Measuring the Effect of Think Aloud Protocols on Workload using fNIRS

Matthew Pike[*], Horia Maior[*†], Martin Porcheron[†], Sarah Sharples[‡] and Max L. Wilson[*]

[*]Mixed Reality Lab, University of Nottingham, Nottingham, UK
[†]Horizon Doctoral Training Centre, University of Nottingham, Nottingham, UK
[‡]Human Factors Research Group, University of Nottingham, Nottingham, UK

{psxmp8, psxhama, psxmp9, sarah.sharples, max.wilson}@nottingham.ac.uk

ABSTRACT

The Think Aloud Protocol (TAP) is a verbalisation technique widely employed in HCI user studies to give insight into user experience, yet little work has explored the impact that TAPs have on participants during user studies. This paper utilises a brain sensing technique, fNIRS, to observe the effect that TAPs have on participants. Functional Near-Infrared Spectroscopy (fNIRS) is a brain sensing technology that offers the potential to provide continuous, detailed insight into brain activity, enabling an objective view of cognitive processes during complex tasks. Participants were asked to perform a mathematical task under 4 conditions: nonsense verbalisations, passive concurrent think aloud protocol, invasive concurrent think aloud protocol, and a baseline of silence. Subjective ratings and performance measures were collected during the study. Our results provide a novel view into the effect that different forms of verbalisation have on workload during tasks. Further, the results provide a means for estimating the effect of spoken artefacts when measuring workload, which is another step towards our goal of proactively involving fNIRS analysis in ecologically valid user studies.

ACM Classification Keywords

H5.2 [Information interfaces and presentation]: User Interfaces. - Evaluation/methodology.

Author Keywords

functional near-infrared spectroscopy, fNIRS, BCI, human cognition, Think Aloud Protocol, HCI

INTRODUCTION

The Think-Aloud Protocol (TAP) is a widely used research method [23], utilised in a variety of research fields including Human Computer Interaction. Since TAP will use resources from verbal working memory, it is fair to assume that the inclusion of spoken protocols will potentially affect cognitive processes due to use of available resources. Consequently, TAPs may affect performance in tasks, and also measures of workload during studies.

To analyse the potential impact that a TAP may have on an individual, we use a direct measure through the brain monitoring technology Functional Near-InfraRed Spectroscopy (fNIRS). fNIRS has received recent focus in HCI research for its amenability for more ecologically valid study conditions [17, 33]. While some brain sensing techniques like functional Magnetic Resonance Imaging (fMRI) require minimal or no movement from users, fNIRS can be used while seated naturally at a computer [33]. Further, because fNIRS measures blood oxygenation and deoxygenation rather than electrical signals like Electroencephalography (EEG), fNIRS permits more natural movements associated with using a computer without being subject to significant artefacts in the data. Although the suggestion is that fNIRS can be used more easily within natural, ecologically valid user study conditions, current research is still limited to performing controlled simple Working Memory tasks (e.g. [33, 30]).

In the context of HCI, TAP is typically used as an evaluation method to elicit insights into participants thoughts and strategies during usability and user studies. TAP, however, has also been used in other settings, such as cognitive psychology and social sciences [7], to understand phenomena such as user mental models, expertise, and problem solving. As well as being a core part of user studies, verbalisations are also closely related with Working Memory, as both the interpretation of words in the task and the integration of thoughts involve the phonological loop [38]. Consequently, to integrate fNIRS measurement within a typical user study that might involve a TAP, we have to be aware of how one will affect the other. There are various forms of TAP, including retrospective, which occurs after a task has been completed, and concurrent, which occurs during a task. Of concurrent forms of TAPs, there is both invasive, which involves directly questioning participants, and passive, which simply encourages participants to maintain verbalisations about their thoughts and actions. Because fNIRS measurements are taken during tasks, this paper focuses on concurrent TAPs.

In the following sections, we first review related work on TAPs, Working Memory and mental workload, fNIRS

sensing and other technologies. The paper continues by describing a user study examining the impact of a) nonsense verbalisations, b) passive concurrent think aloud, and c) invasive concurrent think aloud, compared to a baseline of silent non-verbal working memory. We then present the results of the study, discuss the findings in terms of what we can learn about the impact of TAP on mental workload in general, and recommendations for using fNIRS measurements in an HCI user study.

RELATED WORK

This section presents three key areas of related work: 1) TAP and their effect on Cognition, 2) Working Memory (WM) models and the Prefrontal Cortex (PFC), and 3) Brain sensing techniques in HCI including fNIRS.

Think-Aloud Protocols

Ericsson and Simon's seminal work on verbal reporting of participants thought process is the most cited amongst Think Aloud Protocols [29]. Prior to this work, little consideration was made to the type of verbalisation produced by participants under study conditions [16]. In their original discussion of TAP, Ericsson and Simon [11] distinguish between 3 distinct levels at which verbalisations occur. Levels 1 and 2 are described as being valid representations of a participant's mental state, since they are verbalising information stored in short term memory and are representative of the participant's current state. Level 3 requires access to long term memory and influences what would otherwise be their typical state. Ericsson and Simon's version (Levels 1 and 2) of the protocol is strictly non-intrusive, and researchers implementing the protocol are restricted to simply using the verbal prompt -"Keep talking"- to avoid influencing the participant, and ensuring that the reported data relates solely to the task. To distinguish between other forms, we refer to this type of TAP as Passive (PTAP) for the remainder of this paper.

In practice, however, it has been shown that many researchers incorrectly implement or misreport the TAP they are using [23]. Many practitioners of TAP prefer to question participants at level 3 to obtain coherent, actionable utterances relating to the system under evaluation, rather than inferring results from level 1 and 2 utterances. Researchers have attempted to formalise this level of questioning [9, 16]. We characterise these approaches under the umbrella term Invasive TAP (ITAP). With ITAP, researchers are free to probe the user's mental model, but Ericsson and Simon would disregard the findings at these levels stating that they have been influenced. Under ITAP, a practitioner is able to prompt the participant with more probing questions - "Why did you do X?".

