08 (SD = 0.49), index3 with -0.13 (SD = 0.58) and index1 with the lowest mean Hbt -0.19 (SD = 0.49). The results indicate that Hbo was more responsive for this experiment and gave us higher significance between condition compared to Hbr and Hbt. No correlation was found between the fNIRS mean data and any of the NASA-TLX scales, including the calculated total tlx. As a result, we fail to reject the second null hypothesis which states there is no difference between objective data from fNIRS and subjective data from NASA-TLX.

NASA-TLX

We report data for the NASA-TLX measure for all of the participants without excluding anyone because there was no problem with obtaining the data for this measure. All of the calculated data can be viewed in Table 3.1. There was no statistical significance between each of the NASA-TLX scales, including the total$F(2, 38) = 0.743, p < .482$, partial $\eta^2 = .038$ score as assessed by one way repeated measures ANOVA. Which means statistically we have 51.8% chance of the results of the total NASA-TLX score to be caused by random sampling error. Also, perceived mean mental demand was lowest for index1 9.15 (SD = 4.94), index2 had slightly higher mean 9.40 (SD = 4.68) and index3 has the highest scores 10.8 (SD = 5.38). Also, mental demand had a strong positive correlation with total tlx for the 3 conditions $r(18) = 0.652, p = 0.002$, $r(18) = 0.738, p < 0.001$, and $r(18) = 0.741, p < 0.001$ for index1, index2 and index3 respectfully. Which supports the validity of NASA-TLX measurements. The total calculated value for the NASA-TLX was highest for index3 7.07 (SD = 3.22) decreasing to 6.92 (SD = 2.95) for index1 and 6.47 (SD = 3.11) for index2. This means that index3 requires slightly more attentional resources to complete the task compared to index1 and index2. There was a moderate positive correlation between mental demand scales and task completion times between the three conditions $r(18) = 0.487, p = 0.030, r(18) = 0.484, p = 0.030, r(18) = 0.638, p = 0.002$. The results show that the more participants perceived higher workload the more their performance dropped as it took them more time to complete the task.

SAM - arousal scale

As it can be seen from Figure 3.3 the perceived arousal was lowest for index1 2.8 (SD = 0.95) increasing to 2.95 (SD = 1.05) for index2 and to 3.15 (SD = 1.18) for index3 respectfully. No statistical significance was found when comparing the means between the three conditions $F(2, 38) = 2.462, p < 0.099$ partial $\eta^2 = .115$ using one way repeated measures ANOVA. However, after running

Table 3.1: A table of all of the calculated mean NASA-TLX values for the 6 subscales, including the averaged total tlx

	Index1	Index2	Index3
Mental demand	9.15 (SD = 4.94)	9.40 (SD = 4.68)	10.8 (SD = 5.38)
Physical demand	4.05 (SD = 4.08)	2.90 (SD = 3.21)	3.90 (SD = 3.65)
Temporal demand	7.40 (SD = 4.49)	7.65 (SD = 5.79)	6.55 (SD = 4.91)
Performance	6.65 (SD = 3.79)	5.60 (SD = 3.62)	6.20 (SD = 3.86)
Effort	8.15 (SD = 4.58)	7.35 (SD = 4.68)	8.20 (SD = 5.30)
Frustration	6.10 (SD = 5.11)	6.00 (SD = 3.66)	6.75 (SD = 5.22)
Total	6.92 (SD = 2.95)	6.47 (SD = 3.11)	7.07 (SD = 3.22)

post hoc test without adjustments(LSD) a statistically significant difference was found between index1 and index3 $p = 0.049$. Which means that perceived arousal was significantly higher for index3 compared to index1. Also, the time to complete index1 and index2 positively correlated to perceived arousal for index 1 and index2:$r(18) = 0.551, p = 0.012$ and $r(18) = 0.473, p = 0.035$. However time to complete index3 does not correlate to perceived arousal of index3 $r(18) = 0.269, p = 0.252$. Which indicates that we have a partial positive correlation between the perceived arousal, which can be considered as workload, for the web forms and the time to complete them.

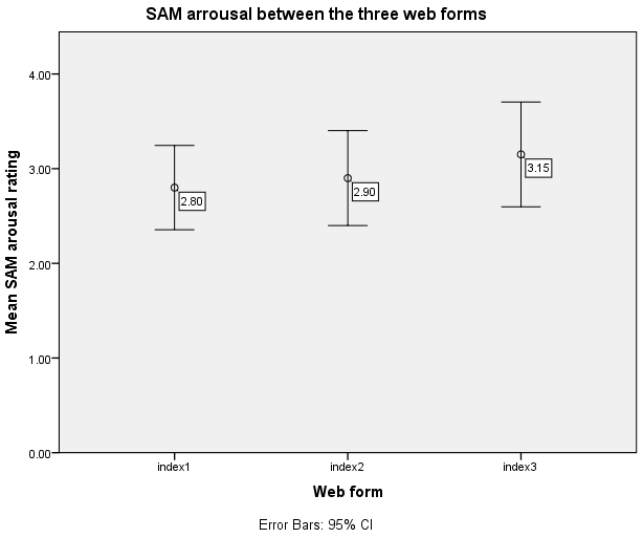

Figure 3.3: The mean SAM arrousal rating obtained from the three web form conditions.

