The effect of web form design on workload

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

Leave Authors Anonymous
for Submission
City, Country
e-mail address

ABSTRACT

In this dissertation we evaluated a web 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 an online insurance claim process. Three variations of the web forms were created: one standard insurance claim form, one which has alternated order of form fields, and one which divides the forms into three pages. We hypothesised that there should be significant difference between them in the objective and subjective measures of mental workload. Also, we tried to elicit emotional state from the fNIRS data and hypothesised that a correlation will be observed between the objective and subjective measures of emotional valence. We found that the control condition elicited the lowest mental workload and positive valence according to the majority of the corresponding measures. In contrast, the divided page approach evoked highest mental workload and positive valence. We contribute to web form design and usability research by proposing implications for design. In addition, we assessed the practicality of fNIRS as limited particularly for low engaging web interfaces due to the higher amount of time, and resources required to run and analyse the data.

ACM Classification Keywords

H.5.2. [Information Interfaces and Presentation] User Interfaces. - Graphical user interfaces

Author Keywords

Mental workload; web forms; usability; functional near-infrared spectroscopy; fNIRS;

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[11, 15] 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[12] 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[10, 16, 14] 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[3] and emotion processing[6]. 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[13] 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.

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.

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?

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.

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.

EXPERIMENT DESIGN

We used fNIRS because it is non-invasive, have low sensitivity to motion artefacts, and with high temporal resolution, thus suitable for user trials. We combine the fNIRS, as an objective measure of mental workload with the subjective NASA-TLX as they are validated and reliable methods. In addition, we decided to measure the emotional valence with the SAM questionnaire. This way we gather comprehensive information about how participants perceive the interfaces and relate it to the psychophysiological data. Furthermore, fNIRS is novel method for conducting usability studies, and as such we aim to assess its practicality.

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 Âč10 Amazon voucher.

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

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 as it can be seen from Figure 1. 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.

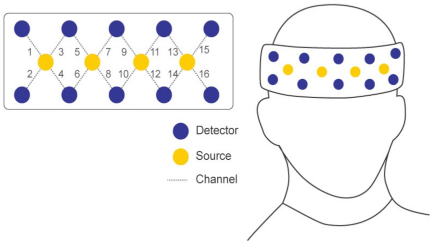

Figure 1. The spatial arrangement of the source-detector pars placed on participants forehead.

Materials

NASA-TLX

The NASA-TLX[8] 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[7], 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[2] 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.

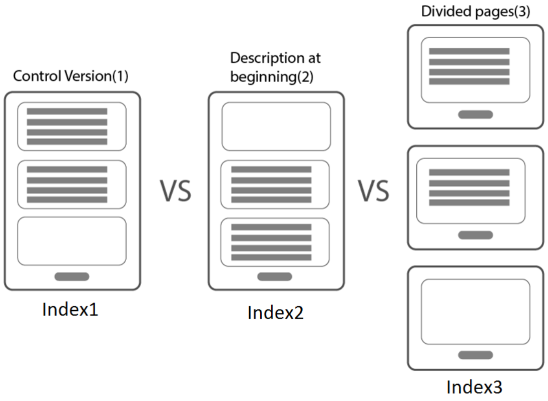

Figure 2. A sketch of the 3 layout variations of the web forms. The control condition index1 asked first for personal information, then accident information, and finally asks for description of the accident. With Index2 the description field was placed at the beginning. Index3 had 3 sections separated on 3 sequential pages.

Layout variations of web forms

A total of 3 HTML/CSS variations of a standard web form for insurance claiming were produced. A generalized sketch is depicted in Figure 2. 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 the attached zip file, that contains all the data.

Video capture

A video capture of the computer screen was recorded during the experiment using Fraps[9]. 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 the attached zip file.

Design

The study used repeated measures within subjects design. The depended variable was the mental workload, 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.

Procedure

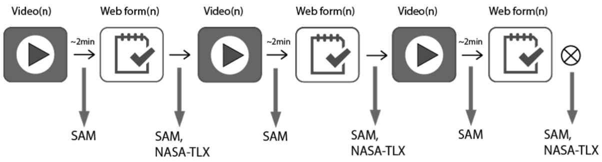

Figure 3. The image illustrates, the study procedure followed in this experiment. First participants, watched video, then fill the SAM measure. After that, they filled the a web form, and then filled both the NASA-TLX and SAM scales. The process was repeated 3 times.

Participants followed the procedure illustrated by Figure3. 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.