Working Memory

In an attempt to characterise and model the cognitive processes involved when a participant is partaking in a TAP, we draw on research into Working Memory (WM)[5], a specific system in the brain which "provides temporary storage and manipulation of information" [2]. WM [5, 3, 4] processes information in two forms: verbal and spatial, and has four main components (Figure 1): a visuo-spatial sketch pad holding information in an analogue spatial form (e.g. colours, shapes, maps, etc.), a phonological loop holding verbal information in an acoustical form (e.g. numbers, words, etc.), an episodic buffer dedicated to linking verbal and spatial information in chronological order and finally, a central executive acting as supervisory system and controlling the information from and to its "slave systems".

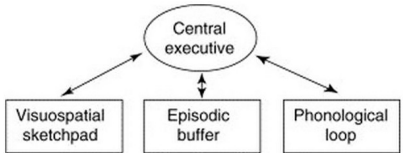

Figure 1. Baddeley's Working Memory Model

Additionally, Baddeley describes the concept of Long-term memory (LTM), which represents a different storage location to working memory. LTM is unlimited in space and is responsible for storing information that is no longer in working memory.

We can relate a number of concepts described by Ericsson and Simon to the working memory model described by Baddeley. For example, Ericsson and Simon note that verbalisations at level 1 and 2 occur within short term memory. We can further characterise this with Ericsson and Simon stating that TAP will utilise the Phonological loop as it is verbal in nature. Tasks under which the TAP is performed may also interact with other components of the working memory model. Tasks involving imagery or mental rotation, for example, will utilise the visuo-spatial sketchpad since they are spatial, whereas verbalising occurs in the phonological loop. For such tasks under TAP conditions the two concepts of the model will be activated, with the central executive mediating information flow between the two. The episodic buffer may also have a role under ITAP conditions, since the protocol will require access to memories that are not in the immediate short term memory. We would not expect the Episodic buffer to be utilised in the PTAP condition.

In addition to the WM model, we can also consider the Information Processing Model [39] and Multiple Resource Model [38] proposed by Wickens. Wickens describes that necessary resources are limited and aims to illustrate how elements of the human information processing system such as attention, perception, memory, decision making and response selection interconnect. We are interested in observing how and when these elements interconnect under TAPs. Wickens describes three different 'stages' (see STAGES dimension in Figure 2) at which information is transformed: a perception stage, a processing or cognition stage, and a response stage. The first stage involves perceiving information that is gathered by our senses and provide meaning and interpretation of what is being sensed. The second stage represents the step

where we manipulate and "think about" the perceived information. This part of the information processing system takes place in WM and consists of a wide variety of mental activities. We can observe that TAP will likely affect each of these elements since the protocol introduces additional sensory inputs, which require potential comprehension and will sometimes require a response (specifically under ITAP).

Wickens also proposes the Multiple Resource Model [38], illustrated in Figure 2. The elements of this model overlap with the needs and considerations of evaluating complex tasks which could be analysed and affected by the inclusion of a TAP. He describes the aspects of cognition and the multiple resource theory in four dimensions: STAGES, MODALITIES, CODES and the VISUAL PROCESSING (see Figure 2).

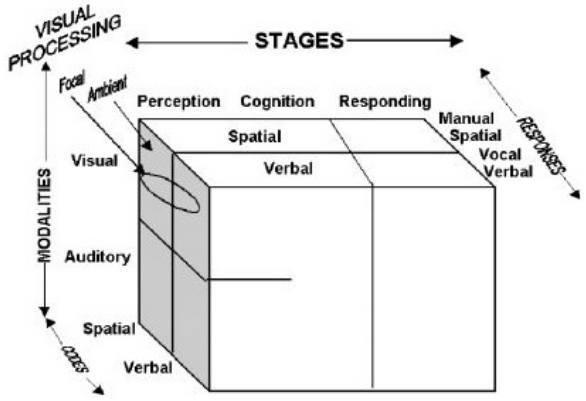

Figure 2. The 4-D multiple resource model, by Wickens

One of the key roles of the Multiple Resource Model is to demonstrate the hypothesised independence of modalities and use this to design tasks. We know for example that the inclusion of TAP will introduce additional Auditory resource requirements, since the participant will hear their own verbalisations. This in turn will require additional Perception from the participant and will draw on their Verbal coding resources and Vocal Verbal responses.

The Prefrontal Cortex (PFC) is the anterior part of the frontal lobes of the brain and is considered central to the function of WM, dealing with executive and attention processes [19]. Miller and Cohen defines the representations in the PFC as "attentional templates, retrieval cues, rules, or goals" [27], and many researchers agree that PFC function is one of Baddley's executive control [1]. Conversely, Rushworth reports that not all PFC subregions are essential for working memory [32]. The PFC is the region of the brain that we targeted during this study since there is significant evidence to support its role in WM [6, 8]. In addition to the PFC, Brocas area is located within the frontal lobe and is linked with speech production [12].

Brain Sensing Techniques

There are several brain sensing technologies available for research, including (but not limited to) fMRI, EEG, and fNIRS, which are summarised in Table 1. Each of these technologies have different strengths and weaknesses, as discussed by Tan and Nijholt [34, Chapter 1].

Functional Magnetic Resonance Imaging (fMRI) is a functional neuroimaging technique that associates detected changes in blood flow (hemodynamic response) to brain activity. fMRI is typically used for applications requiring high spatial resolution, but requires people to lay very still, and precludes the use of a computer. Participants are unable to interact directly with a system, but can respond to visual stimuli through the use of mirrors. Li et. al [22] for example, used real time fMRI to control the animation speed of a virtual human runner.