3.3.2 Emotional Valence

fNIRS differences

Data from the period of web form filling

To begin with, Figure 3.4 shows the valence differences in Hbo concentration

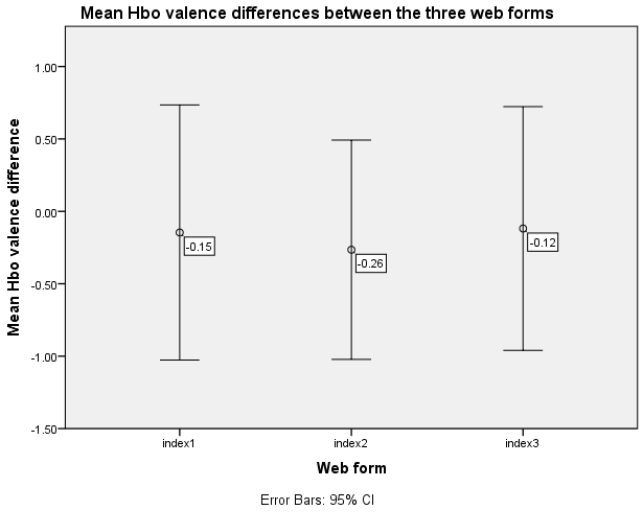

Figure 3.4: The mean Hbo valence differences obtained from the fNIRS for the three web form conditions. Higher values indicate more positive affect, and lower value more negative affect.

between the left and right hemispheres. fNIRS Hbo valence differences was highest for index3 -0.12 (SD = 1.25) decreasing to -0.15 (SD = 1.31) for index1 and the lowest value was for index2 -0.26 (SD = 1.13). Indicating that participants experienced the most positive when completing the index3 condition, slightly less positive for index1 and the least positive when completing index2. However significant statistical difference was not found $F(2, 20) = 0.392, p < 0.681$, partial $\eta^2 = .038$ as assessed by one-way repeated measures ANOVA.

The Hbr valence differences were highest for index2 0.34 (SD = 1.18) decreasing to 0.18 (SD = 1.39) for index3 and to 0.10 (SD = 1.68) for index1. Signifying that index2 elicited the most negative affect compared to the rest of the conditions. However, the difference was not statistically significant. The Hbt mean valence difference values for index1 were lower -0.45 (SD = 0.79) compared to index2 0.06 (SD = 0.56) and index3 0.05 (SD = 0.87) respectfully. There was no statistical significance as assessed by one way repeated measures ANOVA between the three conditions for Hbr valence differences: $F(2, 20) = 0.418, p < 0.664$, partial $\eta^2 = .040$ and Hbt valence differences: $F(2, 20) = 0.302, p < 0.743$, partial $\eta^2 = .029$. Also, there was strong positive correlation between temporal NASA-TLX scale of index1 and the Hbo valence differences of index1 $r(9) = 0.766, p = 0.006$, however, there was no correlation found between index2: $r(9) = 0.581, p = 0.061$ and index3: $r(9) = 0.218, p = 0.519$ which suggests that as participants perceived more temporal demand the obtained mean Hbo data increased.

Data from the period of video clips

The mean Hbo valence difference for video3 was the highest with 0.17 (SD= 0.25) compared to video1 with -0.01 (SD = 1.32) and video2 with -0.8 (SD = 1.20), suggesting that participants experienced more positive emotion when watching video3 compared to video1 and video2. In contrast, mean Hbr valence difference values for video3 were the lowest with -0.25 (SD = 0.47) compared to video1 0.30 (SD = 1.20) and video2 0.32 (SD = 0.89). For the mean Hbt valence difference values video1 was the highest with 0.18 (SD = 0.71) decreasing to 0.07 (SD = 0.64) for video3 and to -0.06 (SD = 0.35) for video2.

A one-way repeated measures ANOVA was conducted to determine whether there was a statistically significant difference in Hbo, Hbr and Hbt valence differences between the three videos. There was no significant statistical difference in the mean Hbo valence difference between the 3 videos $F(2, 20) = 0.051, p < 0.951$, partial $\eta^2 = .005$, the mean Hbr valence difference: $F(2, 20) = 0.062, p < 0.940$, partial $\eta^2 = .006$ and the mean Hbt valence difference: $F(2, 20) = 0.522, p < 0.601$, partial $\eta^2 = .050$. Consequently, we cannot find significant difference in the emotional valence from the objective data between the three videos. There was no correlation found between Hbo valence differences and SAM emotional valence subjective scale for the three videos $r(9) = -0.490, p = 0.126$; $r(9) = 0.095, p = 0.781$; $r(9) = 0.496, p = 0.121$. As a result, we fail to reject the third null hypothesis, which states that there is no correlation between the fNIRS valence difference data and the SAM subjective scale of emotional valence.

SAM emotional valence

Data from the period of web form filling

The perceived mean emotional valence for index1 was the lowest with 3.1 (SD = 0.97) increasing to 3.4 (SD = 0.99) for index2 and to 3.7 (SD = 0.98) for index3, as it can be seen from Figure 3.5. A one-way repeated measures

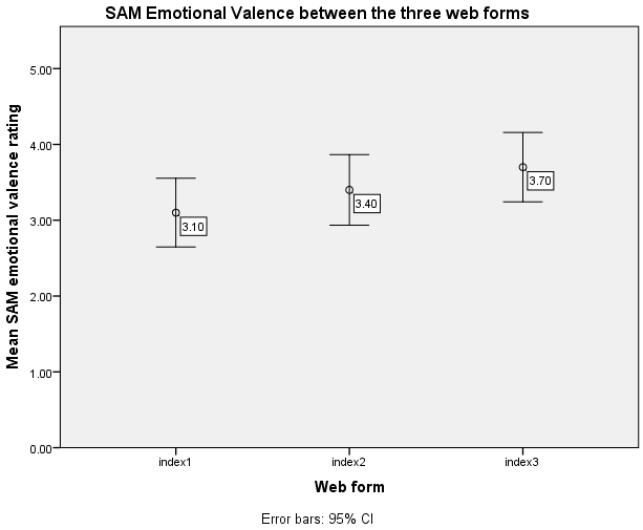

Figure 3.5: Perceived emotional valence between the 3 web form conditions

ANOVA was conducted to determine whether there was a statistically significant difference in SAM emotional valence scale values between the three web forms. The assumption of sphericity was met, as assessed by Mauchly's test of sphericity, $X^2(2) = 0.446, p = 0.800$. There was no significant statistical difference in the SAM emotional valence scale between the 3 web forms $F(2, 38) = 2.803, p < .073$, partial $\eta^2 = 0.129$ with mean SAM emotional valence increasing from 3.1±0.97 in index1 to 3.4±0.99 and 3.7±0.98 for index2 and index3 respectfully. This means we cannot distinguish which web form participants perceived as more positive or negative, however we had marginal statistical significance which gives us high probability that the results were not being biased by random sampling error.

