Real-Time Physiological Measure and Feedback of Workload

The University of
Nottingham

UNITED KINGDOM · CHINA · MALAYSIA

Horia Alexandru Maior

School of Computer Science

University of Nottingham

Advisers: *Max L. Wilson and Sarah Sharples*

This dissertation is submitted for the degree of

Doctor of Philosophy

2017

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where specifically indicated in the text. This dissertation contains less than 100,000 words including appendices, bibliography, footnotes, tables and equations and has less than 150 figures.

Horia Alexandru Maior
2017

Acknowledgements

I would like to thank the Research Councils UK, the Horizon Centre for Doctoral Training (CDT), the Mixed Reality Lab (MRL) and the Human Factors Research Group here at Nottingham University, for all the financial and academic support to conduct the research presented in this thesis. Particular acknowledgement to the two contributing grants EPSRC (EP/M000877/1) and Horizon (EP/G037574/1)

I would also like to thank the two examiners, Robert Jacob and Robert Houghton, who have set aside some of their valuable time to review this thesis. I would also like to thank the Horizon Digital Economy Research Hub for the financial and academic support during my research internship at the International Institute of Information Technology Bangalore (IIIT-B).

I am very grateful for the continuous support and I would like to express my sincere gratitude to my Ph.D. supervisors Max L. Wilson and Sarah Sharples, who have been an amazing team that offered both, professional and personal support during my time at Nottingham University. My deepest gratitude goes to Max, who introduced me to the research field, playing the role of a mentor but also a friend throughout. Thank you for your patience, motivation, immense knowledge and friendship. His guidance helped me find light during the “darkest moments”, or even better.. helped me avoid dark moments! I could not have imagined having a better advisor and mentor for my PhD.

I would especially like to thank Matthew Pike for being THE COMRADE of this whole madness; I hope you know I forgive you and Sorin for the incident in Bremen. We can be friends again :)!

Thanks to my awesome office colleagues Martin Porcheron, Richard Ramchurn, and Chris Ellis for always accepting being interrupted during their daily work with my boring stories :) and my never funny “that’s what she said” jokes. Thanks to Adrian Marinescu, Li Xia, Pepita Stringer, George Filip, Lachlan Urquhart, Dimitri Darzentas

and all the CDT/IDIC colleagues and friends for the stimulating discussions, the sleepless nights and working hard before deadlines, and the precious moments we will never forget.

And finally but most importantly I would like to thank my parents Stefan and Camelia, my brother Tuliu, his wife Alexandra, my aunty Tulia, and the rest of my family and friends including Micusan Family (Adi, Laura, Adina, Mada and Codrin), Sorin Dinu Jr. and Sn. Family, Alexandru Ghitulescu, Alexandru Mardale, Anand Masilamani, Georgiana Avram, Sorin Nica, Andrei Rotaru, Monica Georgescu, Zack Stern, Sandra Mocrei, Maria Karali, Nada El Kateb, Mark Katona, Ciprian Paraschevici, Iulian Ciuborariu, Jillian Donaldson for their continuous support, kind words and happiness they always bring to my heart; without their support and friendship this PhD would have not been possible.

Abstract

Understanding and identifying individuals' capabilities and limitations has always been a challenge within work contexts, but its importance cannot be underestimated.

Humans have a limited mental capacity [142], which means that they can only perform a finite set of tasks at any given period of time. Identifying these limitations is a key factor in the reduction and prevention of what is referred to as Mental Workload Overload. These measures are used in research and industry to evaluate the interaction of users with new systems and tasks. Current techniques involve asking users to subjectively assess and self report their levels of workloads using techniques and questionnaires such as NASA-TLX and Instantaneous Self-Assessment (ISA). The subjective measures become highly important when it comes to evaluating more complex systems and tasks, where performance based measures become highly difficult to measure. Even though they are critical for evaluation of these systems, there are certain limitations that cannot be overlooked when using them. Firstly, subjective measures rely on the participants' ability to judge and report the state throughout the task. This requires not only extra effort from the operator, but also skill and potential training. Secondly, subjective measures, if used in real-time have the potential to interrupt and negatively affect performance; if used post-task, they rely on the operators' ability to recall what happened during certain moments in the past. Direct physiological measures offer an opportunity to capture workload whilst overcoming these limitations. However, new research is needed to understand how physiological data can be interpreted within the context of theories of mental workload. The research presented in this thesis explores the use of one particular physiological approach, functional Near Infrared Spectroscopy (fNIRS), to assess workload in controlled laboratory settings, to overcome the limitations and complement the use of subjective measures; a measure based on participants' brain and physiological responses to task demand, that is independent of the task and/or the operator (without interrupting the task or relying on the operator skill to self report).

We have examined the reliability of the technique, and significantly extended our understanding of how artefacts affect recordings during both - a Verbal memory task of remembering a seven digit number and a Spacial memory task of remembering a 6x6 shaped grid. Our results showed that artefacts have a significantly different impact during the two types of tasks, further contributing insights into the existing guidelines of using fNIRS to assess workload during typical human computer interaction evaluation settings. We have further evaluated the sensitivity of the tool and understand the potential implications of using fNIRS as a measure in real-time. Our findings validated fNIRS as a sensitive workload measure, having consistent results in line with subjective measures, confirming a correlation between fNIRS and subjective workload questionnaires NASA-TLX and ISA. Having shown the relationship between fNIRS and workload, the last part of this thesis explores the use of fNIRS as a novel approach to providing users with concurrent feedback of their Mental Workload based on the measurements obtained objectively from fNIRS. We compare this feedback to traditional methods of asking users to self-assess and report their own mental workload during an Air Traffic Controller simulation game. In line with previous work, we confirm that self-reporting methods affect both perceived and actual performance. Furthermore, we found that our objective concurrent feedback technique allowed participants to reflect metacognitively on their Mental Workload during tasks, without reducing either actual or perceived performance.

fNIRS showed potential to be a useful and reliable additional channel of information about the user during interaction, without further restricting the user during a typical evaluation settings. We found it sensitive to workload, being able to distinguish between various levels of workload, and with great potential for real time, continuous use during tasks. Finally, we explored a new direction of using fNIRS's assessment of workload in real time, and we investigated how users can use feedback of their current workload state during tasks. This proved to allow users to think metacognitively about their workload during tasks, without negatively affecting their performance or workload.

Based on the findings presented in this thesis, scope for future research is proposed and discussed.

Thesis Publications

- Maior, H. A., Wilson, M. L. and Sharples, S. (Under Review) Workload Alerts - Using Physiological Measures of Mental Workload to Provide Feedback during Tasks. *In ACM Transactions on Computer-Human Interaction (TOCHI).*

- Wilson, M. L.,Alsuraykh, Norah and Maior, Horia A. Measuring mental workload in IIR user studies with fNIRS. 2017.

- Lukanov K., Maior, H. A., and Wilson, M. L. Using fNIRS in Usability Testing: Understanding the Effect of Web Form Layout on Mental Workload. *In: CHI'16 ACM SIGCHI Conference on Human Factors in Computer Systems*, San Jose, California, May 2016.

- Maior, H. A. Pike, M., Wilson, M. L., and Sharples, S. Examining the Reliability of Using fNIRS in Realistic HCI Settings for Spatial and Verbal Tasks. *In: CHI'15 ACM SIGCHI Conference on Human Factors in Computer Systems*, Seoul, Korea, April 2015.

- Pike, M., Maior, H. A., Porcheron, M., Sharples, S. and Wilson, M. L. (2014). Measuring the effect of Think Aloud Protocols on Workload using fNIRS. *In: CHI'14 ACM SIGCHI Conference on Human Factors in Computer Systems*, April-May 2014, Toronto.

- Maior, H. A., Pike, M., Wilson, M. L., and Sharples, S. Continuous detection of workload overload: An fNIRS approach. *In Contemporary Ergonomics and Human Factors 2014: Proceedings of the international conference on Ergonomics & Human Factors 2014*, Southampton, UK, April 2014.

- Maior, H.A., Sharples, S., and Wilson, M.L. Subjective and Objective Methods to Continuously Monitor Workload. *Neuroergonomics 2016: The brain at work and in everyday life*, Paris, France 2016.

- Maior, H. A., Wilson, M.L. and Sharples, S. (2015). fNIRS in Human Factors. *2FNIRS Workshop*, Toulouse, France April 2015.

- Maior, H. A., Pike, M. Measuring Work Overload. *The Ergonomist Magazine*, May 2014.

- Maior, H. A., Pike, M., Wilson, M.L. and Sharples, S. (2013). Directly Evaluating the Cognitive Impact of Search User Interfaces: a Two-Pronged Approach with fNIRs. *EuroHCIR 2013 Workshop*, Dublin, Ireland, August 2013.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Although computers are very good at performing repetitive rule-based tasks, humans can better perform 'knowledge-based' tasks that involve decision making processes and new solutions for rapidly changing problems, and developing them based on past experience and innovation [111]. Furthermore, as technology pervades our everyday life, our own tasks are increasingly *"dominated by mental rather than physical task components"* [121]; the human role has moved towards a supervisory and decision making role for such intelligent systems.

This move has the potential to increase the demands on people's mental resources due to the amounts of data being generated and the amounts of concurrent tasks and decisions we make every day. However, humans have a limited capacity [142], which means that they can only perform a finite set of tasks at any given period in time. One problem that may occur during interaction with technology is, therefore, operator overload.

Understanding and identifying individuals' capabilities and limitations has always been a challenge within work contexts, but its importance cannot be underestimated. The evaluation of mental workload plays a key role in the development of Human-Computer Interaction and Human Factors with the focus of ensuring a higher level of comfort, engagement and satisfaction during interaction with day-to-day computers [115], as well as ensuring safety in work contexts, while the user is able to reach goals and have a high standard of performance, avoiding states such as overload.

Years of research have been focused on the development and understanding of methods for assessing workload [16, 42, 115, 147]. This, however, has been a challenge as the concept of mental workload is vaguely defined, with poorly understood mechanism, with indirect measures and "embryonic levels of analytic models" [52]. Cain stresses that despite the interest in the topic for more than 40 years, there is no one accepted definition of workload [26]. For the purpose of this thesis, we suggest a working definition of mental workload based on multiple viewpoints [26, 49, 55, 56, 121, 143] :

"Mental workload refers to the amount of effort or strain imposed by the demands of a task on an individual. The "effort" may have different forms and influenced by multiple factors. It includes the amount of effort perceived by an individual during tasks, any physiological or psycho-physiological changes imposed by task demands on the individual, as well as any impact on individual's task performance, that is caused by changes in task demands or external factors while performing the task."

Many questions have been raised regarding the mental workload measurement, investigating various aspects of the measure as well as the procedures of the measure. Questions regarding the user's involvement in the process, user's ability to perform additional tasks, aspects on the nature of the tasks (e.g. task complexity), the user's feelings whilst performing tasks, and others [115]. Methods for assessing workload have ranged from subjective measures such as NASA-TLX [55], to quantitatively measure individual's cognition utilising a mixture of task performance and physiological measures. Considering these aspects, various tools for measuring workload have been proposed. Meshkati, Hancock, & Rahimi, 1992 [92] categorized them into three main types:

a. performance based measures;

b subjective procedures;

c. and physiological measures.

Performance measures include primary task and secondary task analysis, and are based on the assumption that a drop in performance is caused by an increase in task demands, therefore an increase in mental workload. However, these methods suffer from trade-off effects that are sometimes hard to account for. You can try to bias it with an explicit system of pay-offs but then some would say that it just introduces another level of confound. It is also the case that one can not necessarily safely assume that the secondary task is merely additive and does not alter the ways in which performance is achieved in the primary task.

Subjective procedures are based on the assumption that increased task demand is linked to perceived effort, therefore it can be appropriately assessed by an individual. They work on the assumption that an individual is exerting some extra effort understanding and self-judging their state whenever prompted. Repeatedly evaluating and subjectively reporting workload during tasks may have a self-aware effect on the user/operator, which may lead to introspection of workload, hardly to achieve otherwise. In other words, this may make users/operators more aware of their workload state during tasks, allowing them to prioritize tasks in case of situations such as overload. Although they are highly used in research, subjective techniques tend to be interruptive to the task in hand, causing disruption and/or loss of focus, potentially affecting task performance.