Electroencephalography (EEG) typically uses between 16 and 64 sensors on the scalp to detect varying electrical charge within the brain. With the introduction of commercially available bluetooth EEG sensors, like the Emotiv[1], EEG has become an affordable option for brain sensing [10]. For evaluation, however, EEG data is susceptible to motion artefacts, and so producing averages for periods of interaction provides limited insight. Pike et al [31] proposed, that EEG data was most valuable when combined visually with recorded TAP data, as statements of confusion, or pauses in verbalising ones actions, coincided with and were qualified by EEG data.

fNIRS uses blood oxygenation, rather than electrical levels, for determining the activation of areas in the brain, where more blood flow indicates higher activity. Recent research has shown that because blood-flow in the brain is less affected by body movement, fNIRS may be a more appropriate brain sensing technology for evaluation [17, 30, 21]. Because it takes several seconds for blood to flow to the brain [37], fNIRs has been largely discounted for real-time interaction with systems.

EXPERIMENT DESIGN

The aim of this study was to investigate how verbalisation and TAPs affect cognition and the thought process during user study tasks. We produced three research questions:

- How can we identify the impact of TAPs on human cognition and mental workload using fNIRS?
- What are the most suitable measures to sense such an impact?
- How can we sense the reduction of available resources due to integrating a TAP concurrently with a task?

To answer these research questions, a theoretical understanding of TAPs, human cognition, mental workload and the interconnection between these concepts is required. Wickens' Multiple Resource Model [38] can describe the relationship between the available resources and task demands. When performing a task, a person perceives both their own verbalisation and/or external stimuli as an

[1]http://www.emotiv.com/

Technique	Physical Property	Sensitivity to Motion	Portability	Spatial Resolution	Temporal Resolution	Cost
fMRI	Magnetic	Very High	None	High	Low	Expensive
EEG	Electrical	High	Portable	Low	High	Reasonable
fNIRS	Optical	Low	Portable	High	Low	Moderate

Table 1. Summary of Brain Sensing technologies

Auditory modality. During 'think aloud' we also process information (make decisions, store memories, retrieve memories, etc.), and output them as a response (e.g. as a Vocal Verbal encoding). Therefore, TAPs might have an impact on all three stages (perception, cognition and response) of the Multiple Resource Model. According to the model, a TAP is a verbal/linguistic activity, therefore the codes of its cognition stage is Verbal. Consequently, we chose a task (described further below) that was easy to verbalise and involves continuous use of the phonological loop, such that different verbalisation conditions would interact with the task.

Primarily, we wanted to compare the different concurrent TAPs against a baseline of not verbalising. In order to check whether simply using your voice creates an artefact in the fNIRS data, as opposed to thinking and talking, we also included a second baseline of repeatedly verbalising the word 'blah'. Type of verbalisation, as primary independent variable, created four conditions:

1. Task Only (Baseline - B1)
2. Task + "Blah blah blah" (Baseline - B2)
3. Task + Passive Concurrent TAP (PTAP)
4. Task + Invasive Concurrent TAP (ITAP)

We designed a repeated measures, within-participants study to compare these conditions, where participants solved eight mathematical problems. Conditions and tasks were counterbalanced using a Latin-square design.

Hypotheses

We had a number of hypothesis that we sought to investigate whilst conducting this study relative to performance, cognition, and participants' grouping based on mathematical performance (High and Low performing groups):

HP - There will be a significant difference in performance between verbal conditions.

HC - There will be a significant difference in cognition between verbal conditions.

HP and HC were drawn from Wickens 4D Multiple Resources Model [38]. Both TAP and mathematical tasks should primarily use verbal working memory in the modality, encoding, and processings dimensions. Consequently, the demands imposed by various verbal conditions may affect the total workload element, and workload may then affect performance.

HC.S - There will be a significant difference in cognition between verbal conditions for high performing participants.

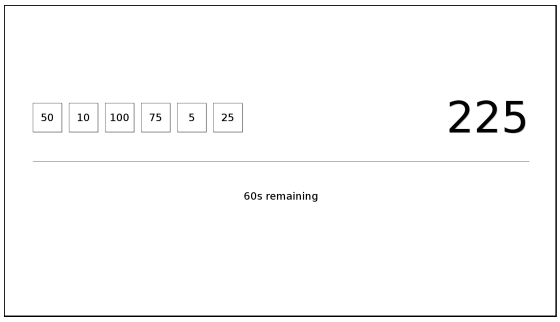

Figure 3. A screenshot of the task

HC.W - There will be a significant difference in cognition between verbal conditions for low performing participants.

Depending on how well participants performed during the four conditions, we distinguished between high performing participants (top half) and low performing participants (bottom half) [26]. These groups were formed to investigate whether TAPs have a different impact on cognition relative to the participants grouping.

Task

In order to determine how TAPs affect the different stages of the Multiple Resources Model, the task had to be chosen carefully such that verbalisation could potentially interrupt the process. The first criterion, therefore, was that the task should primarily use the phonological loop, and thus be a verbally oriented task. Second, the task had to involve continuous use of the phonological loop, and so a simple and discrete memory task was not sufficient. Third, the task had to be verbalisable for the TAPs, which also meant that a memory task was not sufficient. Fourth, the task also had to have various levels of difficulty to enable control over the primary task mental demands; according to the resource vs demands model [25] harder tasks would increase demand and thus reducing participant's resources for engaging in the TAP. Finally, performance on the task had to be measurable in order to determine the effect of verbalisations. Based upon these five criteria, we decided on using a mathematics task. Participants were provided with a set of six numbers and had to get as close as possible to a target final number. This problem is a variation on what is commonly known as the countdown problem[2]. Each number may be used only once (although there is no requirement to use every number), and participants have

[2]based on the mathematical challenge presented to contestants of the popular UK TV quiz show "Countdown"

60s to reach as close to the target number as possible, using four operators: addition, subtraction, multiplication and division.