Data from the period of video clips
The perceived mean emotional valence for video3 was the highest with 3.1 (SD = 1.07) decreasing to 2.9 (SD = 1.33) for video1, and 2.7 (SD = 0.98) for video3. There was no statistical significance as assessed by one way repeated measures ANOVA between the three videos for SAM emotional valence: $F(2, 38) = 0.792, p < 0.460$, partial $\eta^2 = .040$. The data suggest that we cannot compare the perceived emotional valence between the 3 video due to high chance of the results being caused by random sampling error.

3.3.3 User performance and preferences

Task completion time

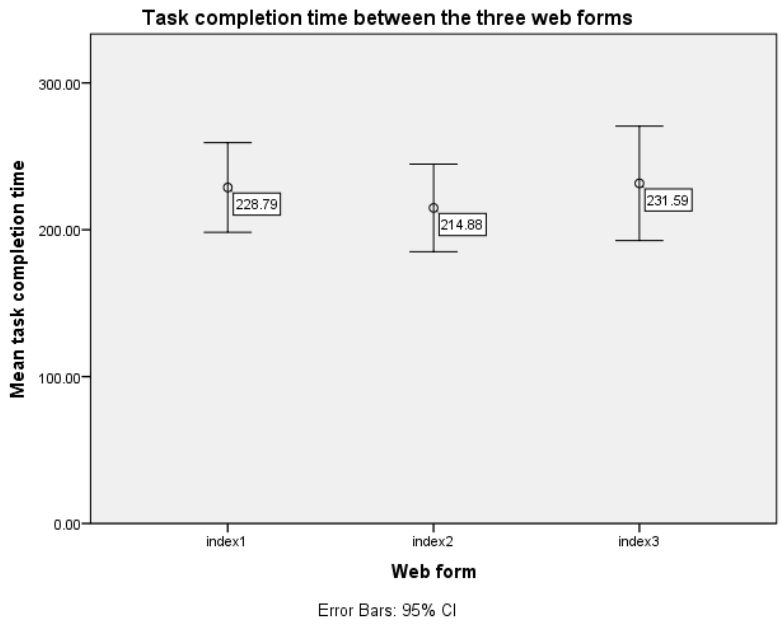

Figure 3.6: Mean task completion time between the three web forms

It can be seen from Figure 3.6 that the mean time to complete index2 was the lowest 214.88 (SD = 63.81) increasing to 228.79 (SD = 65.19) for index1 and to

231.60 (SD = 83.33) for index3. However, there was no significant statistical difference in time to complete between the 3 web forms $F(2, 38) = 0.556, p < .578$, partial $\eta^2 = .028$, as assessed by one-way repeated measures ANOVA. The results indicate that the performance was lowest when participants completed index3, slightly higher for index1 and the highest for index2. Also, the time to complete index2 and index3 had a strong positive correlation with perceived effort(NASA-TLX) for index2 and index3: $r(18) = 0.702, p = 0.016$ and $r(18) = 0.634, p = 0.036$. However, time to complete index1 does not correlate to perceived effort of index1 $r(18) = 0.216, p = 0.524$. This suggests that when participants perceived more effort it took them more time to complete the web form.

User preferences from the short-interview question

After the end of the experiment participants were asked which of the three web forms they prefer the most. The bulk of them preferred index3 and index1 with 10 and 9 votes respectively compared to index2 which was only preferred by 3 participants. The given answer were transcribed as can be seen in Appendix number.

Chapter 4

Discussion

4.1 Comparison between the three web forms

4.1.1 Control condition - Index1

When analysing the results it can be noted that index1 was the least mentally demanding web form, according to the mental demand scale of NASA-TLX, and objective measure of mean Hbo activation, and the SAM arousal scale(statistical significance $p = 0.049$ compared to index3). However, the calculated total tlx indicated higher perceived workload than index2, and lower than index3. Index1 can be said to support the consistency heuristic[68] because it was the control condition and represented a typical insurance claim form. The comparatively low workload measured for index1 can be attributed to the fact that the task was familiar, and therefore participants had to spending less working memory resources on planning because the task sequence was already mapped in the brain. One participant supported this argument by saying: *"I consider it more logical, like logical approach..to say who you are, what happened to your car and then what happened exactly, describe it from your own perspective."*. Also, 4 users reported that the sequence of small question first, leading to the description at the end, helped them to recall the situation better: *"I preferred the first(index1) and third(index3) examples, as they asked more general questions at the beginning. This made me think about what aspects to include in the event summary at the end, such as number of cars involved"*. Which means that the sequence of web forms or their order of presentation is important aspect to consider.

Furthermore, index1 induced the least perceived emotional valence from the SAM questionnaire, compared to the rest of the conditions, with marginal significance of $p = 0.073$ when compared to index3. However, the objective data from the fNIRS differences demonstrated higher left hemisphere activation than index2, and slightly less than index3. Similarly, it took more time to complete the web form compared to index2, and slightly less time than index3.

4.1.2 Description at beginning - Index2

Generally, index2 elicited the least workload inferred from the mean Hbo concentration, total tlx, perceived effort, physical demand, and frustration measures.