Physiological measures are typically task independent and therefore provide a way of having an objective quantifiable measure whilst not interrupting the task. These measures are based on the assumption that variations in mental workload experienced by an individual will cause variations in levels of physiological activation. Whether directly or not, the physiological measures aim to characterise users' mental state experienced during the task, without relying on users' ability to subjectively report this information. Physiological measures include brain related measures, eye related measures (blink interval and blink rate, pupil dilatation), heart related measures (heart rate, blood volume, heart rate variability), as well as skin related measures (e.g. skin temperature variations in various parts of the body, galvanic skin response). Advances in brain imaging technologies opened doors when studying interaction with technology, by allowing the collection of useful information during interaction whilst remaining portable and non-invasive.

1.2 Thesis Statement and Research Questions

Recent research has been examining the potential of functional Near Infrared Spectroscopy (fNIRS) as perhaps one of the ways to assess workload. This thesis focuses on using this emerging neuroimaging technique in the field of Human Computer Interaction and Human Factors, to objectively assess users' mental workload during interaction with technology. This will allow a better understanding of the users' abilities and workload capacity in an objective way that is complementary to subjective techniques.

We explore the use of fNIRS as a potential, useful measure during the evaluation of technology because it is safe, less-invasive compared to other brain based techniques (it allows normal interaction with a computer without further restricting the user, or requiring additional discomforts such as gel on the sensors), and relatively portable. Furthermore, with the recent advances in the area of sensors development, this research becomes more relevant as sensors become available to a larger population. Because the nature of fNIRS technology was originally designed for clinical use, this thesis will focus on understanding the practicality of the measure during Human Computer Interaction and Human Factors. The first aim of this thesis is therefore:

- Explore how fNIRS can be used to gain insights into mental workload during interaction with technology within realistic lab-based evaluation settings by continuing the work started by Solovey et. al. [127].

If we know that workload is a useful notion and construct, that is probably a combination of memory and attentional demands, and fNIRS will appear to correlate with this measure, it will be therefore useful to understand the demands placed upon people, and how the demands impact users' (available) mental resources during real time task completion. The second aim of this thesis is:

- Test the sensitivity and validity of the fNIRS measure in the context of real-time, continuous use for assessing workload during tasks.

Moreover, an individual self-assessment of workload, would potentially allow people to regulate their resource allocation to the primary task, this way avoiding situations where the demands placed upon them will exceed their capabilities and limitations in terms of mental workload; therefore in the last aim of this thesis is:

- Explore the use of objective assessments of workload using fNIRS, to provide real-time concurrent feedback of mental workload to users during tasks.

Based on the aims presented above, we devise and present the thesis research questions below. The research questions follow on from each other, meaning that each question is somehow dependent on the previous questions.

1. How can fNIRS be used to assess workload during interaction with technology within realistic lab-based evaluation settings?

   (a) Investigate the possibility of using fNIRS' measures of oxygenated (HbO) and de-oxygenated (Hb) hemoglobin to distinguish between a "busy" state (participant performing a task) and a "rest" state (participant performing no task).

   (b) Investigate how fNIRS can be used in the presence of artefacts produced during interaction with technology within realistic lab-based evaluation settings. Understand the impact of various artefacts on the two different task encodings: spatial task vs verbal tasks.

   (c) Investigate the Reliability, Replicability, Sensitivity, Validity, of the fNIRS measure: Understand the sensitivity of the measure to both spatial and verbal memory tasks; investigate methods to distinguish between various levels of workload using fNIRS.

2. How can fNIRS be used as a sensitive and valid technique in the context of continuous, real-time use, to gain insights into mental workload during tasks?

   (a) Investigate the validity of fNIRS measure in contrast to the subjective techniques including NASA-TLX and the continuous Instantaneous Self Assessment technique (ISA).

   (b) Investigate the implications of moving beyond block design, towards the real time-continuous measure of workload (using fNIRS).

3. How can a real time, continuous version of fNIRS be used to give workload feedback to the user?

   (a) Explore the impact of workload feedback on task performance.

   (b) Explore the impact of workload feedback on subjective ratings.

1.3 Thesis Contributions

To support the aims and questions of this thesis, my research touches the fields of Human-Computer Interaction (HCI), Human Factors, signal processing and brain computer interfaces, and contribute mainly to the first two.

Human Factors is where the notion of workload has been extensively studied, and it focuses on understanding how humans behave physically and psychologically in relation to particular environments, tasks and jobs. In this thesis we explored and evaluated models from Human Factors literature, and contributed with several findings in terms of mental workload and the measurement of mental workload.

On the other side, HCI is where the interaction between people and technology sits, and it is focused on studying how people interact with computers and to what extent computers are or are not developed for successful interaction with human beings. The thesis contributes to the HCI literature by exploring ways in which the notion of workload can be used to better study the interaction between people and technology. It also provides practical guidelines and examples on how the measurement of mental workload can be useful for HCI. There is a massive overlap between the two disciplines, with various different schools of thought, but this work sits in the intersection between the two.

The contribution of this thesis can therefore be summarized as follows:

- This thesis contributes to the measurement and assessment of workload using fNIRS. The reliability of the measure was tested within lab-based evaluation settings, and we extended the understanding of its use during both verbal and spatial tasks.

- This thesis further contributes to the real time measurement and use of fNIRS during more natural tasks. We further tested the sensitivity and validity of the measure, and extended our understanding of workload in relation to performance measures, subjective techniques and physiological methods using fNIRS.

- The last contribution of this thesis is focused on exploring the potential impact of workload feedback during tasks. We investigated how feedback of mental workload (based on real time measurements during tasks using fNIRS) could be useful to people, and we showed how people think metacognitively about their state during tasks.

1.4 Thesis Overview

Table 1.1 presents the structure of this thesis based on chapter contributions to the research questions.

Chapter 2 and Chapter 3 provide an overview of the related work that lays the foundation of this dissertation. Chapter 2 presents the core models of mental workload from a Human Factors perspective and the methods used to measure it. Chapter 3 explores and presents physiological techniques used to measure mental workload, and presents how fNIRS is used in this thesis for assessing workload.

Chapter 4 presents a user study addressing the research question RQ1, investigating the reliability of fNIRS in a typical user study scenario. It is the baseline - proof of concept - study as it investigates the suitability and capability of using fNIRS to collect useful information about the users during interaction with technology.

This research then leads into the study presented in Chapter 5, further investigating the validity and sensitivity of fNIRS, as an objective, continuous technique to assess workload during tasks in the context of real-time continuous use. This chapter addresses aspects of research question RQ2a and RQ2b, but also contributes to RQ1b.

Chapter 6 addresses the final research questions of the thesis, and explores the use of feedback of mental workload during tasks.

The thesis ends with the discussions and conclusions in Chapter 7.

Table 1.1 Thesis Chapter Overview

	Description	**Methodology used/** **Research Questions Addressed**
Chapter 2	Theoretical background of workload	Literature Review
Chapter 3	Physiology and fNIRS	Literature Review and Methodology
Chapter 4	How can fNIRS be used to assess workload during interaction with technology within realistic lab-based evaluation settings?	• Empirical study (15 participants). • Research Questions involved: RQ1a, RQ1b, RQ1c. • Task: Simple Verbal and Spatial Memory task (Low Complexity). • Main Publication: Examining the Reliability of Using fNIRS in Realistic HCI Settings for Spatial and Verbal Tasks. *In: CHI'15 ACM SIGCHI Conference on Human Factors in Computer Systems*, Seoul, Korea, April 2015.
Chapter 5	How can fNIRS be used as a sensitive and valid technique in the context of continuous, real-time use, to gain insights into mental workload during tasks?	• Empirical study (20 participants). • Research Questions involved: RQ1b, RQ2a, RQ2b. • Task: Mathematical Problem Solving (Countdown problem) Verbal memory task (Medium Complexity). • Main Publication: Continuous detection of workload overload: An fNIRS approach. *In Contemporary Ergonomics and Human Factors 2014: Proceedings of the international conference on Ergonomics & Human Factors 2014*, Southampton, UK, April 2014.
Chapter 6	How can a real time, continuous version of fNIRS be used to give workload feedback to the user during tasks?	• Empirical study (32 participants). • Research Questions involved: RQ1c, RQ2a, RQ2b, RQ3a, RQ3b. • Task: Air Traffic Control Simulator Game (High Complexity). • Main Publication: (Under Review) Workload Alerts - Using Physiological Measures of Mental Workload to Provide Feedback during Tasks. *In ACM Transactions on Computer-Human Interaction (TOCHI).*
Chapter 7	Discussions and Conclusions	Lessons Learned and Future Directions

Chapter 2

Mental Workload

This chapter presents two major parts of the thesis related works. The first part will be focused on the concept behind mental workload. As the term was extensively studied in the field of Human Factors, multiple models of mental workload from the field will be presented and discussed. The final part of the chapter will focus on the measurement of mental workload. We will discuss in detail the methods used to capture the experienced mental workload but also how workload relates to other factor such as users' physiology.

2.1 Workload characteristics

Mental workload is a concept used to describe how much mental effort is being experienced by an individual when completing a task. It is described by Hart and Staveland (1988) [55] as a relationship between the mental processing capabilities and the demands imposed by a task. Non-optimal workload levels may result in human performance issues such as slower task performance and increase in error rates such as slips, lapses or mistakes.

Although the topic has been around for more than 40 years, there is no clear definition of mental workload. Huey and Wickens discussed the origin of the term, which did not appear in many dictionaries until 1970 [64]. Psychologists have used the term in the context of attention and performance, engineers have used it in the context of aircraft design as a critical factor in system effectiveness, however the term workload is something that every one of us has experienced in one way or another - we have all experienced periods of high and low task demands within a specific time period. Huey and Wickens pointed out the close relationship between workload and performance, and discussed workload using four characteristics:

- The relationship between task demands and workload - as difficulty of the task increases, or the demands imposed on an operator increase, the workload is expected to increase.
- The relationship between task performance and workload - as performance deteriorates (error rates increase, or the task precision decreases), workload is expected to increase.
- The relationship between mental and physical effort of an the operator and workload - workload reflects the impact of the task demands on operator rather than the task demands directly;
- The relationship between the perceived effort by the operator and workload - when an operator feels effortful then the workload is expected to be high.

One assumption found in most of the workload definitions, is that workload is a concept that exists in a relationship between an individual (operator) and a specific assigned task, within a specific time constraint; it is much more about the way the task was experienced by the individual, and the impact of the task demands on the individual rather than workload as an absolute measure. Sharples and Megaw [121] described the effect of workload as *"the relationship between primary task performance and the resources demanded by the primary task"*. The consequences of optimal/non-optimal workload levels may have a direct impact on performance. It is expected, but not always, that when the task difficulty increases, performance degrades. This is typically the case when the demands placed upon the operator increase with the difficulty of the task to a level where the operator can no longer cope with the work (potentially reaching a level close to the operators' maximum capacity). This scenario may be reflected, as mentioned, into performance degrade, but it could also be captured in the operator's subjective experience of the task, or reflected in participants' physiological data - an increase in arousal is expected when workload increases (e.g. increase in heart rate).

In order to discuss and better understand workload, a few core models at the heart of the concept are presented in this chapter. Moreover, these models were explored as they are used in this thesis to understand and break down the elements of workload in the studies presented in later chapters. These include:

- the Working Memory Model first proposed by Baddeley and Hitch [9–13],
- the originally described models of information processing,
- the Multiple Resource Model first proposed and later developed by Wickens et al [143, 144],
- and the Limited Resource Model, adapted by Sharples and Megaw [121].
- the Framework for Mental Workload Measurement [121]

2.1.1 Working memory

"The concept of working memory proposes that a dedicated system maintains and stores information in the short term, and that this system underlies human thought processes."(A. Baddeley [11])

In an attempt to characterise and model the cognitive processes involved when a participant is performing a task demanding mental resources, we draw on research into psychology models of memory, such as *short-term memory*, *long-term memory*, and *working memory*, and we will be focusing mainly on the latter one.