36 versions of the task were created to be used across the four conditions, at various levels of difficulty. To classify their difficulty, one researcher and two independent judges rated the difficulty of each problem. Difficulty was judged in four categories: easy, quite easy, quite hard, and hard. Inter-rater agreeability was confirmed with a Cohen's Kappa test, where the researcher achieved scores of 0.6419 (substantial agreement [20]) with the first independent judge, and 0.8571 (almost perfect agreement) with the second. This agreement was used to ensure that problem difficulty was balanced between conditions.

Participants

Twenty participants (14 male, 6 female) with an average age of 28.55 years were recruited to take part in the study. Participants were recruited from the University of Nottingham, and included a mix of staff members and students from a range of disciplines. All participants had normal or corrected vision and reported no history of head trauma or brain damage. The study was approved by the school's ethics committee. Participants provided informed consent, and were compensated with £15 in gift vouchers.

Procedure

Participants were first introduced to the task that they would be completing during the study. They were given two practice runs of the task (under baseline conditions) to familiarise themselves and reduce the impact of learning in their first condition. Once comfortable with the requirements of the task, participants were fitted with the fNIRS brain imaging device, which was placed upon their forehead targeting the PFC. At this point participants entered the recorded section of the study. During this stage, participant input was captured, verbalisations were recorded via microphone, and brain data was captured on a separate, calibrated machine.

Participants partook in four conditions which were counterbalanced using a latin square rotation. Each condition began with a tutorial and practice session. The tutorial session was used to train the participant on how to verbalise according to the specific TAP being used in the particular condition. The practice session would then serve as an opportunity to trial the technique prior to beginning the test itself and thus reducing the interference on the first task in each condition. Each condition included eight of the tasks described above.

For each of the eight tasks in each condition, participants were given sixty seconds to attempt the problem. All calculations were performed mentally; pen and paper was not provided. After the sixty seconds had elapsed (or if the participant decided to proceed prior to this), participants were prompted to enter the number they had achieved during the calculation period. To avoid participants simply entering the target number, they were prompted to recall their solution. The solutions provided by participants were recorded by the researcher on paper and later digitalised.

After each condition, participants completed a standard NASA TLX form to subjectively rate their mental workload during the task. Each condition concluded with a thirty second rest period where the participants were asked to remain still, relax and empty their mind of thoughts.

The study was conducted in an office-like environment. This was an important consideration as many brain based studies are conducted under strictly controlled lab settings. The office environment provides a more naturalistic and ecologically valid setting.

Measurements and Equipment

We collected various types of data during the study. The data can be categorised into two groups: Performance during the study (P), and Cognition (C).

Task Accuracy - P

We measured task performance according to distance from the target answer for each of the 36 problems across the four conditions. Because the target varied, we used measured distance from the target as a percentage, which was subtracted from 100%. 100% represented the correct answer, 95% as being 5% from the target, and so on. As the results tended towards the target, task accuracy was analysed. To provide incentive to submit actual rather than ideal answers, we also measured whether participants could recall the solution to their answer.

Task Time - P

Task time was measured for each of the 36 problems performed across the four conditions. We note that participants were not encouraged to solve the problem in the shortest possible time, rather, they were asked to get as close possible to the target.

NASA-TLX questionnaire - C

We used the NASA-TLX questionnaire, a subjective workload assessment tool [15], based on the weighted average ratings of six subscales including, in order: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration. Each participant was asked to self rate their mental workload using the NASA-TLX once after each condition. We additionally investigated each of subscales independently.

fNIRS data - C

fNIRS data was recorded using an fNIRS300 device and the associated COBI Studio recording software provided by Biopac Systems inc. The headband shaped device is a sixteen-channel transducer for continuous Near Infrared Spectroscopy (NIRS). The headband consists of four infrared (IR) emitters operating on a range between 700 to 900 nm, and ten IR detectors. The device is placed on the PFC targeting the Brodmann area 10 (BA10). Oxygenated hemoglobin (HbO) and deoxygenated-hemoglobin (Hb) are both strong absorbers

of light, whereas skin, tissue and bone are mostly transparent to NIR light, this property is typically referred to as the *optical window* [18]. The tissue is radiated by the light sources and the detectors receive the light after the interaction with the tissue. See Figure 4 [13] for an illustration of how the headband is positioned, and to visualise the path that the light follows during operation.

Preprocessing was performed to transform raw data from the device into oxygenation values using the Modified Beer-Lambert law (MBLL) [36]. We also applied filtering algorithms to remove high-frequency noise, physiological artefacts such as heartbeats and motion derived artefacts. To perform this preprocessing step we used the Matlab Toolbox, NIRS-SPM [40]. We performed de-trending using a discrete cosine transform with a high frequency cut off of 128 seconds. The baseline was removed, and low pass filtering was performed with a Gaussian filter with a width of 1 second. We also considered the delay induced by the hemodynamic response [36] by omitting the first 10s of the trial when processing the data [30].

The Biopac fNIRS device used in this study provides 16 channels of brain data readings. A channel is defined by the relationship between a source and detector pair as shown in Figure 4. From the MBLL we receive Hb, HbO and TotalHb (Hb + HbO) values for each channel. Measures were synthesised by combining specific channels averages to form a single measurement. Channels 3,4,5,6 were used to represent the left side and channels 11,12,13,14 formed the right side in these measurements. An overall measurement was produced by averaging the data from all 16 channels(see Figure 4).