Also, users performance was superior in index2, as the task completion time was the lowest, although not statistically significant. This can be attributed to the fact that index2 supported participants working memory better, specifically for this study design, because index2 presents the description field first, thus lowering the time to hold the visual information in the episodic memory. Consequently, users memory was relatively fresher when filling the description field in index2, which demanded the most attentional capacity, compared to the rest of the conditions.

Furthermore, index2 elicited approximately neutral emotional valence rating(3.40) which was higher than index1(3.10) and lower than index3(3.70). In contrast, the objective data from fNIRS differences reveal more right hemisphere activation which can be interpreted as more negatively valenced affect or avoidance motivation. The last finding can be supported with the fact that only 3 out of 20 participants expressed preference for index2. A possible explanation is that index2 did not support the consistency heuristic[68, 89] because the description field is normally situated at the end of insurance claim forms. This statement can be supported by the following participant comments: *"The only one I didn't like is the first one(index2) particularly. It was just describing the event before filling out the details it is a bit awkward."*, and another participant stated: *"Index1 and Index3 were Ok, but I didn't like index2 because I was not used to it."*.

4.1.3 Divided form - Index3

The purpose of this study was to improve usability of web form filling of insurance claims and find which of three web forms layouts was the most usable. After analysing the results we found out that index3 elicits the highest mental workload according to the fNIRS Hbo data which is correlated to workload, SAM arousal scale(significant difference with index1), total tlx, mental demand scale from NASA-TLX, and takes the most time to complete the task compared to the rest of the conditions. However, index3 was evoking highest positive valence(SAM), fNIRS differences were indicating more left hemisphere activation, and users preferred it the most when asked after the finish of the experiment. The results were unexpected according to the Norman[69] positive affect should enhance operator creativity, thus increasing the performance, however we found the opposite results.

Also, it can be noted that perceived temporal demand for index3 was lower than index1 and index2(no statistical significance). This can be attributed to the fact that index3 had lower number of forms(visual cues) on each page. It can be inferred that the more web forms are present on an interface, the more the user feels temporal demand because it appraises the situation as one that needs more work to be performed. This claim can be supported with the feedback from the participants: *"so the second one(index3) I felt having to click the links broke it down a little bit, like you didn't have to think about everything in one go..."* or other user expressed *"you give all the details and then you take your time to write about the incident so you don't feel very rushed or something"*.

4.1.4 Divided vs single page approach

A division can be made between the web forms which contain all forms on one page(index1, index2), and web forms which are consolidated or divided on several pages(index3). We define these as the *single page approach*, and the *divided page approach*, respectively.

A possible explanation of the higher measured positive emotional valence across all measures for the divided page approach is that users had to process less informational cues at a time, compared to the single page approach, which they appraise as a positive experience. As one of the participants mentioned *"because you don't have to think about the other forms"*. So having less web forms displayed on one screen improves visual search, decreases clutter, which imposes less temporal demand on the task. Consequently, users appraise the task as more positively valenced because it spends less attentional resources for meta-cognition.

One advantage of the single page approach is that participants can go back and check what information they have already entered: *"I remember in the second one(index1) I'm not sure whether the option provided left or right, so I rechecked"*. This way working memory resources are saved because participants has the ability to quickly recheck what they have already entered, relying on recognition, rather than recall[68] from working memory the same information again. Another advantage of the single page approach is that participants can choose which form to start first: *"the good thing about the number two(index1) is everything is on the same page I can choose whatever I like"*. Generally, some of the participants preferred to fill in the description field first, and then the rest of the forms, so researchers and practitioners have to give users the power to choose from where to start. However, in our case the participants watched a recorded accident that they had to recall after approximately 2 minutes. In reality, the delay between the accident and the claim form filling will be much higher, thus the arrangement of the web forms depends on the time between the accident and the claim form filling.

In our opinion, index1 is most suitable because it elicited less mental workload in 3 out of 4 measures for workload, and the 10 out of 20 users expressed that their prefer it. Also, index1 met the consistency heuristic, it gave more control and freedom to the user by letting her choose from which form to start first. Furthermore, index3 should also be considered because it produced most positive valence according to our subjective(SAM) and objective(mean Hbo differences) measures. In addition, 11 out of 20 users favoured this web form. We could not conclude which of the two approaches provides better user experience because singe paged approach was less mentally demanding, and slightly less preferred, in contrast to index3 which required the most attentional resources, but it received the most positive results. Furthermore, as it is not clear how mental workload or emotional valence affects the user experience we cannot conclude which is web form is more important. Also the qustion what should evaluators try to minimize: the perceived workload or the negative valenced responses, remains unanswered.

4.2 fNIRS and valence asymmetry hypothesis

Unfortunately, we could not prove the valence asymmetry hypothesis[29] because we had no power and statistical difference for both, the videos and the web forms. We also, could not find a correlation between SAM emotional valence scale and the mean Hbo differences between the left and right hemispheres. One of the reasons we could not obtain statistical significance is that we followed the opposite approach to what other studies did. They used already researched and proven frameworks which contain previously classified affective stimuli while we present them with an unclassified interface which was not previously evaluated as neither positive and negative, and tried to infer about the emotional valence based on the valence asymmetry hypothesis. Unfortunately, we obtained far from significant difference between the three conditions $p = 0.681$. This suggests that the approach we took requires testing with more participants. Which is not suitable for real life HCI evaluation studies because the limited time availability for such projects.

4.3 fNIRS practicality for HCI evaluations

One of the aims of the study was to assess the feasibility of fNIRS for HCI evaluation studies. After analysing the fNIRS data from 20 participants we had to exclude 9 because of bad data. However, we were close to reaching statistical significance between the three conditions as the p value was $p = 0.054$ with only 11 subjects examined. This gives promising results that fNIRS can be used as reliable objective measure for workload in usability studies. We propose that a study needs approximately 20 participants, in order to reach statistical significance, however that is four times more than the suggested 5 users tests by Virzi[95]. It should be noted that this recommendation highly depends on study design, and some studies may require only 10 subjects, while others more than 30.