As described by Cowan [32], the term short-term memory was first used by Broadbent (1958) [22] and Atkinson and Shiffrin (1968) [5] in order to describe "... faculties of the human mind that can hold a limited amount of information in a very accessible state temporarily". Baddeley and Hitch 1974 developed an alternative model of short-term memory, called Working Memory [13], a specific system in the brain which "provides temporary storage and manipulation of information" [9]. They were first to discuss the limitations of short-term memory, and "... demonstrated that a single module could not account for all kinds of temporary memory" [32], therefore their proposed model of working memory was composed of multiple components [13]. Its main characteristics are focused on the way in which information is processed and encoded in our brain, namely he distinguishes between two types of encodings: verbal and spatial information encoding. While working memory processes information in the two aforementioned forms: verbal and spatial, Baddeley first divided the process through three main components (Figure 2.1) [13]: a visuo-spatial sketch pad holding information in an analogue spatial form (e.g. colours, shapes, maps), a phonological loop holding verbal information in an acoustical form (e.g. numbers, words), and finally, a central executive acting as a supervisory system and controlling the information from and to its "slave systems".

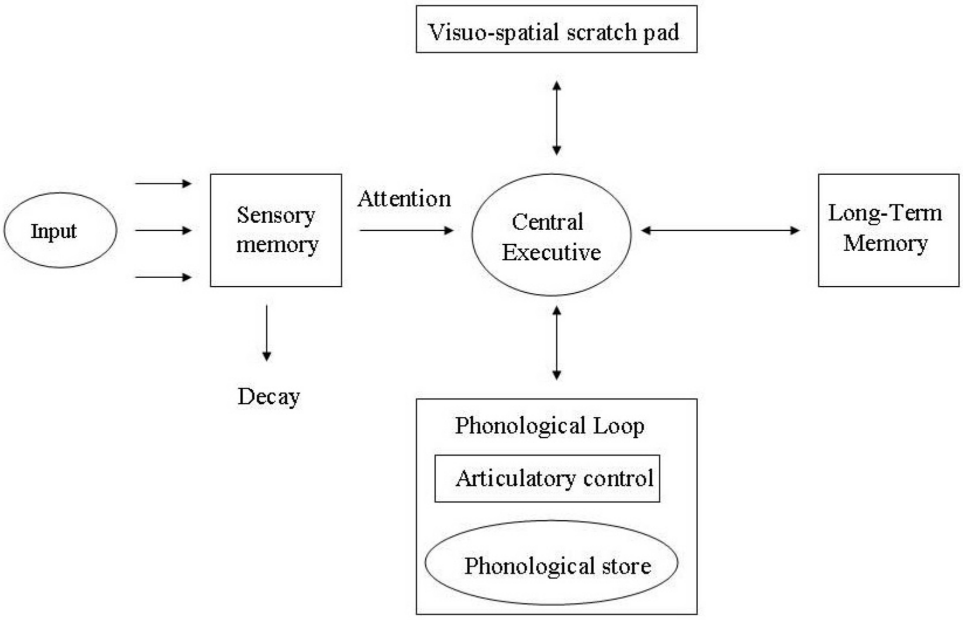


Figure 2.1 Baddeley and Hitch 1974 Working Memory Model [13]

Later work completed the working memory model by introducing the episodic buffer [10, 12], which is dedicated to linking verbal and spatial information in chronological order (presented in Figure 2.2).

Therefore, the four major components of Baddeley's model of working memory are:

- A **central executive** managing attention, acting as a supervisory system and controlling the information from and to its "slave systems".
- A **visuo-spatial sketch pad** holding information in an analogue **spatial** form (e.g. colours, shapes, maps); specialised on learning by means of Visuo-Spatial imagery.
- A **phonological loop** holding **verbal** information in an acoustical form (e.g. numbers, words); specialised on learning and remembering information using repetition.

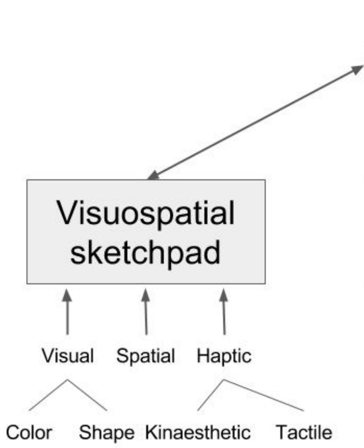


Figure 2.2 Baddeley complete model of Working Memory

- An **episodic buffer** dedicated to linking verbal and spatial information in chronological order. It is also assumed to have links to long-term memory.

In the same model, Baddeley describes the concept of long-term memory, which represents a different storage location to working memory. Long-term memory is presented as being unlimited in space and is responsible for storing information that is no longer in working memory. Typically, information moves from working memory to long-term memory by repetition or rehearsal, or through repeatedly processing the same information. Similarly, Wickens [144] described the working memory as the temporary holding of information that is "active", while long-term memory involving the unlimited, passive storage of information that is not currently in working memory.

Using this model as a foundation, we can develop tasks to target various components through different task encodings, allowing us to investigate whether measurement techniques can detect them. 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.

Although Baddeley's model could be used for the understanding of the processes involved during interaction with technology, it is limited to the decomposition of tasks based on the either verbal or spatial encoding. There are associated limitations with this model, when one considers tasks that are highly complex, involving both verbal and spatial encodings. Moreover, Baddeley's model does not provide a good understanding on the way information is manipulated during interaction. However, one of the model's most important limitations is the one regarding the central executive. Although it is the most important component of the working memory system we know considerably less about this component than the two subsystems it controls.

Having a critical role of managing attention of the working memory, when comes to tasks that are performed simultaneously, such as secondary task techniques, a conflict may arise when the tasks require more - or close to - the maximum available resources of the working memory. An example can be a driving task, where the driver decides to set up the satnav during driving. During normal driving conditions the driver can cope with both tasks, however, when additional resources are required due to a hazard on the road, the primary task may be affected, resulting in poorer performance in the primary task rather than instantly abandoning the secondary task. Although we used the working memory model to understand the cognitive processes during the driver scenario, this shows once again that primary and secondary measures are affected by trade-offs and strategies and the risk of denaturing the primary task.

A particular attention should be drawn when discussing secondary tasks in relation to the phonological loop. As previously discussed, the phonological loop it is that part of the working memory model that is responsible with dealing with sound or phonological information. Macken and Jones [86] further discussed the two parts of the phonological loop: the short-term phonological store dealing with auditory memory traces and an articulatory rehearsal component (sometimes called the articulatory loop) that can revive the memory traces. The phonological store acts as an "inner ear", in a way responsible with remembering speech sounds in a temporal order, whilst the articulatory process acts as an "inner voice" responsible with repeating the series of words in a loop in order to prevent the lose of information [9, 86]. Therefore, performing tasks simultaneously may cause effects such as articulatory suppression - the process of inhibiting memory performance by speaking while being presented with an item to remember [2]. This effect is known to be caused as the articulatory rehearsal processes are being blocked by the irrelevant speech, leading memory traces in the phonological loop to decay [78].

A similar interference effect (to the articulatory suppression) during serial recall is "tapping on a specified point" [71, 117]. Both types of secondary tasks can be used to understand the mechanisms underlying working memory by overloading domain-specific resources [2].

In addition to the working memory model, we consider the Information Processing Model [145] and Multiple Resource Model [144] proposed by Wickens.

2.1.2 Information Processing Model (IPM)

WM model ties well with what were originally described as models of information processing. One shared characteristic between the two, is the limited capacity that we have as human beings, meaning that we can only process a limited amount of information at any one time. Two of the early known models are Welford's [140] and Whiting's [141], both models reflecting the same process, however, using slightly different terminology. Further work developed by authors such as Kahneman [74] and Wickens [62, 144] refined the models, having important implications for the definition and measurement of mental workload. Limited processing capacity was replaced by the term attentional resources, which have to be shared between a number of psychological processes such as perception, WM, and response execution [121]. Wickens describes how 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. These are illustrated in the general model of human information processing proposed by Wickens (see Figure 2.3).

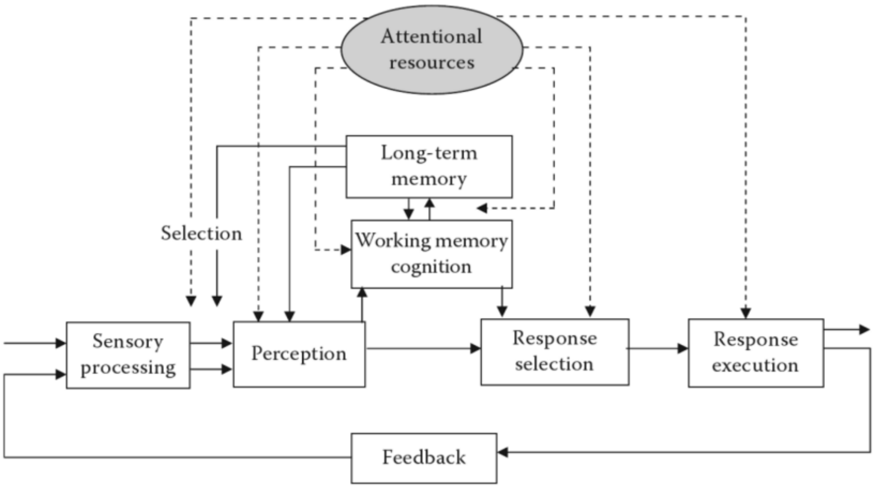


Figure 2.3 A general model of human information processing - from [121] (Adapted from [62])

In a later model, Wickens describes the need of three different 'stages' (see STAGES dimension in Figure 2.4) 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. The response is described based on the modality of its nature.

2.1.3 Multiple Resource Model

The Multiple Resource Model (MRM) proposed by Wickens [144] illustrates how resource limitations and coordination affects the interrelation of mental workload in tasks. MRM is illustrated in Figure 2.4. We are interested in observing how and when these elements interconnect under various tasks that users perform. The elements of this model overlap with the needs and considerations of evaluating complex tasks. Wickens describes the aspects of cognition and the multiple resource theory in four dimensions: STAGES, MODALITIES, CODES and the VISUAL PROCESSING (see Figure 2.4).

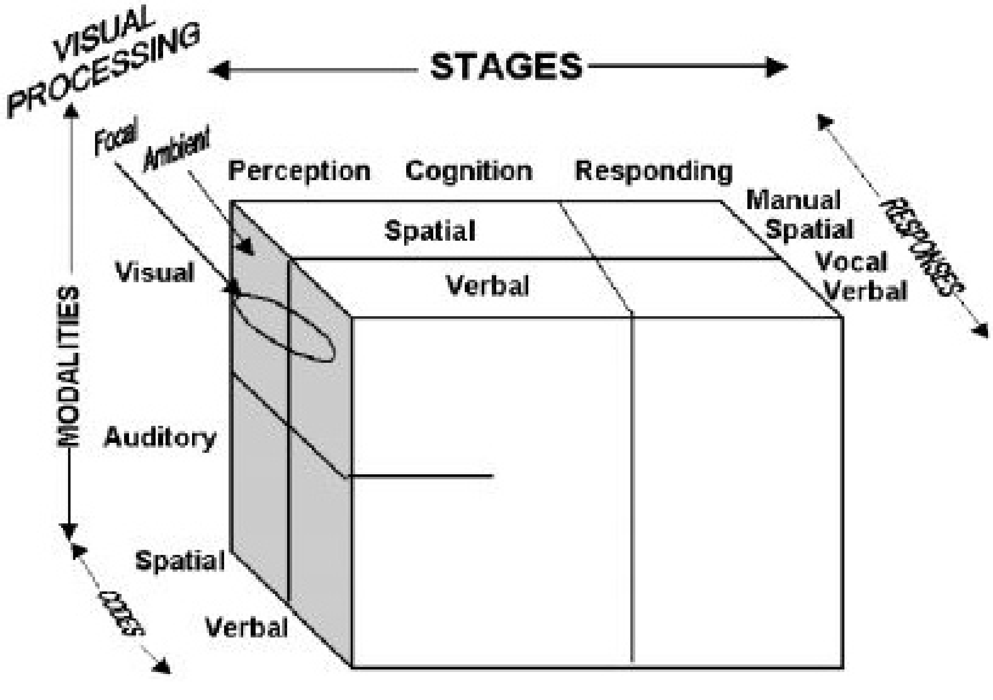


Figure 2.4 The 4-D multiple resource model, by Wickens

- The STAGES dimension refers to the three main stages of information processing system as described above (Figure 2.3).
- The MODALITIES dimension indicating that auditory and visual perception have different sources.
- The CODES dimension refers to the types of memory encodings which can be spatial or verbal.
- The VISUAL PROCESSING dimension refers to a nested dimension within visual resources distinguishing between focal vision (reading text) and ambient vision (orientation and movement).