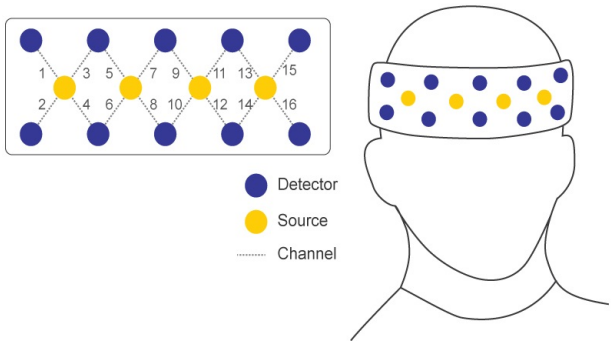

Figure 4. Sensor layout for the Biopac fNIRS used

Experiment Software

When designing the study we placed a strong emphasis on automating the running of the study and collection of the associated data. With the exception of the brain data, all other measures were collected from a single program. We developed this program using PEBL: The Psychology Experiment Building Language [28]. The language provides a convenient set of features including accurate experiment timing and predefined psychology/study procedures such as demographic questionnaires. Of particular relevance to this study was the pre-defined, computerised version of NASA-TLX.

RESULTS

We began by checking for ordering effect. A one way repeated measure ANOVA showed that participants performed significantly slower in the first condition they experienced, while average time to complete the subsequent conditions was even ($F(19,3) = 2.816, p < 0.05$). An LSD post-hoc ANOVA test also showed that average scores also improved between the first condition they experienced and the last ($F(19,3) = 2.271, p < 0.05$).

Performance

Against hypothesis HP, our analysis showed no significant difference in task accuracy between conditions. We found no significant difference in performance between any of the four conditions, however, under the TAP conditions, participant performance slightly improved. There was also no difference in the number of tasks correctly calculated in each condition. We hypothesised that, ITAP under time pressure would cause performance to drop, but instead these results support the findings of McDonald et al. [24] who found that neither form of TAP affected performance.

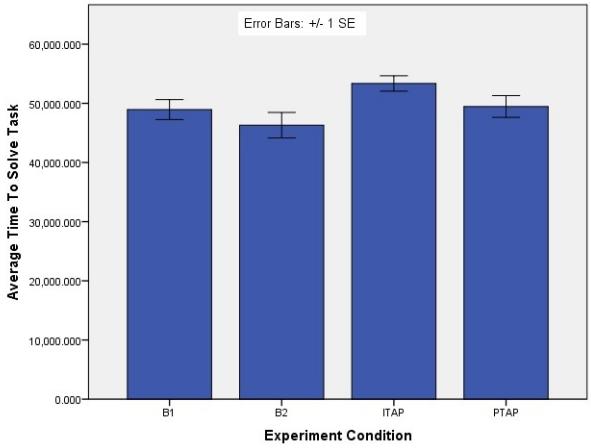

Figure 5. Mean time to solve a set of tasks

A significant difference was found in terms of time to complete tasks (Figure 5). As perhaps expected, participants took significantly longer to solve tasks in the ITAP condition ($F(17,3) = 9.895, p < 0.01$) relative to the other three conditions (B1: $p < 0.005$, B2: $p < 0.001$, PTAP: $p < 0.05$). PTAP was not significantly different to B1 or B2. This time difference was likely created by the additional time required to explain decisions being made. Participants were not asked to solve the tasks in the shortest amount of time, but were encouraged to get as close to the target answer as possible. As such this metric is a measure of the participants natural behaviour under a given condition.

Mental Workload: Subjective measure

In support of hypothesis HC, we found significant differences between conditions in the NASA-TLX scales: Mental Effort, Mental Demands, and Physical Demands.

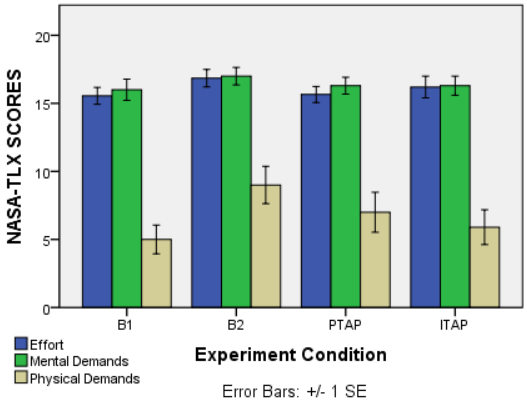

Figure 6. Means for three sub-scales of NASA-TLX

Against our own intuitions, each of these measures demonstrated higher demands for B2 compared to the other conditions (see Figure 6). A Wilcoxon signed-rank test showed that B2 created significantly more mental effort than B1 ($Z = -2.058, p < 0.05$), and it required more mental demands ($Z = -2.292, p < 0.05$). The difference between B2 and PTAP was only $p = 0.075$ and there was no significant difference between ITAP and the other conditions. Participants also rated B2 as being physically more demanding than the other conditions (B1: $p < 0.05$, PTAP $p = 0.067$, and ITAP $p < 0.05$). This is to say that participants found the additional utterance of a nonsense word whilst solving the maths problems induced a greater physical demand than other conditions (see Figure 6).

Correlations between performance scales from unweighted NASA-TLX and performance data were found. This includes a negative Pearson correlation between NASA-TLX Performance scale and distance from target $r = -252, n = 80, p = 0.024$, indicating that participants were rating their performance as worse, when in fact it was better. Two positive Pearson correlations between NASA-TLX Mental Demands and Temporal Demands when compared with time to solve a problem were also found: $r = 0.340, n = 80, p = 0.002$, and $r = 0.408, n = 80, p = 0.001$ respectively.

Mental Workload: fNIRS

Further supporting HC, our analysis found a significant difference in brain region activation in both right and left inferior PFC during the experiment conditions.