Other advantage of fNIRS is that it provides continuous measure of the Hbo concentration in the PFC, thus it can be highly suitable for detecting distinct periods of overload. For example, it can be particularly useful in game studies where participants should stay in the flow[64] state. Also, functionally the flow experience has been suggested to be situated in the prefrontal cortex and medial temporal lobe[32], consequenly fNIRS can be used to measure it. As a result, researchers should be able to detect moments where the challenge of the game is either too difficult or too easy.

When considering the measurement of emotional valence by comparing the hemodynamic activation between the left and right hemispheres was can conclude that we were unsuccessful as we were far from statistical significance($p = 0.681$). More focused research should be conducted in order to test this approach.

In conclusion, we can infer about design issues from user trials, from which we can receive rich data for relatively less amount of money compared to the more expensive and time consuming fNIRS evaluation studies. From them we receives single continuous measure which can be interpreted in various ways. Consequently, fNIRS has limited practicality for evaluating neutral web interfaces, like web forms because it requires more time and financial resources.

However, as noted above fNIRS appears to be particularly useful in game evaluation studies where the continuous measure will provide valuable information about the level of engagement[41] and the flow states[105] of players.

4.4 Implications for Design

Based on the obtained results we propose the following recommendations for design of long web forms:

- Include visual representations of information where possible, rather than display it as text only:
  Some participants did not knew the names of the different body parts of a car and a few even used search engine to see what the different words like, bonnet and bumper meant. To address this problem, we propose that a visualisation of car divided on its main body parts should be depicted on the form in order to support recognition.

- Sequence of web forms is important aspect to consider, so testing the order of presentation of forms may yield important information:
  Some of the participants wanted to fill different form fields first, and scrolled down to find them. Consequently, researchers should consider which questions to ask first, and which last in their web forms.

- Divide a long web forms onto separate pages if the information entered in each field is not relative to the other fields, or when the users do not need to check what they have entered previously:
  During the web form filling process it can be noted that participants occasionally inspect what they have written before, in order to verify that information or check what they had written before. Designers should take into account the behaviour described above as it often observed.

4.5 Disadvantages of the study

The study tries to simulate real conditions, and therefore lacks ecological validity. Participants wait approximately 2 minutes after they watch a video to start filling the web form. In real situations the delay between the accident and the form filling is usually much bigger. Therefore, it lacks ecological validity. Also, watching videos of automotive accidents may have biased the subjective ratings of participants when they had to rate the web forms.

One of the disadvantages of near infrared scanning technology is that, in order to collect reliable measurements the emitter-detector pairs have to be in contact with the skin constantly, which causes pain and discomfort, after approximately 40 minutes of wearing the device.

We had difficulties with the fNIRS device, as some of the channels did not worked reliably. Therefore, we had to exclude a large amount of participant data that was defective or incomplete. As a result, we obtained less statistical power from the results. However, this was a problem encountered because of malfunctioning of the particular device, and not problem encompassing all fNIRS based systems.

4.6 Future work

To produce more valid results, experiments without the inclusion of videos should be conducted. For example testing the insurance quote process, instead of insurance claims. Also, he divided page approach should be further examined because it produced encouraging and positive results, except that it required the most attentional resources. However, as it is not clear which improves the user experience more, the workload experienced or the positive or negative appraisal of the interface, we should further test and diagnose the single versus divided page approaches.

4.7 Conclusion

In summary, we used functional near infrared spectroscopy in a user study of the web form filling process of insurance claim. We aimed to obtain objective and subjective measurements of mental workload and emotional valence. We tested the web page layout by producing 3 variations. After conducting the experiment the standard single page web form which is widely used, showed the least mental workload, according ot the majority of objective and subjective measures that we used to infer it. On the other hand, the web form which was divided on three separate pages, yielded highest workload compared to the rest of conditions but elicited most positively valenced ratings, and was most preferred by participants. However, we could not determine which of the web forms is more suitable for implementation, as one there is debate on which is more important for this type of interface: to elicit positive affect, or to be less mentally demanding. Moreover, experiments should be conducted to examine the difference between single and divided page approaches for long web forms in more details. Also, in our opinion, the practicality of fNIRS for similar web interface studies is limited. In addition, we could not reliably measure the level of emotional valence with fNIRS in this particular experiment. Finally, we outlined implications for design for long web forms.

References

[1] Hasan Ayaz. Functional near infrared spectroscopy based brain computer interface. *PhD Thesis, Drexel University, Philadelphia, PA*, 2010.

[2] Hasan Ayaz, Patricia A Shewokis, Scott Bunce, Kurtulus Izzetoglu, Ben Willems, and Banu Onaral. Optical brain monitoring for operator training and mental workload assessment. *Neuroimage*, 59(1):36–47, 2012.

[3] Alan Baddeley. The episodic buffer: a new component of working memory? *Trends in cognitive sciences*, 4(11):417–423, 2000.

[4] Alan D Baddeley and Graham J Hitch. Working memory. *The psychology of learning and motivation*, 8:47–89, 1974.

[5] Michela Balconi and Adriana Bortolotti. Detection of the facial expression of emotion and self-report measures in empathic situations are influenced by sensorimotor circuit inhibition by low-frequency rtms. *Brain Stimulation*, 5(3):330–336, 2012.

[6] Michela Balconi, Elisabetta Grippa, and Maria Elide Vanutelli. What hemodynamic (fnirs), electrophysiological (eeg) and autonomic integrated measures can tell us about emotional processing. *Brain and Cognition*, 95:67 – 76, 2015.