One of the key roles of the MRM is to demonstrate the hypothesised independence of modalities and use this to design tasks. Our aim is to understand how elements of MRM link together and compose more complex components/tasks. On the other hand, we want to consider how complex tasks can be divided into primary components according to the models described, in order to better understand task factors impacting the demands placed upon operators and workload. These will help identify possible problems in design as well as indicate solutions such as (suggested implications by Wickens [145]):

- Minimize working memory load of the interactions and consider working memory limits in instructions;
- Provide more visual echoes (cues) of different types during interaction (verbal vs spatial);
- Exploit chunking (Miller, 1956 [95]) in various ways: physical size, meaningful size, superiority of letters over numbers;
- Minimize confusability;
- Avoid unnecessary zeros in codes to be remembered;
- Encourage regular use of information to increase frequency and redundancy;
- Encourage verbalization or reproduction of information that needs to be reproduced in the future;

- Carefully design information to be remembered;

2.1.4 Limited Resource Model (LRM) and Mental Workload

Mental workload is a concept that refers to the amount of resources and necessary "effort" required by all the processes mentioned above in relation to a task, and the demands required by the task. Sharples and Megaw [121] described the effect of workload as *"the relationship between primary task performance and the resources demanded by the primary task"*.

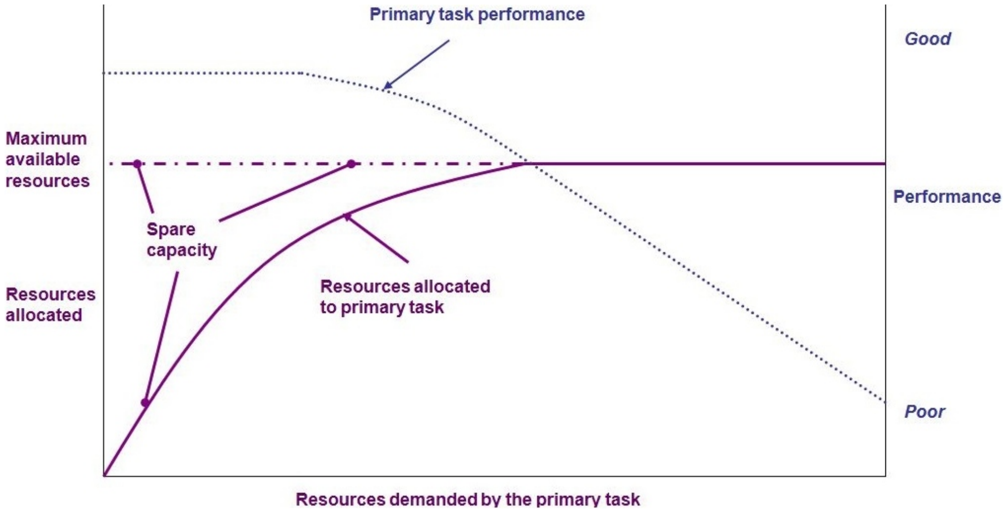


Figure 2.5 The relationship between the resources allocated to the primary task and the resources demanded by the primary task, and the relationship between primary task performance and the resources demanded by the primary task (Adapted from Wickens, C. D.,et al [146])

The Limited Resource Model (LRM) in Figure 2.5 [121], presents the concept of workload as the relationship between the resources allocated to the primary task, and the resources demanded by the primary task, and how performance is "affected" at different stages of demand. The vertical axes on the left indicates the resources being used by the task, but also points out that these resources are limited, having a maximum level of available resources. The vertical axes on the right indicates the performance of the primary task, and the horizontal axes indicates time. When task demands increase, more resources need to be allocated (therefore the spare capacity decreases). When allocated resources reach a point near the maximum available resources, a drop in performance is expected as the operator cannot cope with the task demands.

Sharples and Megaw further contributed to the original model (see Figure 2.6) by identifying three (rather than one) points where performance can be negatively affected in relation to workload: the impact of underload on performance, the dip in performance whilst there is still spare capacity due to data limitation, and the less graceful decline in performance due to reaching the maximum available resources - overload. Later in this thesis, we will investigate how providing feedback of workload for the two extremes, namely underload and overload - could support operators during tasks.

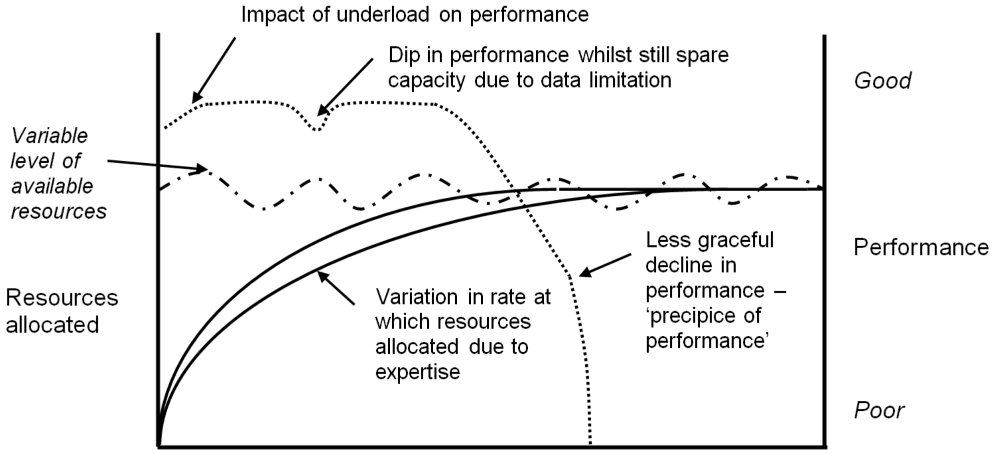

Impact of underload on performance
Dip in performance whilst still spare capacity due to data limitation
Good
Variable level of available resources
Performance
Less graceful decline in performance – 'precipice of performance'
Resources allocated
Variation in rate at which resources allocated due to expertise
Poor

Figure 2.6 Further developed model presented by Sharples and Megaw [121] - The relationship between the resources allocated to the primary task and the resources demanded by the primary task, and the relationship between primary task performance and the resources demanded by the primary task (Adapted from Wickens, C. D., et. al. [146])

2.1.5 Other related conceptualisations

The term *mental workload* is interpreted differently by different researchers in different disciplines. *Cognitive Load*, for example, is popularly used in Educational disciplines to evaluate the effectiveness of learning resources [41]. Like mental workload, Cognitive Load accommodates different modalities, limited capacity, and the role of effort required for comprehending both the task and the materials used to achieve it. While the term Cognitive Load suggests a consideration of the cognitive aspects of a work task, Sharples and Megaw point out that its origins of use are in laboratory-based problem-solving tasks, with a cognitive psychology approach to the issue of load. Conversely, mental workload is used in relation to real-world tasks or jobs, *"where expertise, memory, attention, situational awareness and social and organisational factors all combine to contribute to the individual's experience of workload and thus the concept of mental* *workload needs to reflect the real-world complexity"* [90]. Given the strong empirical validation for Mental Workload in the Human Factors community, we chose to ground our work in this framework rather than in Cognitive Load.

2.1.6 Syntactic VS Semantic Workload: the relationship between workload-task and workload-interface

Most of the workload literature in this thesis presents workload as a relationship between a task in hand, the demands placed upon an individual accomplishing the task, and the associated effects in workload, performance and the individual's physiological changes due to the efforts expelled in completing the task. An interesting perspective of workload is presented by Girouard et. al. [48], who separates the workload concept into two components, based on Shneiderman's theory of semantic and syntactic components of a user interface [124]. Shneiderman's theory discusses the efforts expended by the user to complete a certain task, and this is denoted as the semantic component of a user interface, and the effort associated with operating the user interface, denoted as the syntactic component of the user interface

Based on this principle, Girouard et. al. [48] proposed de-composing the workload required to perform a task using a computer into a portion attributable to the difficulty and demands of the task in hand, plus a portion attributed to the means of interacting with the task, operating the user interface (see Figure 2.7).

Figure 2.7 shows how the approach presented above could be used to evaluate user interfaces. Interface A and Interface B, they both share the same task, therefore the workload generated by the task itself remains constant between the two interfaces. However, they have different hypothetical ways of interaction during task completion. Therefore the difference in workload between the two, must be based on the workload associated with the user interface. Depending on the required need for the task, a high/low workload interface might be suitable. A low workload associated to the user interface (Interface B) is typically preferred, however, a high workload is not always associated with negative effects (for example if the task demands are low and tend to lead the user towards boredom, a high workload interface might be considered).

Interface A, presented in Figure 2.7 is associated with higher workload caused by the interface itself, and not by the task. Interface B would be typically the preferred

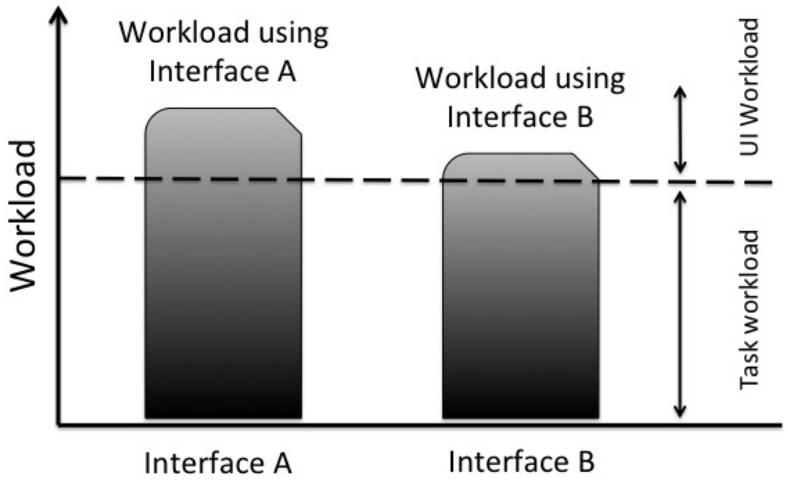


Figure 2.7 Separating the workload generated from the task and the workload generated by the user interface (UI)

option.

Lukanov et. al. [85] used a similar approach to evaluate the workload of three versions of a web form filling process. The workload itself was measured using both physiological and subjective techniques.

——————————————————————————-

2.2 Framework for Mental Workload Measurement

To better understand the implications for the actual measurement process of mental workload in relation to the studies presented in this thesis, we will present a framework for mental workload measurement in Figure 2.8, discussed in [121]. The framework consists of three main components: the physical and cognitive task demands; the operator's workload/effort; and performance. The relationship between these three components, as well as the external and internal influences on workload are the "essence" of mental workload definition and measurement.