As shown in Figure 7, OverallHbO were significantly higher during B2 compared to all other conditions (PTAP: $p < 0.05$, ITAP: $p = 0.064$). We also noted an effect on the rest time at the end of each conditions: values at rest after B2 were significantly higher than values at rest after B1 ($p = 0.05$).

Peck et al [30] found a negative correlation between fNIRS levels of Hb and the subjective ratings from NASA-TLX Mental Demands scale. Tasks that created more mental effort were accompanied by lower levels of Hb. We were

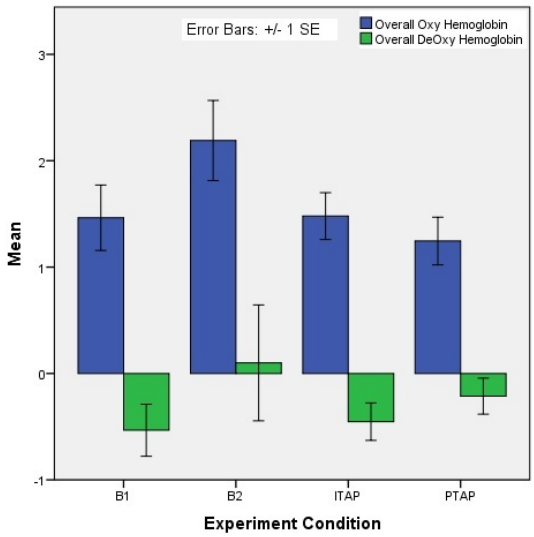

Figure 7. Overall HbO and Hb levels for each condition

unable to confirm these finding across all participants, however we found a positive correlation between performance data (distance from target) and fNIRS overall Hb, $r = 0.228, n = 80, p = 0.04$. This possibly complements Peck's correlation assuming that when mental demands are high to the point of overload, performance decreases and therefore Hb follows. This assumption ties well with the Limited Resource Model presented by Megaw [25].

There was also a strong positive Pearson correlation ($r = 0.474, n = 80, p < 0.001$) between the fNIRS readings Hb left and Hb right.

Mathematics Skill

Peck et al [30] found differences in participants depending on their ability to analyse both pie and bar charts. Similarly, we believed that mathematical propensity would affect an individuals performance under differing TAPs, with the assumption that high performers would better cope with TAPs, while lower performers would be impaired as a result of reduced resources (from TAP).

High Performers

The high performing group rated ITAP as being more mentally demanding (Wilcoxon signed-rank test with $Z = -1.89, p = 0.059$) and requiring more mental effort (Wilcoxon signed-rank test with $Z = -1.98, p = 0.048$) when compared against PTAP. A Spearman negative correlation for the strong mathematicians between the NASA-TLX Mental Demands scale and the fNIRS Hb levels on the right side of the PFC ($r = -0.348, n = 40, p = 0.028$) confirms Peck's [30] findings. High performers also demonstrated a positive Spearman correlation between distance from target and fNIRS Hb on the left side of the PFC ($r = 0.344, n = 40, p = 0.03$). Weighted NASA TLX score also positively correlated with time taken to solve a problem ($r = 0.399, n = 40, p = 0.01$).

Low Performers

For the low performing group we observed an agreement between weighted NASA-TLX score and fNIRS overall Hb. There was the same significant difference from a Wilcoxon Sign Rank test in both NASA-TLX and fNIRS ($Z = -1.78, p = 0.074$) between PTAP and ITAP. Participants workload measured with both NASA-TLX and fNIRS is marginally higher in PTAP than ITAP. This result is opposite to what was observed with the high performing group.

DISCUSSION

fNIRS, Language and Mental Workload

Activations in the left side of the prefrontal cortex are known to occur during semantic, relative to nonsemantic, tasks that relate or involve "the generation of words to semantic cues or the classification of words or pictures into semantic categories" [14]. Due to the physical placement of our fNIRS device on participants foreheads, we can discount the interaction between Broca's area and our results as it does not fall within the reach of our device. Because fNIRS was sensitive to the B2 condition, we developed two premises (interpretations) of the results:

1. fNIRS is particularly picking up the part of the brain that is activated during B2 and therefore the signal received by fNIRS is higher, or
2. fNIRS is picking up an indicator related to mental workload and that B2 induces more workload. The reason behind this is the non-compatibility and non-complementarity of B2 with the mathematical reasoning task, rather than the compatibility of verbalisation protocols from PTAP and ITAP with the mathematical reasoning task.

One way to distinguish between these two is to look at the participants performance data and subjective ratings (the NASA-TLX scores) together with fNIRS. If the first premise is true, you would not expect a difference in mental workload (in the subjective scores) between the verbalisation conditions. Additionally, you would not expect any relationship between performance or NASA-TLX data with fNIRS readings. We found significant difference between verbalisation conditions in NASA-TLX scores and we also found correlations between fNIRS data with both performance and NASA-TLX. If fNIRS would pick up information related to language generation, you would expect significant difference in fNIRS data between verbalisation conditions and the silent condition (which we did not find, see Figure 7). With this in mind, we propose that fNIRS is not an indicator of how many words you are saying, but is sensitive to mental workload and human cognition (therefore provides support for the second premise).

Using the fNIRS alone we were unable to identify the significant differences we were expecting. However we found the fNIRS data to be complementary to existing measures such as performance and NASA-TLX. Considering the number of marginally significant results($p < 0.075$), we believe that increasing the number of participants would increase power, reduce type II error, and positively impact our findings.

High and Low Performers

If generalisable, our findings suggest that for high performers PTAP is the more suitable protocol and that ITAP is better suited to low performers. One possible explanation for this is that high performers have an existing procedural structure in which they operate, so interrupting this procedure (as is experienced under ITAP) potentially interferes with their natural behaviour. For low performers, however, such structure is not present and verbalising via PTAP is potentially troublesome, as they are being forced to verbalise a process that is absent or unnatural for them. The introduction of carefully chosen prompts, however, may encourage non-experts to describe how they are struggling and provide useful insight into how researchers may help these types of users in the future.