[7] Michela Balconi, Elisabetta Grippa, and Maria Elide Vanutelli. What hemodynamic (fnirs), electrophysiological (eeg) and autonomic integrated measures can tell us about emotional processing. *Brain and cognition*, 95:67–76, 2015.

[8] Jackson Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, 91(2):276, 1982.

[9] Katie Bessiere, Irina Ceaparu, Jonathan Lazar, John Robinson, and Ben Shneiderman. Social and psychological influences on computer user frustration. *Media access: Social and psychological dimensions of new technology use*, pages 169–192, 2004.

[10] George E Billman. Heart rate variability–a historical perspective. *Frontiers in physiology*, 2, 2011.

[11] Isabelle Blanchette and Anne Richards. The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning. *Cognition & Emotion*, 24(4):561–595, 2010.

[12] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

[13] Elvira Brattico, Vinoo Alluri, Brigitte Bogert, Thomas Jacobsen, Nuutti Vartiainen, Sirke Nieminen, and Mari Tervaniemi. A functional mri study of happy and sad emotions in music with and without lyrics. *Frontiers in psychology*, 2, 2011.

[14] Todd S Braver, Jonathan D Cohen, Leigh E Nystrom, John Jonides, Edward E Smith, and Douglas C Noll. A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage*, 5(1):49–62, 1997.

[15] D Broadbent. Perception and communications. 1958.

[16] D Broadbent. Decision and stress. 1972.

[17] Scott D Brown. Common ground for behavioural and neuroimaging research. *Australian journal of psychology*, 64(1):4–10, 2012.

[18] Scott C Bunce, Kurtulus Izzetoglu, Hasan Ayaz, Patricia Shewokis, Meltem Izzetoglu, Kambiz Pourrezaei, and Banu Onaral. Implementation of fnirs for monitoring levels of expertise and mental workload. In *Foundations of augmented cognition. Directing the future of adaptive systems*, pages 13–22. Springer, 2011.

[19] James C Byers, Alvah C Bittner, Susan G Hill, Allen L Zaklad, and Richard E Christ. Workload assessment of a remotely piloted vehicle (rpv) system. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 32, pages 1145–1149. SAGE Publications, 1988.

[20] Brad Cain. A review of the mental workload literature. Technical report, DTIC Document, 2007.

[21] Daniel Casasanto. Embodiment of abstract concepts: good and bad in right-and left-handers. *Journal of Experimental Psychology: General*, 138(3):351, 2009.

[22] Ann Chadwick-Dias, Michelle McNulty, and Tom Tullis. Web usability and age: how design changes can improve performance. In *ACM SIGCAPH Computers and the Physically Handicapped*, number 73-74, pages 30–37. ACM, 2003.

[23] George E Cooper and Robert P Harper Jr. The use of pilot rating in the evaluation of aircraft handling qualities. Technical report, DTIC Document, 1969.

[24] M Cope and David T Delpy. System for long-term measurement of cerebral blood and tissue oxygenation on newborn infants by near infra-red transillumination. *Medical and Biological Engineering and Computing*, 26(3):289–294, 1988.

[25] Kelly P Cosgrove, Carolyn M Mazure, and Julie K Staley. Evolving knowledge of sex differences in brain structure, function, and chemistry. *Biological psychiatry*, 62(8):847–855, 2007.

[26] Xu Cui, Signe Bray, Daniel M Bryant, Gary H Glover, and Allan L Reiss. A quantitative comparison of nirs and fmri across multiple cognitive tasks. *Neuroimage*, 54(4):2808–2821, 2011.

[27] Xu Cui, Signe Bray, and Allan L Reiss. Functional near infrared spectroscopy (nirs) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *Neuroimage*, 49(4):3039–3046, 2010.

[28] Antonio R Damasio, BJ Everitt, and D Bishop. The somatic marker hypothesis and the possible functions of the prefrontal cortex [and discussion]. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 351(1346):1413–1420, 1996.

[29] Richard J Davidson. Emotion and affective style: Hemispheric substrates. *Psychological science*, 3(1):39–43, 1992.

[30] Richard J Davidson. Anxiety and affective style: role of prefrontal cortex and amygdala. *Biological psychiatry*, 51(1):68–80, 2002.

[31] David T Delpy, Mark Cope, Pieter van der Zee, SR Arridge, Susan Wray, and JS Wyatt. Estimation of optical pathlength through tissue from direct time of flight measurement. *Physics in medicine and biology*, 33(12):1433, 1988.

[32] Arne Dietrich. Neurocognitive mechanisms underlying the experience of flow. *Consciousness and Cognition*, 13(4):746–761, 2004.

[33] Gautier Durantin, J-F Gagnon, Sébastien Tremblay, and Frédéric Dehais. Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural brain research*, 259:16–23, 2014.

[34] James A Easterbrook. The effect of emotion on cue utilization and the organization of behavior. *Psychological review*, 66(3):183, 1959.

[35] Charles W Eriksen and James D St James. Visual attention within and around the field of focal attention: A zoom lens model. *Perception & psychophysics*, 40(4):225–240, 1986.

[36] Lisa Feldman Barrett and James A Russell. Independence and bipolarity in the structure of current affect. *Journal of personality and social psychology*, 74(4):967, 1998.

[37] Susan Folkman, Richard S Lazarus, Christine Dunkel-Schetter, Anita DeLongis, and Rand J Gruen. Dynamics of a stressful encounter: cognitive appraisal, coping, and encounter outcomes. *Journal of personality and social psychology*, 50(5):992, 1986.

[38] Elaine Fox. *Emotion science cognitive and neuroscientific approaches to understanding human emotions*. Palgrave Macmillan, 2008.

[39] Eija Haapalainen, SeungJun Kim, Jodi F Forlizzi, and Anind K Dey. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 301–310. ACM, 2010.

[40] Eddie Harmon-Jones, Philip A Gable, and Tom F Price. Toward an understanding of the influence of affective states on attentional tuning: Comment on friedman and förster (2010). 2011.