Physical and cognitive task demands reflect the characteristics of a task undertaken by a person, and thus, imposed on a person. It is therefore important to quantify work

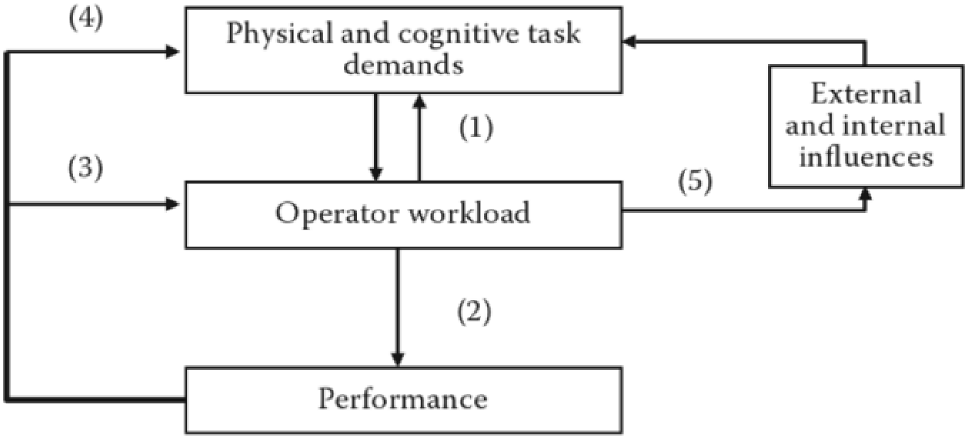

(4)
Physical and cognitive task demands
(1)
(3)
External and internal influences
(5)
Operator workload
(2)
Performance

Figure 2.8 A Framework for mental workload definition and evaluation [144]

demands in the context of mental workload measurement. As the demand may have different impact on different individuals, it is important to not only capture the externally imposed demands, but also consider measuring the perceived demand.

Operator workload is concerned with an operator performing a task, and it is "equivalent to measures of operator strain or effort" [121]. A lot of workload measurement techniques are therefore designed to capture the operator's perceived experience during and after the task via subjective questionnaires, but measures of effort from behaviour indices and the impact of effort on physiology are often used. Both subjective and objective measures for mental workload will be discussed later in the section.

Performance refers to the measures often described in terms of speed and error rates. However, performance measures can become problematic as task complexity increases, but also when closely analysing the relationship between the three workload components of demand, workload and performance; "contrary to what is expected, as task demands increase there is not necessarily an increase in operator workload, or decrease in task performance" [121]. Sharples and Megaw, present an indication of why such simple relationships do exist between the components (see the 5 relationships numbered in Figure 2.8):

1. Operator workload is influenced depending on how the task is perceived by the operator, and it is not just a simple relationship between demand and workload. It can be seen as a consequence of demand created by not only the task demands, but also by a combination of physical and cognitive task demands and external and

internal influences [106]. Pickup et. al. presents workload as a being influenced by intrinsic factors such as operator skill, amount of training, and attitude towards a task. Therefore, these intrinsic factors can influence the strategies the operator can take towards performing a task and the workload perceived by the operator, and indicate that measures of operator workload will not necessarily reflect an objective measure of task demand.

2. Although there is an expected relationship between operator workload and performance (as presented in Figure 2.6), with higher workload associated with relatively poorer performance, this is not always the case. Sharples and Megaw discuss that even with a highly sensitive performance measure, it is likely to be unable to determine differences in how hard an individual is working in order to maintain a good level of performance.

3. Feedback. Both the unconscious and explicit, operators monitor their own performance at every stage during task completion. This may change the way they perceive a task, but also their decision making, strategy and attitude towards a task. The final aims of this thesis will be focused on understanding the impact of not only performance feedback, but also workload feedback on operators, therefore feedback is further discussed later in Chapter 6.

4. Performance outcomes may impact task demands. An error that may have occurred due to high demands on the operator, can lead subsequent task demands to increase, thus further increasing the demands placed upon the operator. On the other hand, a good performance may lead to lower demands on subsequent tasks.

5. Sharples and Megaw present this relationship, and the whole framework in the context of work, and therefore describe most of the external and internal influences as factors from a workspace perspective. In this thesis, we will discuss and consider the external and internal factors in terms of the task and the settings of the tasks, to better state hypothesis, but also understand and interpret the results. Therefore we will consider external factors that will influence the operator's behaviour and experience or perception of workload, but also internal factors, skill and motivation when drawing hypothesis and conclusions in relation to this thesis.

In this thesis we will use this framework to understand the underlying processes that happen during the studies presented in relation to mental workload measurement. We will use relationship five presented in the framework to break down and control the external factors in each study, to better understand their impact on various aspects of workload.

2.3 Measuring Workload

Measuring workload has been a challenge, and multiple attempts from different authors have been made to establish the appropriate criteria for the measurement of mental workload.

Measures such as primary task performance, secondary task performance, and subjective ratings are commonly used techniques of assessing workload. Subjective ratings are usually obtained after the task has been completed, commonly missing essential information about user's experiences during the task. In this section, we will discuss various empirical measures which can be divided into primary and secondary measures, subjective and psychophysiological measures.

There are a variety of subjective and objective methods used for assessing workload including performance measures (subdivisions of primary and secondary task measures) [91], physiological or psychophysiological techniques [42, 59, 76, 108, 127], as well as self-assessment or subjective rating scales [55, 72]. As workload is such a complex concept, lacking of a clear definition, it has already been shown that different "accepted" measures capture various aspects of workload, and the measures usefulness may vary depending on their application.

Cain [26] suggests that even though questionnaires and interviews are informative techniques, they can not be classed as workload measurements as they are "complex to design to avoid unwanted biases, awkward to validate, and difficult to generalize". [133].

2.3.1 Criteria for the Mental workload measurement

One of the hot topics in workload literature is exactly what makes a good measurement of workload, and a possible criteria for evaluating mental workload measurement techniques, in order to separate good from bad measures of workload.

Multiple authors have considered the following factors that contribute towards the assessment of workload measurements techniques [42, 62, 121]:

- **Sensitivity:** Evaluating the technique in detecting changes in task difficulty or task demands experienced by an individual.
- **Diagnosticity:** Evaluating the technique in not only detecting the changes in workload, but potentially also the reason for why workload changed.
- **Selectivity:** The capability of a technique to identify changes in workload in terms of cognitive demands, rather than other variables not directly related to the change in workload (e.g. emotional stress, physical workload)
- **Validity:** Ensuring the measure is actually detecting changes in workload. One way of doing this is by exploring it's relationship to other workload measurement techniques.
- **Reliability:** Techniques used for measuring workload should provide consistent results, that are reproducible and replicable (R (reproducing similar study conditions and stimuli, techniques should have consistent results).
- **Intrusiveness:** The techniques used for measuring workload should be intrusive to the task, not interfering with the primary task performance.
- **Subject acceptability:** This refers to the participants' perception of validity and usefulness of the measurement technique.

In addition to the above mentioned considerations, there are other things that could be considered when discussing ways to evaluate or value workload measurement techniques such as:

- The costs in time and effort towards using a particular technique
- The flexibility of the measurement to accommodate the assessment of workload in various environments (lab-based techniques vs. more naturalistic environments (in-the-wild approaches)).

- The type of workload captured: whether it is a short-term vs. long-term workload captured; identifying if a momentary or overall workload was captured using a particular technique.

- Aspects of workload captured: some of the techniques for assessing workload might capture different aspects of task workload, as presented in Figure 2.8); thus employing more than one measure may be more helpful for diagnostic purposes.

- Adaptability of the technique: Some workload measurement techniques might allow the use of other workload measurement techniques simultaneously, while others might restrict it.

- Relativity of the measure: There is no standard "absolute" rage of high vs low workload, and this is due to the relative, nature of the workload, being influenced by multiple factors of the individual. However, a typical "baseline" of low workload task could be used in order to class various levels of workload.

2.3.2 Primary and secondary task measures

Primary measures rely on measures of primary task performance to predict workload. As previously discussed, performance measures can reflect performance, however, they have limitations to the extent of using primary measures alone: it is difficult to discriminate between levels of effort the operator is going through while the demand changes and performance does not. Consequently, primary task measures should be combined with other workload measures. Secondary techniques involve the inclusion of a additional task to the primary one. The secondary measures are used in cases where the primary task demands would allow enough available resources (see Figure 2.6) for a secondary task to be completed concurrently. In this case, the secondary performance measures can reflect the amount of workload imposed on the operator.

2.3.3 Subjective Measures of Mental Workload

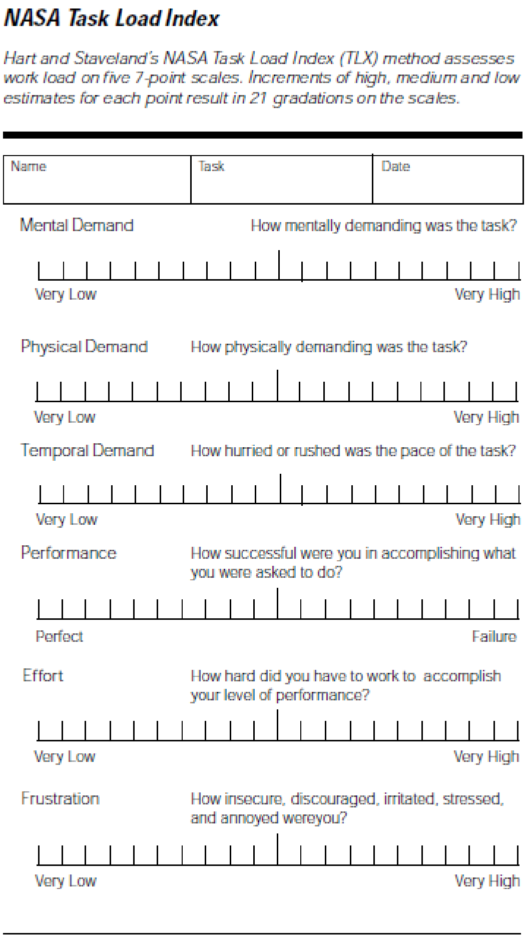
Even though much effort has been invested in developing objective measures of workload, subjective techniques are still the most popular due to their acceptance (high face validity), easy to administer, non-intrusiveness and low cost. Subjective measures of workload attempt to capture the users reflection and perspective on how much effort was expelled and perceived by the operator during the task completion. Subjective measures are significant, important tools in evaluation, used for assessment of operators'/users' workload, but not only. In the context of workload, subjective measures cannot be underestimated, especially when it comes to evaluating more complex systems and tasks, where performance based measures become highly complex to measure [115]. Even though subjective measures are critical for the evaluation of these systems, there are certain limitations that cannot be overseen. Casali and Wierwille [27] point out that "... properly designed rating scales with associated instructions ... are particularly sensitive measurement instruments, especially with highly-trained populations ...". Cain says it is rather "appropriate" for mental workload to be measured using subjective means, as it is a psychological construct, however, Gopher and Donchin suggest "... an operator is often an unreliable and invalid measuring instrument" [49].

Subjective measures do rely on the operator's ability to self-judge and report the state throughout the task. This requires not only extra effort from the operator, but also skill and potential training. Moreover, if used in real-time, subjective techniques may interrupt the operator performing the tasks in hand, and negatively affect performance; if used post-task, they rely on the operator's ability to recall what happened during certain moments in the past.

When applying a subjective rating scale for assessing workload, participants are typically instructed that the scale should be used for reporting their perceived workload based on mental rather than physical work. However, some scales such as NASA-TLX (presented below) have a dedicated sub-scale for the Physical Demand users go through during the task. Depending on the need for the evaluation, the subjective ratings of workload could be categorized into uni- and multidimensional scales, retrospective and concurrent measures.

The most widely used techniques for subjectively assessing workload include NASA Task Load Index (TLX) [55] and the ISA Instantaneous Self Assessment [21, 72, 73] used in the studies presented in this thesis, but also others such as Subjective Workload Assessment Technique (SWAT) [113], The Integrated Workload Scale (IWS) [106], and Workload profile [135].

NASA-TLX



NASA Task Load Index

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

Name	Task	Date

Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low Very High

Figure 2.9 NASA-TLX workload assessment technique.

NASA-TLX (Task Load Index), developed by Hart and Staveland [54, 55], is an example of a widely used subjective measure of mental workload. NASA-TLX is a multi-dimensional tool -see Figure 2.9- that uses perceived workload ratings in order to assess a task after performing it [84, 108, 110]. This makes the measure suitable for providing an overview of the task retrospectively, however it does not provide insight into users workload at a given moment during the task.

The NASA-TLX questionnaire is used to capture participants' subjective workload as a self-assessment [55], based on the weighted average ratings of six subscales including, in order: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort and Frustration. In a typical study, participants are asked to self rate their mental workload using the NASA-TLX questionnaire once after each study condition.