Data Analysis

The fNIRS data was analysed 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[4], in order to calculate oxygenated, and deoxygenated hemoglobin values. Finally, to remove motion artefacts, the correlation based signal improvement(CBSI)[5] 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.

RESULTS

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 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.

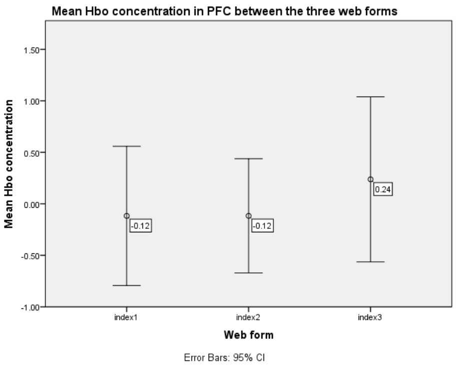

Figure 4. Mean Hbo activation between the three web form conditions as measured by fNIRS. Higher Hbo values indicates higher mental workload experienced by the operator.

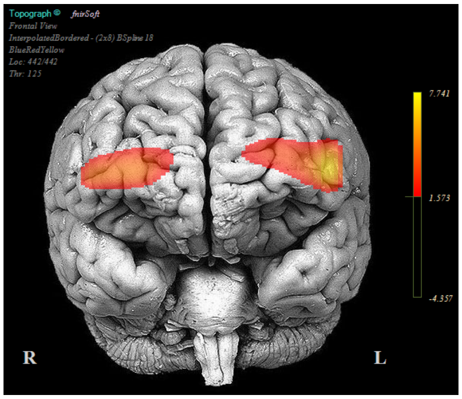

Figure 5. Brain view of Participant's 2(P2) Hbo activation during web form filling of index3. It can be noted more hemodynamic activation in the left hemisphere. P2 rated its emotional valence as positive (4/5).

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

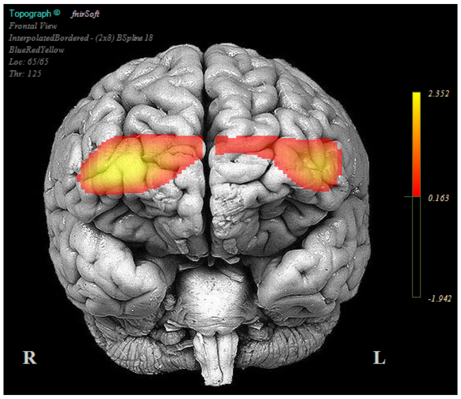

Figure 6. Brain view of Participant 2 while viewing video of road accident(video3). More arousal can be noted and more right hemisphere activation, which is expected as automotive accidents increase arrousal and emotional state becomes negative.

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 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.

Table 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 =
Physical demand	4.05 (SD = 4.08)	2.90 (SD = 3.21)	3.90 (SD =
Temporal demand	7.40 (SD = 4.49)	7.65 (SD = 5.79)	6.55 (SD =
Performance	6.65 (SD = 3.79)	5.60 (SD = 3.62)	6.20 (SD =
Effort	8.15 (SD = 4.58)	7.35 (SD = 4.68)	8.20 (SD =
Frustration	6.10 (SD = 5.11)	6.00 (SD = 3.66)	6.75 (SD =
Total	6.92 (SD = 2.95)	6.47 (SD = 3.11)	7.07 (SD =

SAM - arousal scale

As it can be seen from Figure 7 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 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.

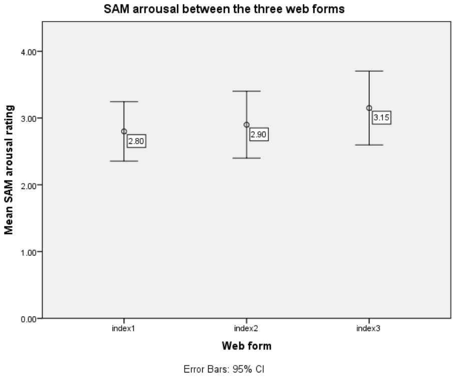

Figure 7. The mean SAM arrousal rating obtained from the three web form conditions.