Using fNIRS to Measure Mental Workload

In this study we looked at evaluating the cognitive impact of various TAPs using fNIRS as a novel measurement. fNIRS was chosen for its non invasive application, portability and relative resilience to motion artefacts. We found the device to be suitable ecologically for HCI style user study settings, with the device providing minimal distraction and interference. After completing the study, we informally questioned the participants regarding their experience with wearing the fNIRS device. No participant described feeling particularly uncomfortable during the study, some did however state that they began to experience some discomfort towards the end of the study. We advise that studies utilising fNIRS should aim to keep sessions below 1 hour in a single sitting.

We believe that fNIRS is well suited to HCI evaluation and usability testing. We believe that the inclusion of this novel new measurement complements existing evaluation measures such as NASA-TLX. fNIRS benefits from having the properties of being both an objective and continuous measure allowing for accurate, time correlated recording during evaluation and testing studies, especially when compared to the subjective one time snapshot rating achieved via NASA-TLX. We must also note the potential negatives associated with this type of technology. fNIRS is an emerging technology and as such does not have the associated supporting research proving its correctness. Studies have correlated the measurements to those observed with fMRI [35], specifically the BOLD signal. Additionally, in the current state of technology, fNIRS can only be used to detect a level of workload (high or low), leaving a distinct lack of mapping between the readings recorded with fNIRS and the actual cognitive or emotional states. For example, detecting frustration under a evaluation study would be a useful measure, but is not currently obtainable from fNIRS.

Another point of interest, that can possibly be considered a shortcoming of this study is the exclusion of performing the study task without wearing the fNIRS device. Doing so would allow us to determine whether fNIRS affected performance or behaviour in anyway. We did ask however, as a part of the informal post study interview, whether participants felt that they were influenced in some way by wearing the device; no one reported such an effect. This does leave the potential for a follow up study to examine whether there was indeed an effect.

Running a TAP

One of our research questions was to investigate two think aloud protocols (namely PTAP and ITAP). The study results should be seen as a positive indicator that both TAPs do not significantly affect or influence participants ability to solve the tasks presented in the study. We used a high demand tasks and participants performance was not negatively affected in any way. Contrarily, we observed a slight improvement in participants' performance under TAP conditions, confirming with McDonald [24] that using the TAPs during the task did not have a negative influence on participants' performance.

Reflecting on Wicken's Multiple Resource Model, using multiple resources that are complementary and compatible with the task in hand might have a positive impact on performance in the case of non multitask resource overload. Between the four conditions, participants performed the worse in Condition B2 where they had to repeatedly say 'Blah' during task solving. This was due to a higher workload generated by the condition, sensed with both fNIRS and NASA TLX subjective scale.

The TAPs conditions differed when compared between the expertise level of participants. The high performing group rated ITAP as being more mentally demanding requiring more mental effort when compared against PTAP. This result was also confirmed with the fNIRS data. Conversely for the low performing group, PTAP was the one that was more mentally demanding.

CONCLUSIONS

The aims of this research were a) to investigate how *verbalisations* might affect the use of fNIRS in increasingly ecologically valid user studies, and b) to provide insights into how different forms of verbalisations affect mental workload and performance in user studies. In order to achieve our aims, we compared nonsense verbalisations with different forms of concurrent TAP: passive and invasive. One of our primary findings was that non-complementary verbalisations, as opposed to complex verbalisations, created higher levels of mental workload. In particular, nonsense verbalisations created higher mental workload, across measures, than Invasive TAP where participants discussed their mathematical problem solving options. Consequently, we can conclude that the use of TAPs in user studies is fine as long as the discussion uses words relating to solving the task. We saw a slight increase in mental workload for Invasive TAP compared to Passive TAP, indicating that some Invasive TAP verbalisations may not have been directly conducive to solving the task. None of the nonsense verbalisations supported the task.

The findings about non-complementary language were hidden within the subjective, reflective, self-assessments included in NASA TLX; ratings had high variance, and results were only evident in some of the sub-scales. Further, we saw no difference in task performance between conditions. The objective measure obtained from the fNIRS however, provides a clear indication of the participants' mental workload whilst completing the study tasks. Because there were no differences between the silent baseline and TAP conditions, we can conclude a) that fNIRS measurements were not largely affected by verbalisation itself, and b) that fNIRS can be used to determine mental workload objectively during tasks if verbalisations remain task-related.

Overall, we provide three main contributions: 1) we provide novel insights into the underlying cause of increased mental workload created by TAPs *during* tasks; 2) we provide a novel example of using fNIRS to measure cognition during a more complex task than prior work; and 3) we provide an example to show that fNIRS is suitable for use with tasks that involve verbalisation. Our results make a positive step towards proactively using fNIRS as an evaluation tool within realistic HCI user studies.

REFERENCES

1. Aron, A. R., Robbins, T. W., and Poldrack, R. A. Inhibition and the right inferior frontal cortex. *Trends in cognitive sciences 8*, 4 (2004), 170–177.
2. Baddeley, A. Working memory. *Science 255*, 5044 (1992), 556–559.
3. Baddeley, A. The episodic buffer: a new component of working memory? *Trends in cognitive sciences 4*, 11 (2000), 417–423.
4. Baddeley, A. D. Is working memory still working? *European psychologist 7*, 2 (2002), 85–97.
5. Baddeley, A. D., and Hitch, G. Working memory. *The psychology of learning and motivation 8* (1974), 47–89.
6. Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., and Noll, D. C. A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage 5*, 1 (1997), 49–62.
7. Charters, E. The use of think-aloud methods in qualitative research an introduction to think-aloud methods. *Brock Education* (2013).
8. D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., and Grossman, M. The neural basis of the central executive system of working memory. *Nature 378*, 6554 (1995), 279–281.
9. Dumas, J. S. *A practical guide to usability testing.* Intellect Books, 1999.
10. Duvinage, M., Castermans, T., Dutoit, T., Petieau, M., Hoellinger, T., Saedeleer, C., Seetharaman, K., and Cheron, G. A P300-based quantitative