The tool has been successfully used across a wide range of domains, including civil and military aviation [14, 80, 130], driving [3, 53], power plant control room operation [63, 66], and air traffic control (ATC) [39, 93].

ISA

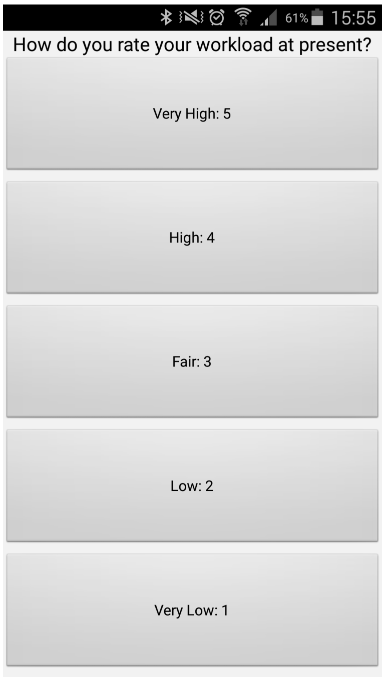
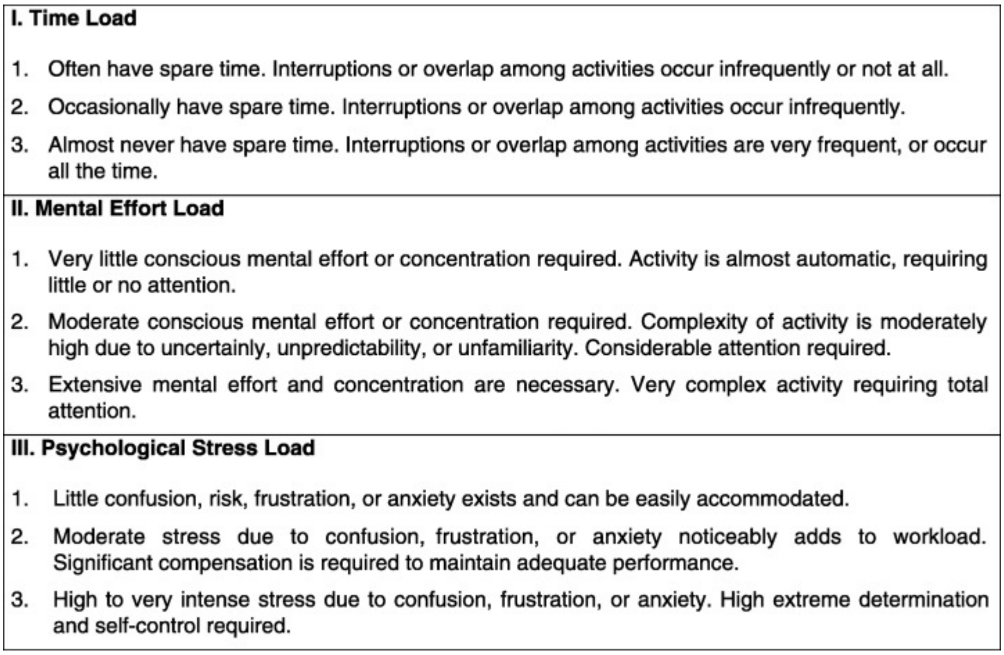

How do you rate your workload at present?

Figure 2.10 Instantaneous Self Assessment (ISA) Recorder App

Concurrent alternatives to NASA TLX exist, including ISA (Instantaneous Self Assessment), developed by Jordan and Brennen [21, 72, 73] and have been validated as being a reliable workload measure [81, 133]. ISA, presented in Figure 6.4, derived for use within an Air Traffic control setting, and is one of the most frequently used measures of workload in real-time simulations, being measured using a five-point rating scale to provide immediate subjective ratings of work demands during the performance of primary work tasks.

Users are prompted at regular time intervals during the task to rate their current workload levels. There are, however, questions as to the interference caused by the measure, with conflicting findings in the existing literature: there are cases of detectable [133] and non-detectable [81] task intrusions. Regardless, however, the measure has been considered preferable to other measures, such as Subjective Workload Assessment Technique (SWAT) [112], and Workload profile[135].

SWAT



I. Time Load

1. Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.
2. Occasionally have spare time. Interruptions or overlap among activities occur infrequently.
3. Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.

II. Mental Effort Load

1. Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention.
2. Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainly, unpredictability, or unfamiliarity. Considerable attention required.
3. Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.

III. Psychological Stress Load

1. Little confusion, risk, frustration, or anxiety exists and can be easily accommodated.
2. Moderate stress due to confusion, frustration, or anxiety noticeably adds to workload. Significant compensation is required to maintain adequate performance.
3. High to very intense stress due to confusion, frustration, or anxiety. High extreme determination and self-control required.

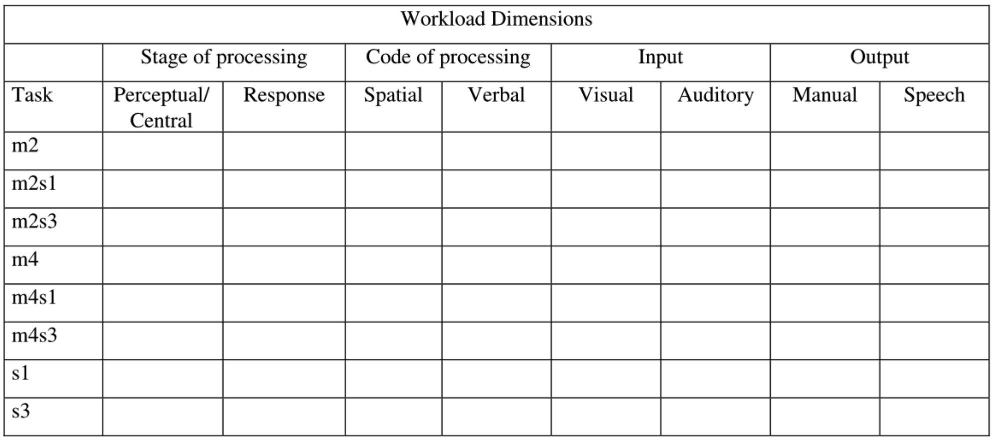
Figure 2.11 The SWAT workload assessment technique.

The Subjective Workload Assessment Technique was developed by Reid and Nygren (1988) [113], and it is a subjective measurement technique of assessing workload while performing a task. The technique is based on three levels of operator ratings: (1) low, (2) medium, and (3) high, for each of the following 3 dimensions (also see Figure 2.11):

- time load,
- mental effort load,
- and psychological stress load.

Workload profile

Workload profile of Tsang and Velazquez (1996) [135] is yet another technique to asses workload subjectively. The multidimensional tool is based on Wickens' Multiple Resource Model [144] presented above, and it attempts to use the benefits of secondary techniques combined with subjective techniques in order to capture operators' attentional resources used during a series of tasks. Workload profile (see Figure 2.12) requires operators to rate the the proportion of attentional resources for each task individ-



Workload Dimensions								
	Stage of processing		Code of processing		Input		Output	
Task	Perceptual/ Central	Response	Spatial	Verbal	Visual	Auditory	Manual	Speech
m2								
m2s1								
m2s3								
m4								
m4s1								
m4s3								
s1								
s3								

Figure 2.12 The workload profile technique for assessing workload.

ually using the dimentions used by Wickens in the Multiple Resource Model. Operators tasks in a random order, and only when all required tasks are completed.

Physiological measures

The physiological techniques of assessing mental workload stand on the assumption that as workload is increased, there will be a corresponding effect (increase/decrease) in the person's level of arousal, measurable and reflected into the activity of the autonomic nervous system, and therefore reflected in a number of physiological parameters. With a controversial term such as mental workload, and with no general relationship between various physiological measures and workload, it has been shown that certain measures capture different workload components [121]. As physiological parameters do not involve any extra work to be done by the users during interaction (such as subjective measures), and have the potential to continuously monitor fluctuations in workload throughout a long period of time, there are clear advantages in favour of using these techniques. Moreover, if the environment studied would be a "in the wild" as part of a natural critical system job such as Air Traffic Control, subjective techniques and secondary performance techniques would create additional demands on the individuals, potentially generating a negative impact on performance. This is not acceptable in such environments, and therefore the use of non-invasive techniques could allow the study of individuals at work, even in such critical scenarios. However, subjective techniques are still the most widely used in industry, hence the significance of further exploring physiological measurements for assessing workload.

2.4 Chapter Summary

This chapter explored concepts surrounding the structure of workload, and presented some of the most used techniques for measuring workload. Chapter 3 explores in better detail how physiological techniques work, and present the latest, most widely used physiological techniques for assessing mental workload.

Chapter 3

Measuring Physiological Workload: Background and Methodology

This chapter aims to present the second part of the literature and related works, describing how physiological measures work. It begins by describing the connection between workload and physiology, laying and discussing the foundations of physiology. Further sections will then present various methods to capture human physiological responses to mental workload, focused on brain monitoring techniques. The last part of the chapter presents fNIRS measurement methodology in relation to mental workload, and the techniques used to process fNIRS signal in relation to mental workload.

3.1 Physiological measurement of workload

Research has shown that task demands can induce complex and dynamic processes influencing a variety of physiological changes [44]. Physiological measures are used to give an objective perspective on mental workload by not relying on subjective scales or performance measures. They can be obtained by recording cardiac activity [18, 19, 51, 100, 129, 132, 148], electrodermal activity [31, 122, 123], eye function [15, 67, 74, 88] or imaging the brain [19, 58, 88, 96, 127]. These techniques detect the change in the arousal from the autonomic nervous system level which could be used to infer mental workload levels. Stemberger et al [128] developed and evaluated a system for estimation of workload levels based on analysis of facial skin temperature. Previous studies that have looked at inferring the level of mental workload by using facial thermography have shown a high correlation of workload with the decrease in nose temperature [102].

Wilson and Russell [148] presented ways in which the physiological signals are known to change with the state of the operator: heart rate increases as the cognitive demands on the operator increase, and the rate of eye blinking decreases as the visual demands increase. Different physiological measures capture different aspects of workload [26], therefore consideration should be put in choosing the most appropriate measure for the given studied task, setting, and research interest.

Beyond just objectively assessing workload using physiological measures, research has shown that mental workload itself can be valuable as an input into a system [127]. Afergan et al [1], for example, demonstrated the ability to distinguish and use different mental workload states during an Unmanned Aerial Vehicle simulation task by looking at the physiological changes in the brain. These measures were used in real time as an input to dynamically adjust the difficulty of the task in order to improve user engagement and performance. With a similar aim, but without adapting the task or the interface, the last study presented in this thesis is focused on exploring a new area of research, providing real time workload feedback to alert users of low and high moments of mental workload, in order to help users reflect metacognitively, adapt to the task, focus and better manage when prioritizing resources to tasks. We attempt to understand how people can use this feedback, and the effects this may have on people during tasks, as described in the relationship 3 in the framework for the measurement of mental workload.

3.1.1 Prefrontal Cortex (PFC) and Brodmann area 10

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 [75]. The PFC has been first associated with higher cognitive functions by studies examining brain damaged individuals [119, 125], however, experiments on healthy subjects using the n-back task have supported this claim [20]. Activation was observed in the dorsolateral prefrontal cortex, inferior frontal and parietal when the task demanded more resources. However, it is difficult to point out which brain region is involved with which processes because one brain area is usually involved in multiple cognitive tasks [23]. Miller and Cohen define the representations in the PFC as "attentional templates, retrieval cues, rules, or goals” [94], and many researchers agree that PFC function is one of Baddeley’s executive control [4]. Conversely, Rushworth reports that not all PFC subregions are essential for working memory [116]. The PFC, more specifically the Brodmann area 10 (BA10), is the region of the participants’ brain that we are targeting using fNIRS during the studies presented in this thesis (See Figure 3.1), since there is significant evidence to support its role in WM [20, 36]. Brodmann area 10 is thought to be the gateway between processing perceptual or reflective information. In addition to the PFC, Brocas area is located within the frontal lobe and is linked with speech production [43].