[41] Angela R Harrivel, Daniel H Weissman, Douglas C Noll, and Scott J Peltier. Monitoring attentional state with fnirs. *Frontiers in human neuroscience*, 7, 2013.

[42] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage Publications, 2006.

[43] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[44] Christina Hirth, Hellmuth Obrig, Kersten Villringer, Andeas Thiel, Johannes Bernarding, Werner Mühlnickel, Herta Flor, Ulrich Dirnagl, and Arno Villringer. Non-invasive functional mapping of the human motor cortex using near-infrared spectroscopy. *Neuroreport*, 7(12):1977–1981, 1996.

[45] Justin G Hollands and Christopher D Wickens. Engineering psychology and human performance. *Journal of surgical oncology*, 1999.

[46] TJ Huppert, RD Hoge, SG Diamond, Maria Angela Franceschini, and David A Boas. A temporal comparison of bold, asl, and nirs hemodynamic responses to motor stimuli in adult humans. *Neuroimage*, 29(2):368–382, 2006.

[47] Carroll E Izard. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, 2(3):260–280, 2007.

[48] Kurtulus Izzetoglu, Scott Bunce, Banu Onaral, Kambiz Pourrezaei, and Britton Chance. Functional optical brain imaging using near-infrared during cognitive tasks. *International Journal of human-computer interaction*, 17(2):211–227, 2004.

[49] Caroline Jarrett and Gerry Gaffney. *Forms that work: Designing Web forms for usability*. Morgan Kaufmann, 2009.

[50] Daniel Kahneman. *Attention and effort*. Citeseer, 1973.

[51] Ryuta Kawashima, Masato Taira, Katsuo Okita, Kentaro Inoue, Nobumoto Tajima, Hajime Yoshida, Takeo Sasaki, Motoaki Sugiura, Job Watanabe, and Hiroshi Fukuda. A functional mri study of simple arithmetica comparison between children and adults. *Cognitive Brain Research*, 18(3):227–233, 2004.

[52] Feng Kong. Space–valence associations depend on handedness: evidence from a bimanual output task. *Psychological research*, 77(6):773–779, 2013.

[53] Carol L Krumhansl. An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, 51(4):336, 1997.

[54] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, pages 39–58, 1997.

[55] Jose Leon-Carrion, Jesús Damas, Kurtulus Izzetoglu, Kambiz Pourrezai, Juan Francisco Martín-Rodríguez, Juan Manuel Barroso y Martin, and Maria Rosario Dominguez-Morales. Differential time course and intensity of pfc activation for men and women in response to emotional stimuli: a functional near-infrared spectroscopy (fnirs) study. *Neuroscience letters*, 403(1):90–95, 2006.

[56] Jose León-Carrión, Juan Francisco Martín-Rodríguez, Jesús Damas-López, Kambiz Pourrezai, Kurtulus Izzetoglu, Juan Manuel Barroso y Martin, and María Rosario Domínguez-Morales. A lasting post-stimulus activation on dorsolateral prefrontal cortex is produced when processing valence and arousal in visual affective stimuli. *Neuroscience letters*, 422(3):147–152, 2007.

[57] Lorna Lines, Oluchi Ikechi, K Hone, and Tony Elliman. Online form design for older adults: Introducing web-automated personalisation. In *Proceedings of HCI, the Web and the Older Population, workshop at HCI 2006*, 2006.

[58] Luca Longo, Fabio Rusconi, Lucia Noce, and Stephen Barrett. The importance of human mental workload in web design. In *WEBIST*, pages 403–409, 2012.

[59] Beepa Pty Ltd. Fraps game capture video recorder fps viewer. `http://www.fraps.com/`, 2004 (accessed July 13, 2015).

[60] H Maior, Matthew Pike, Sarah Sharples, and Max L Wilson. Examining the reliability of using fnirs in realistic hci settings for spatial and verbal tasks. *Proceedings of CHI*, 15:3807–3816, 2015.

[61] Horia A Maior, Matthew Pike, Max L Wilson, and Sarah Sharples. Continuous detection of workload overload: An fnirs approach. *Contemporary Ergonomics and Human Factors 2014: Proceedings of the international conference on Ergonomics & Human Factors 2014, Southampton, UK, 7-10 April 2014*, page 450, 2014.

[62] Valerie Mendoza and David G Novick. Usability over time. In *Proceedings of the 23rd annual international conference on Design of communication: documenting & designing for pervasive information*, pages 151–158. ACM, 2005.

[63] George A Miller. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[64] Jeanne Nakamura and Mihaly Csikszentmihalyi. The concept of flow. *Handbook of positive psychology*, pages 89–105, 2002.

[65] Noman Naseer, Melissa Jiyoun Hong, and Keum-Shik Hong. Online binary decision decoding using functional near-infrared spectroscopy for the development of brain–computer interface. *Experimental brain research*, 232(2):555–564, 2014.

[66] U Neisser. *Cognitive psychology*. Prentice-hall, 1967.

[67] Jakob Nielsen and Jonathan Levy. Measuring usability: preference vs. performance. *Communications of the ACM*, 37(4):66–75, 1994.

[68] Jakob Nielsen and Rolf Molich. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 249–256. ACM, 1990.

[69] Don Norman. Emotion & design: attractive things work better. *interactions*, 9(4):36–42, 2002.

[70] Donald A Norman and Daniel G Bobrow. On data-limited and resource-limited processes. *Cognitive psychology*, 7(1):44–64, 1975.

[71] Donald A Norman and Stephen W Draper. User centered system design. *Hillsdale, NJ*, 1986.

[72] R Parasuraman and D Caggiano. Neural and genetic assays of human mental workload. *Quantifying human information processing*, pages 123–149, 2005.