comparison between the Emotiv Epoc headset and a medical EEG device. In *IASTED* (2012).
11. Ericsson, K. A., and Simon, H. A. *Protocol analysis*. MIT press, 1985.
12. Fadiga, L., Craighero, L., and DAusilio, A. Broca's area in language, action, and music. *Annals of the New York Academy of Sciences 1169*, 1 (2009), 448–458.
13. fNIR Devices LLC. *fNIR Imager & COBI Studio Manual*. Biopac, 2013.
14. Gabrieli, J. D., Poldrack, R. A., and Desmond, J. E. The role of left prefrontal cortex in language and memory. *Proceedings of the national Academy of Sciences 95*, 3 (1998), 906–913.
15. Hart, S. G., and Staveland, L. E. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload* (1988).
16. Hertzum, M., Hansen, K. D., and Andersen, H. H. Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology 28*, 2 (2009), 165–181.
17. Hirshfield, L. M., Solovey, E. T., Girouard, A., Kebinger, J., Jacob, R. J., Sassaroli, A., and Fantini, S. Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In *Proc. CHI*, ACM (2009), 2185–2194.
18. Izzetoglu, M., Bunce, S. C., Izzetoglu, K., Onaral, B., and Pourrezaei, K. Functional brain imaging using near-infrared technology. *IEEE Engineering in Medicine and Biology Magazine 26*, 4 (2007), 38.
19. Kane, M. J., and Engle, R. W. The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic bulletin & review 9*, 4 (2002), 637–671.
20. Landis, J. R., and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* (1977), 159–174.
21. Leanne, M., and Robert, J. Using brain measurement to evaluate reality based interactions. *Challenges in the Evaluation of Usability and User Experience in Reality Based Interaction 5* (2009), 19–20.
22. Li, X., Xu, L., Yao, L., and Zhao, X. A novel HCI system based on real-time fmri using motor imagery interaction. In *Foundations of Augmented Cognition*. Springer, 2013, 703–708.
23. McDonald, S., Edwards, H. M., and Zhao, T. Exploring think-alouds in usability testing: an international survey. *Professional Communication, IEEE Transactions on 55*, 1 (2012), 2–19.
24. McDonald, S., and Petrie, H. The effect of global instructions on think-aloud testing. In *CHI*, ACM (2013), 2941–2944.
25. Megaw, T. The definition and measurement of mental workload. *Evaluation of human work, Eds. Esmond N. Corlett, and John R. Wilson* (2005), 525–551.
26. Miles, J., and Shevlin, M. *Applying regression and correlation: A guide for students and researchers*. Sage, 2001.
27. Miller, E. K., and Cohen, J. D. An integrative theory of prefrontal cortex function. *Annual review of neuroscience 24*, 1 (2001), 167–202.
28. Mueller, S. Pebl: The psychology experiment building language (version 0.10).[computer experiment programming language]. *Retrieved Nov* (2012).
29. Olmsted-Hawala, E. L., Murphy, E. D., Hawala, S., and Ashenfelter, K. T. Think-aloud protocols: a comparison of three think-aloud protocols for use in testing data-dissemination web sites for usability. In *CHI*, ACM (2010), 2381–2390.
30. Peck, E. M., Yuksel, B. F., Ottley, A., Jacob, R. J., and Chang, R. Using fNIRS Brain Sensing to Evaluate Information Visualization Interfaces. In *CHI*, ACM (2013).
31. Pike, M., Wilson, M. L., Divoli, A., and Medelyan, A. CUES: Cognitive Usability Evaluation System. In *EuroHCIR2012* (2012), 51–54.
32. Rushworth, M. F., Nixon, P. D., Eacott, M. J., and Passingham, R. E. Ventral prefrontal cortex is not essential for working memory. *The Journal of Neuroscience 17*, 12 (1997), 4829–4838.
33. Solovey, E. T., Girouard, A., Chauncey, K., Hirshfield, L. M., Sassaroli, A., Zheng, F., Fantini, S., and Jacob, R. J. Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines. In *Proc. UIST2009*, ACM (2009), 157–166.
34. Tan, D. S., and Nijholt, A. *Brain-Computer Interfaces: applying our minds to human-computer interaction*. Springer, 2010.
35. Toronov, V., Webb, A., Choi, J. H., Wolf, M., Michalos, A., Gratton, E., and Hueber, D. Investigation of human brain hemodynamics by simultaneous near-infrared spectroscopy and functional magnetic resonance imaging. *Medical physics 28* (2001), 521.
36. Villringer, A., and Chance, B. Non-invasive optical spectroscopy and imaging of human brain function. *Trends in neurosciences 20*, 10 (1997), 435–442.
37. Villringer, A., Planck, J., Hock, C., Schleinkofer, L., and Dirnagl, U. Near infrared spectroscopy (NIRS): a new tool to study hemodynamic changes during activation of brain function in human adults. *Neuroscience letters 154*, 1 (1993), 101–104.
38. Wickens, C. D. Multiple resources and mental workload. *The Journal of the Human Factors and Ergonomics Society 50*, 3 (2008), 449–455.
39. Wickens, C. D., Gordon, S. E., and Liu, Y. *An introduction to human factors engineering*. Pearson Prentice Hall Upper Saddle River, 2004.
40. Ye, J. C., Tak, S., Jang, K. E., Jung, J., and Jang, J. NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy. *Neuroimage 44*, 2 (2009), 428–447.