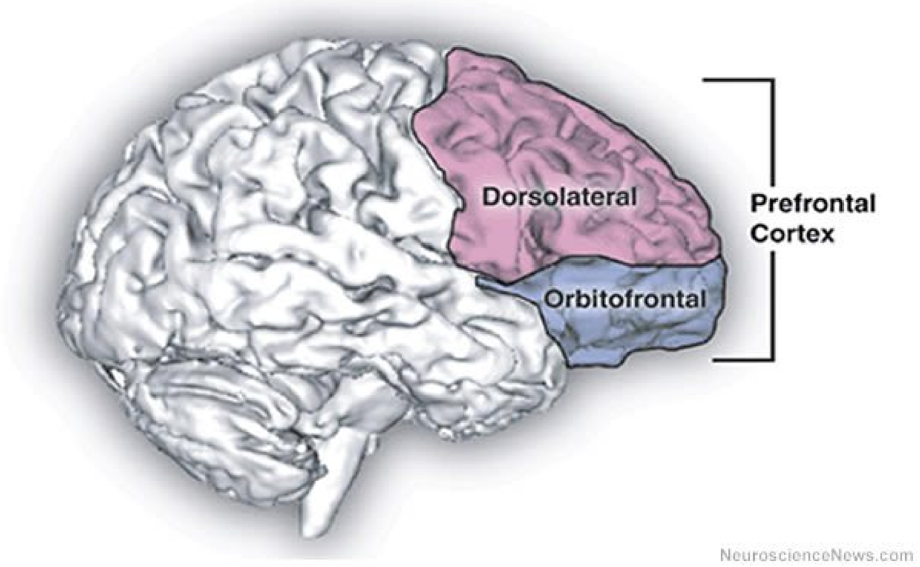


Figure 3.1 A brain drawing with the prefrontal cortex highlighted [1]

3.1.2 Brain sensing techniques and BCI

Brain sensing technologies were first created for medical and diagnostic use, however it has been only recently that advances in cognitive neuroscience and brain imaging technologies have made a new era possible: interfacing directly with the human brain. Firstly driven by the needs of people with disabilities, researchers have used these technologies to build Brain Computer Interfaces (BCI), systems where users explicitly manipulate their brain activity as an alternative interface to the outside world. Recently, researchers in the field of HCI, have adopted BCI techniques to learn more about peo-

[1] https://i1.wp.com/neurosciencenews.com/neuroscience_images/prefrontal-cortex-public.jpg

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 3.1 Summary of Brain Sensing technologies

ple during interaction with technology [1, 85, 127, 152]. It is agreed, however, that to be valuable for HCI, measures and techniques used for measuring workload should provide useful information about the user while allowing normal interaction with the computer [7, 108, 127, 152]. Pike et. al. presented a comparison between various BCI techniques, with a focus on their suitability in HCI settings [108].

Brain imaging technologies can be therefore classed in two general categories: invasive techniques, and non-invasive techniques. Depending on the level of invasiveness, the sensors can range from being implanted directly on, or in the brain, putting a human subject inside a sensor (fMRI), to less invasive techniques that only require wearing a headband shaped device with external sensors only [131]; the later ones being preferred for studying the interaction with technology.

There are several brain sensing technologies available for research, including (but not limited to) fMRI, EEG, and fNIRS, which are summarised in Table 3.1. Each of these technologies have different strengths and weaknesses, as discussed by Tan and Nijholt [131]. In this thesis the focus is on technologies for assessing brain activity using external sensors only.

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. Experiments that have used fMRI for studying the interaction with technology, typically place a mirror above the participant such that they can see a display in another room. Participants are unable to interact directly with a system, but can respond to visual stimuli through the use of mirrors. Li et. al. [82] 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[2], EEG has become an affordable option for brain sensing [40]. Researchers have successfully utilised EEG as a form of input, rather than evaluation, as major events can be more easily detected as triggered (see [134]). For evaluation, however, EEG data is susceptible to motion artefacts, and so producing averages for periods of interaction provides limited insight. Pike et al [107] proposed, that EEG data was most valuable when combined visually with recorded think aloud data, as statements of confusion, or pauses in verbalising ones' actions, coincided with and were qualified by the EEG data.

fNIRS - Functional Near Infrared Spectroscopy - 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 [60, 79, 105]. Because it takes several seconds for blood to flow to the brain [65, 137], fNIRs has been largely discounted for real-time interaction with systems, or for direct control during active BCI.

3.2 fNIRS

fNIRS is an emerging neuroimaging technique that offers a non-invasive, portable and low-cost method of monitoring brain activity. fNIRS is based on the use of near infrared spectroscopy (NIRS), introduced by F.F. Jobsis in 1977 [70], which started to be used later for functional brain imaging [28, 136, 137]. Around 1990, fNIRS started to increase its popularity, and it was further introduced to overcome the limitations of using EEG and other techniques, mainly due to its non-invasive nature, allowing a more naturalistic study setting [29].

fNIRS uses near infrared (NIR) light to measure regional hemodynamic responses associated with neuron behaviour, namely changes in blood volume and oxygenation. This is possible due to the properties of our biological tissue, in our case the skull, that is relatively permeable to electromagnetic (EM) radiation of different frequencies

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

and intensities. Light penetrates the skull well at near infrared range, allowing the NIR light to reach different molecules that are known to absorb different wavelengths of EM radiation to different degrees (in this case light). For fNIRS imaging, the concerned molecule is haemoglobin, which is the oxygen carrier for the red blood cells. These are the primary absorbers of near-infrared light in tissues. During the hemodynamic and metabolic processes, the light values change proportionally with neural activity in the brain [29, 68] fMRI studies [37] have confirmed that a decrease in deoxygenated hemoglobin indicates an increase in brain activity. When a brain region becomes active, it requires more oxygen. To meet these demands, there is an increase in oxygenated hemoglobin, resulting in a decrease of deoxygenated hemoglobin.

fNIRS has been successfully used to measure brain activity in different brain regions such as, prefrontal cortex [7], motor cortex[61] and auditory cortex [109].

3.2.1 Using fNIRS to assess Workload

fNIRS was used in a wide range of disciplines, but for the sake of this thesis, the focus will be the use of fNIRS for measuring a range of cognitive processes and states of operators during their interaction with technology in normal working conditions. A few studies in the area include [1, 30, 57, 105, 127, 153]. The general aim of this thesis is to understand how fNIRS can be used in relation to mental workload, to assess and reflect the demands placed upon individuals whilst performing tasks, by observing the relative changes in oxygen concentration in the PFC. A typical fNIRS equipment consists of a probe, a data processing unit for pre-filtering and pre-processing the raw data, a power supply, and two computer units, one used to collect and process the data, and the other to present stimuli to participants (e.g. an Air Traffic Control task).

fNIRS has been used to evaluate various tasks, including remotely operating vehicles [7, 38], mental arithmetic [108], n-back tasks [7, 38], and other complex cognition tasks such as video games [7, 24, 68]. Its robust nature to ambient electrical noise, (that is the main artefact affecting the EEG signal, fNIRS allows a more naturalistic study of interaction compared to fMRI, while measuring essentially the same signals, and the relatively reduced hardware requirements, make fNIRS well suited for BCI systems.

3.2.2 fNIRS during human computer interaction

Recent research has proven fNIRS to be more suitable (compared to other brain sensing techniques) for assessing mental workload in HCI user studies [1, 87, 127, 152] due to the robustness in noisy environments. HCI researchers are concerned with using fNIRS to assess mental workload, and use this as an additional channel of information about users during interaction with technology. Lukanov et. al. used fNIRS to assess workload during usability testing of three versions of an insurance claim form [85]. fNIRS has also been used for implicit input. Afergan et. al., used fNIRS as implicit input to control and dynamically adjust task difficulty based on user's mental workload state [1]. This was the first step towards using fNIRS in the context of adaptive user interfaces. Yeksel et. al. used the same technology for adapting learning during piano lessons [153]. With a similar aim, we investigated how fNIRS can be used as input to provide users with feedback of their mental workload levels in real time during tasks, and how the users can better think about their state.

The BCI studies (including the ones with fNIRS) in human computer interaction move away from traditional psychology experiments, "where one can isolate, manipulate and measure mental workload with great precision" [48], into the use of the techniques for assessing more realistic user interfaces. This enables researchers to study and create better technology and user interfaces as a result of using an objective perspective of interaction, without relying on participants' ability to self-report, or create additional burden to the interaction (such as a secondary task). However, studying more complex tasks and interactions, makes it harder to understand and interpret the data and the results. This is why fNIRS is used as a complementary measure during human computer interaction to provide an additional channel of information about the users during interaction, and often researchers combine the results of various measures together for a better understanding and interpretation of results. The studies presented in this thesis involve a transition of increased level of task complexity, from very traditional memory tasks in the first study, to more naturalistic air traffic control game simulator in the last study.

Peck (2013) [105] found a correlation between the NASA-TLX subjective questionnaire [55] and deoxygenated hemoglobin levels in fNIRS data during a visual task. Peck (2013) [105] successfully managed to distinguish between various levels of n-back visual tasks with fNIRS, suggesting that different levels of n-back induced different levels of mental workload. Confirming with the fMRI findings, fNIRS deoxygenated hemoglobin levels had lower values during 3-back task compared with 1-back task. These are great findings further validating fNIRS as a measurement of workload. We will attempt to validate these results and further understand the sensitivity of fNIRS to various levels of workload.

3.3 fNIRS methodology

Throughout the experiments presented in this thesis, we used fNIRS to assess users' workload state during tasks. This section presents a few insights into how fNIRS was used, including some general processing stages for a typical fNIRS experiment.

We recorded measures of brain activity using an fNIRS300 device and the associated Cognitive Optical Brain Imaging (COBI) Studio hardware integrated software platform 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.

3.3.1 Hemodynamic response

The hemodynamic response is a slow one, and changes measured by fNIRS occur in a time span of 6-8 seconds [25]. It is therefore important to consider and account for the delay when designing brain based interfaces, or simply ant experiments using fNIRS. This is one of the main reasons fNIRS has been largely discounted for direct control in these types of BCIs. Research has been exploring the potential of using event-related fNIRS (for direct control of a BCI) [34], however, most of the studies are taking advantage of the delay in order to monitor the user's state during tasks.

3.3.2 Processing stages

Oxygenated (HbO) and deoxygenated (Hb) hemoglobin 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 [69]. The tissue is radiated by the light sources and the detectors receive the light after the interaction with the tissue. See Fig. 3.2 [6] for an illustration of how the headband is positioned, and to visualise the path that the light follows during operation.

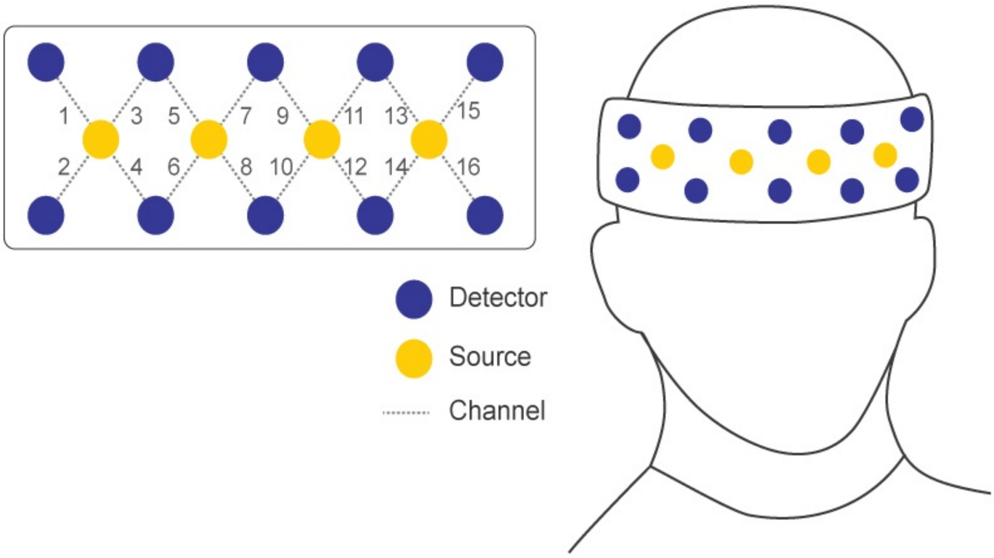


Figure 3.2 Sensor layout for the Biopac fNIRS used [3]

fNIRS data could be used to detect changes in blood flow and oxygenation in real-time, or the data could be processed and analysed after the experiment has been completed. For real time use, COBI Studio, the associated platform for processing fNIRS signal, requires collecting baseline signal levels (typically asking participants to rest), that are used in order to calculate oxygenation in real time, via the Modified Beer Lambert Law (MBLL) [136].

fNIRS data can be further processed and analysed post-experiment, with the advantage of using additional filters and feature extraction techniques discussed below. For the post-experiment fNIRS analysis we processed the data using fnirSoft, the Comprehensive Signal Processing, Analysis and Visualization Platform for Optical Brain Imaging [8]. A low pass filter with cut off frequencies of 0.2 Hz can be used in order to remove high-frequency noise, physiological artefacts such as heartbeats and motion derived artefacts. In addition, the Correlation Based Signal Improvement (CBSI) [35], a technique designed for fNIRS technology, can be used as a filtering method to improve detection of workload. In the analysis, we also considered the delay associated with the hemodynamic response [136], using various techniques including: averages across

[3] Many thanks to Hyosun Kwon for the designs of this image.

blocks of data, omitting the first few seconds of the trials when processing, or simply delaying the trial data by a few seconds [105, 108]. These were further detailed in each study chapter.