[73] Edward J Peacock and Paul TP Wong. The stress appraisal measure (sam): A multidimensional approach to cognitive appraisal. *Stress Medicine*, 6(3):227–236, 1990.

[74] Evan M M Peck, Beste F Yuksel, Alvitta Ottley, Robert JK Jacob, and Remco Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 473–482. ACM, 2013.

[75] K Luan Phan, Tor Wager, Stephan F Taylor, and Israel Liberzon. Functional neuroanatomy of emotion: a meta-analysis of emotion activation studies in pet and fmri. *Neuroimage*, 16(2):331–348, 2002.

[76] Matthew Pike, Horia A Maior, Martin Porcheron, Sarah Sharples, and Max L Wilson. Measuring the effect of think aloud protocols on workload using fnirs. In *ACMCHI*, 2014.

[77] Michael M Plichta, Antje BM Gerdes, GW Alpers, W Harnisch, S Brill, MJ Wieser, and Andreas J Fallgatter. Auditory cortex activation is modulated by emotion: a functional near-infrared spectroscopy (fnirs) study. *Neuroimage*, 55(3):1200–1207, 2011.

[78] MM Plichta, MJ Herrmann, CG Baehne, A-C Ehlis, MM Richter, P Pauli, and AJ Fallgatter. Event-related functional near-infrared spectroscopy (fnirs): are the measurements reliable? *Neuroimage*, 31(1):116–124, 2006.

[79] Gary B Reid and Thomas E Nygren. The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in psychology*, 52:185–218, 1988.

[80] Yvonne Rogers, Helen Sharp, Jenny Preece, and Michele Tepper. Interaction design: beyond human-computer interaction. *netWorker: The Craft of Network Computing*, 11(4):34, 2007.

[81] Edmund T Rolls. The orbitofrontal cortex and reward. *Cerebral cortex*, 10(3):284–294, 2000.

[82] James C Root, Philip S Wong, and Marcel Kinsbourne. Left hemisphere specialization for response to positive emotional expressions: A divided output methodology. *Emotion*, 6(3):473, 2006.

[83] William B Rouse, Sharon L Edwards, and John M Hammer. Modeling the dynamics of mental workload and human performance in complex systems. *IEEE Transactions on Systems, Man, & Cybernetics*, 1993.

[84] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145, 2003.

[85] Sergio Sayago and Josep Blat. Some aspects of designing accessible online forms for the young elderly. In *WEBIST (2)*, pages 13–17, 2007.

[86] Sergio Sayago, José-María Guijarro, and Josep Blat. Selective attention in web forms: an exploratory case study with older people. *Behaviour & Information Technology*, 31(2):171–184, 2012.

[87] Felix Scholkmann, Stefan Kleiser, Andreas Jaakko Metz, Raphael Zimmermann, Juan Mata Pavia, Ursula Wolf, and Martin Wolf. A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. *Neuroimage*, 85:6–27, 2014.

[88] Tim Shallice. *From neuropsychology to mental structure*. Cambridge University Press, 1988.

[89] Ben Shneiderman. *Designing the user interface: strategies for effective human-computer interaction*, volume 3. Addison-Wesley Reading, MA, 1992.

[90] Craig A Smith and Leslie D Kirby. Putting appraisal in context: Toward a relational model of appraisal and emotion. *Cognition and Emotion*, 23(7):1352–1372, 2009.

[91] Edward E Smith and John Jonides. Working memory: A view from neuroimaging. *Cognitive psychology*, 33(1):5–42, 1997.

[92] Erin Treacy Solovey, Audrey Girouard, Krysta Chauncey, Leanne M Hirshfield, Angelo Sassaroli, Feng Zheng, Sergio Fantini, and Robert JK Jacob. Using fnirs brain sensing in realistic hci settings: experiments and guidelines. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*, pages 157–166. ACM, 2009.

[93] Paul M Thompson, Tyrone D Cannon, Katherine L Narr, Theo Van Erp, Veli-Pekka Poutanen, Matti Huttunen, Jouko Lönnqvist, Carl-Gustaf Standertskjöld-Nordenstam, Jaakko Kaprio, Mohammad Khaledy, et al. Genetic influences on brain structure. *Nature neuroscience*, 4(12):1253–1258, 2001.

[94] Pamela S Tsang and Velma L Velazquez. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, 39(3):358–381, 1996.

[95] Robert A Virzi. Refining the test phase of usability evaluation: how many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 34(4):457–468, 1992.

[96] David Watson, Lee A Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[97] Christopher D Wickens. Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, 3(2):159–177, 2002.

[98] Christopher D Wickens. Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3):449–455, 2008.

[99] christopher D Wickens and J M Flach. Information processing. *Human factors in aviation*, pages 110–156, 1988.

[100] Walter W Wierwille and John G Casali. A validated rating scale for global mental workload measurement applications. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 27, pages 129–133. Sage Publications, 1983.

[101] John R Wilson and Sarah Sharples. *Evaluation of human work*. CRC Press, 2015.

[102] Luke Wroblewski. *Web form design: filling in the blanks*. Rosenfeld Media, 2008.

[103] Erik Wstlund, Torsten Norlander, and Trevor Archer. The effect of page layout on mental workload: A dual-task experiment. *Computers in Human Behavior*, 24(3):1229 – 1245, 2008. Instructional Support for Enhancing Students' Information Problem Solving Ability.

[104] Tal Yarkoni, Russell A Poldrack, Thomas E Nichols, David C Van Essen, and Tor D Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature methods*, 8(8):665–670, 2011.

[105] Kazuki Yoshida, Daisuke Sawamura, Yuji Inagaki, Keita Ogawa, Katsunori Ikoma, and Shinya Sakai. Brain activity during the flow experience: A functional near-infrared spectroscopy study. *Neuroscience letters*, 573:30–34, 2014.

Appendix