Emotional Valence

fNIRS differences

Data from the period of web form filling

To begin with, Figure 8 shows the valence differences in Hbo concentration 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

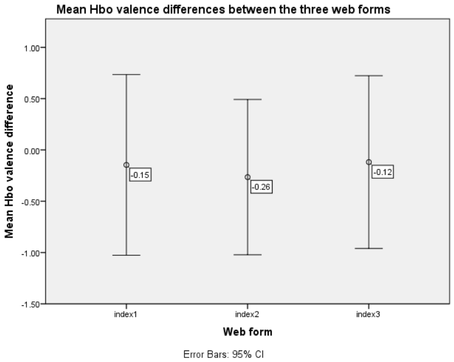

Figure 8. The mean Hbo valence differences obtained from the fNIRS for the three web form conditions. Higher values indicate more positive affect, and lower value more negative affect.

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 9. A one-way repeated measures 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.

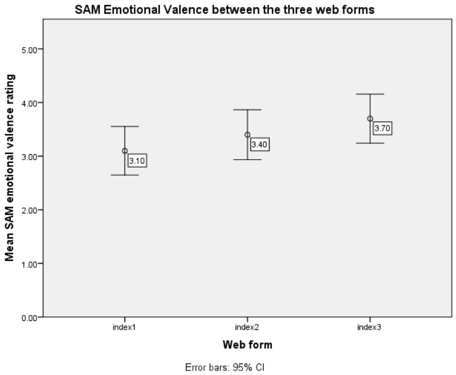

Figure 9. Perceived emotional valence between the 3 web form conditions

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.

User performance and preferences

Task completion time

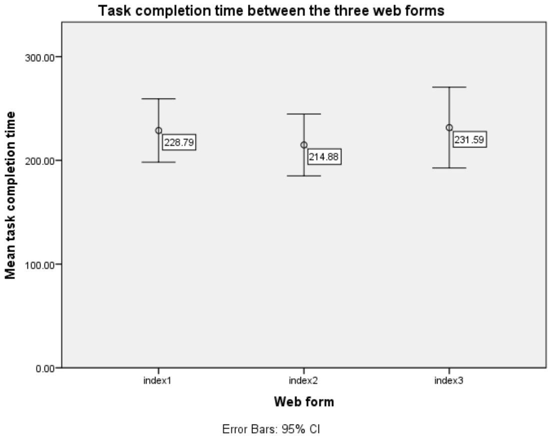

Figure 10. Mean task completion time between the three web forms

It can be seen from Figure 10 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 which is depicted in Figure 11. The bulk of them preferred index3 and index1 with 10 and 9 votes respectively compared to index2 which was only preferred by 3 participants. user-prefe The given answer were transcribed as can be seen in Appendix **??**.

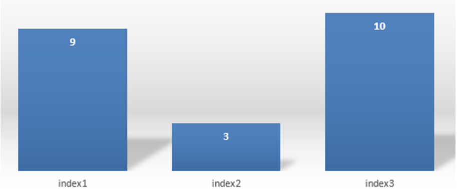

Figure 11. At the end of the experiements participants were asked which of the 3 web forms they preferred the most. Th sum of the amswers is depicted in a bar chart.

CONCLUSIONS

REFERENCES

1. Hasan Ayaz. 2010. Functional Near Infrared Spectroscopy based Brain Computer Interface. *PhD Thesis, Drexel University, Philadelphia, PA* (2010).

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

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

4. M Cope and David T Delpy. 1988. 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 (1988), 289–294.

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

6. Antonio R Damasio, BJ Everitt, and D Bishop. 1996. 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 (1996), 1413–1420.

7. Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage Publications, 904–908.

8. Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52 (1988), 139–183.

9. Beepa Pty Ltd. 2004 (accessed July 13, 2015). FRAPS game capture video recorder fps viewer. http://www.fraps.com/. (2004 (accessed July 13, 2015)).

10. H Maior, Matthew Pike, Sarah Sharples, and Max L Wilson. 2015. Examining the reliability of using fNIRS in realistic hci settings for spatial and verbal tasks. *Proceedings of CHI* 15 (2015), 3807–3816.

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

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

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

14. Matthew Pike, Horia A Maior, Martin Porcheron, Sarah Sharples, and Max L Wilson. 2014. Measuring the Effect of Think Aloud Protocols on Workload using fNIRS. In *ACMCHI*.

15. Ben Shneiderman. 1992. *Designing the user interface: strategies for effective human-computer interaction*. Vol. 3. Addison-Wesley Reading, MA. 60–63 pages.

16. Erin Treacy Solovey, Audrey Girouard, Krysta Chauncey, Leanne M Hirshfield, Angelo Sassaroli, Feng Zheng, Sergio Fantini, and Robert JK Jacob. 2009. 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*. ACM, 157–166.