For studying mental workload, the device is typically placed on the participants' forehead, targeting the PFC and Brodmann area 10 (BA10) (see Fig. 6.2). As displayed in Fig. 3.2, the Biopac fNIRS device provides 16 channels of brain data reading. Each channel is defined by the relationship between a IR source and a near by IR detector pair. After the MBLL is calculated, COBI studio provides a file with two measures HbO and Hb, and for the post experiment analysis the TotalHb can be calculated (Hb+HbO), and one additional measure of oxygenation (OXY) can be obtained from fnirSoft, that is oxygenation (OXY).

3.4 Thesis Overview

Chapter 2 and Chapter 3 presented the underlying background of this thesis. Various specific aspects of background literature will be provided in chapters when needed. The following chapter presents the proof of concept study, further investigating the use of fNIRS to assess workload during interaction with technology within realistic lab-based evaluation settings, therefore continuing the work of Solovey et. al. [127]. Chapter 5 presents the challenges of using fNIRS as a sensitive and valid technique in the context of continuous, real-time use, to gain insights into mental workload during tasks. The study presented in the same chapter starts the investigation of fNIRS's relationship with subjective workload techniques and performance measurements (that will continue in the last study presented in the following chapter). The last study of this thesis is presented in Chapter 6. The study is focused on using the continuous, real time assessment of workload based on fNIRS objective measurements to provide users with workload feedback during tasks. We then investigate the impact of workload feedback on both, users' performance and perception of workload, but we also assess their experience throughout the experiment and potential future use of the feedback element.

Chapter 4

Using fNIRS within realistic lab-based evaluation settings: Investigating the Reliability of the measure in the presence of artefacts

How can fNIRS be used to assess workload during interaction with technology within realistic lab-based evaluation settings?

4.1 Introduction

This thesis investigates the use of fNIRS as a technique to assess workload continuously during tasks. However, to be valuable for HCI/HF research, sensors and techniques used for collecting data from users should be as transparent as possible, while still providing useful information about them. This is to ensure that the studied interaction would be affected as little as possible, in order to simulate naturalistic study settings.

Because the nature of fNIRS technology was originally designed for clinical use, in this chapter, we present a study to explore the reliability of fNIRS, and how it can be used to assess workload during interaction with technology, within realistic lab-based evaluation settings; therefore an attempt to answer the first research question RQ1 with the sub questions RQ1a, RQ1b, and RQ1c. See the overview Table 4.1 below.

Table 4.1 Experiment 1 - Experimental Program Development

No	Research Questions and Aims	Main findings	What next?
RQ1(a)	Investigate the possibility of using fNIRS’ measures of oxygenated (HbO) and de-oxygenated (Hb) hemoglobin to distinguish between a “busy” state (participant performing a task) and a “rest” state (participant performing no task).	• For both Verbal and Spatial tasks, a paired-sample t-test, within participants, revealed significant differences over multiple channels between rest periods and task periods. • In both task conditions HbO was significantly higher compared to rest states.	• fNIRS was useful to detect moments of high workload (participants performing a cognitive task compared to participants at rest). One of the future challenges is moving beyond comparing between the binary conditions rest and task, towards the assessment of workload during more complex tasks.
RQ1(b)	Investigate how fNIRS can be used in the presence of artefacts produced during interaction with technology within realistic lab-based evaluation settings. Understand the impact of various artefacts on the two different task encodings: spatial task vs verbal tasks. • It was hypothesised that non-related verbalisation will negatively impact performance during the Verbal task, as demonstrated by Pike et al. [108], however, we also hypothesised that performance will not change during the spatial conditions, as the resources used during the spatial task are complementary. • Investigating whether there are differences between the presence and absence of the artefacts in the fNIRS signal. • In addition to reproducing the distinction between rest and Verbal task times (as identified by Solovey et al.), there will also be a significant difference between rest and Spatial task times.	• Participants performed significantly worse under the typing artefact compared to all other conditions during the Verbal task. • Participants also performed significantly worse in the Verbalisation artefact condition compared to the no artefact one during the Spatial task. • For both Verbal and Spatial tasks, a paired-sample t-test, within participants, revealed significant differences over multiple channels of fNIRS’ HbO and Hb, distinguishing between rest periods, task periods and artefact periods.	• The two types of tasks, Verbal and Spatial, were affected differently for each artefact. This was reflected in both performance and fNIRS measures. For the Verbal task, the greatest interference was typing, which could be interpreted as being a Spatial input modality since the keys have a physical mapping. Whereas for the Spatial task, the verbalising artefact had the greatest interference providing a crossing of resource modalities, which is the opposite of our original hypothesis. • Consider the task encoding and artefacts nature in future experiments.
RQ1(c)	Investigate the Reliability, Replicability, Sensitivity, Validity, of the fNIRS measure: Understand the sensitivity of the measure to both spatial and verbal memory tasks; investigate methods to distinguish between various levels of workload using fNIRS.	• Results confirmed Solovey et al. findings and showed the reliability of fNIRS to distinguish between users’ rest states and users performing a Verbal memory task. • The replicability of fNIRS experiments was also tested by reproducing a previous fNIRS experiment.	• fNIRS proved to have potential for detecting periods of performing a task compared to periods where participants were at rest. The results also showed the potential of using fNIRS during human computer interaction lab-based experiments. Further, it would be essential to explore how fNIRS could be used as real-time, continuous measure, to assess workload during tasks. • Therefore, it is also essential to test the sensitivity of the measure to distinguish various levels of workload during tasks, but also validate the measurement in relation to existing workload assessment techniques.

To understand the implications of using fNIRS during lab-based experiments, this work builds on the initial findings of Solovey et. al. [127]. In 2009, Solovey showed that functional near-infrared spectroscopy (fNIRS) has potential value for brain sensing in HCI user studies, being more suitable than other brain sensing techniques such as EEG, PET, fMRI and others. Their research has shown that, although large head movement significantly affects fNIRS data, typical interaction with a computer does not affect the fNIRS measurements. During a *Verbal* memory task, they studied a number of typical artefacts present in HCI lab-based settings, including: large/normal head movement, facial movement, ambient light and ambient noise, respiration and heartbeat, muscle movement, and slow hemodynamic response.

This chapter replicates and extends Solovey's study and aims to examine the *Reliability* of fNIRS, by 1) confirming these prior findings (Solovey et. al. [127]), and 2) significantly extending our understanding of how artefacts affect recordings during *Spatial* tasks, since much of user interfaces and interaction is inherently spatial.

This chapter is based on the "Examining the Reliability of Using fNIRS in Realistic HCI Settings for Spatial and Verbal Tasks" by Horia A Maior, Matthew Pike, Sarah Sharples, Max L Wilson, paper which was presented at ACM Conference on Human Factors in Computing Systems (CHI2015) in Seoul, Korea.

4.2 Reliability of fNIRS

Sharples and Megaw [121] state the appropriate criteria for choosing the technique for the measurement and assessment of mental workload: Validity, Reliability, Generalisability, Sensitivity, Interference, Diagnosticity, Selectivity, Granularity/Bandwidth, Feasibility of use, Acceptability/Ethics and Resources.

Pike et. al. and Peck et. al. [105, 108] provide evidence of fNIRS correlating with NASA-TLX, a widely used measure of MWL (**Validity**). fNIRS is inherently generalisable as it simply measures oxygenation and is not specific to a particular domain (**Generalisability**). Afergan et. al. [1] demonstrated the ability to distinguish between different workload states (**Sensitivity**) during a UAV simulation task. Additionally, the study identified workload changes over time (**Bandwidth**). Pike et. al. [108] identified non-related verbalisations as being a contributing factor to increased mental workload

(**Diagnosticity**). Solovey et. al. [127] demonstrated that fNIRS was able to distinguish between common human behaviours (typing, mouse movement, head and facial movement) and a Verbal Memory task (**Selectivity**). fNIRS has been deployed in a number of studies and has caused minimal **Interference**, with many reporting ecological validity whilst using fNIRS (also demonstrating **Feasibility of use, Acceptability/Ethics** and **Resources**) [1, 108, 127].

In the context of this study, however, exploring the **reliability** of fNIRS during human computer interaction is a focus. As fNIRS is an emerging technology in this field, replicating the findings of existing work is one step towards establishing the reliability of the technology.

4.3 Experiment Design

The aim of this study is to identify the reliability of fNIRS as a measure of mental workload in the context of human computer interaction. This study examines and tests the work of Solovey et. al. [127] and Pike et. al. [108] on *verbal* memory, but significantly extends our understanding of the impact of artefacts on fNIRS measurements, by also examining *Spatial* tasks. Reliability of the measure is one of the criteria identified by Sharples and Megaw [121] as being appropriate for measuring workload. In this study, much of the original procedure was followed, as described by Solovey et. al., however, some of the behaviours under study were removed, in order to focus on the behaviours that had the greatest impact on the fNIRS signal (Typing and Head Movement). More over, a new human behaviour was included in the study - Verbalisation, a common part of a typical evaluation study, that was not investigated by Solovey et. al.

This study also addressed the issue of task types in terms of information encoding, and how various artefacts investigated in this study could potentially affect the fNIRS signal during different task encodings. In the original study, the task of memorising a seven digit number was Verbal since the encoding of the digits would reside within the Phonological Loop of Baddeley and Hitch's model of WM [13]. To extend our understanding in terms of the impact of different task encodings and artefacts on fNIRS, we introduced a Spatial memory task of memorising a 6x6 grid. This task will be encoded in the Visuo-spatial Sketchpad (according to the same model [13]), allowing us to investigate whether there are differences in results according to the encoding type of the task.

4.3.1 Study Conditions

We devised a study focused on investigating three behaviours and potential artefacts, typical during typical interaction evaluation settings, two of them selected from Solovey et. al. (head movement and keyboard typing) and one original (Verbalising). Therefore, we designed a study with four conditions (three artefacts plus a control condition having no artefacts present), which were tested under both task types (Verbal memory and Spatial memory):

C1 Task Only (No Artefact)

C2 Task + Head Movement

C3 Task + Typing

C4 Task + Verbalising.

The same repeated measures, within-participants approach was followed to compare conditions, as in Solovey et. al. [127].

4.3.2 Study Task - Verbal and Spatial Memory

As previously described, there were two devised tasks in this study, a verbal memory task and a spatial memory task. The verbal memory task involved memorising and reproducing a series of 7 digit numbers, see below the study protocol and procedure, and it closely followed the task presented in Solovey et. al.

The second task, was meant to be similar in terms of demand and complexity, however, having a spatial encoding. Therefore we devised a spatial memory task, where participants were asked to memorize and reproduce a series of 6x6 black and white shaped grids using an on-screen form (See Figure 4.1). Similarly, in the Verbal task conditions, participants submitted the memorized number using an on-screen form.

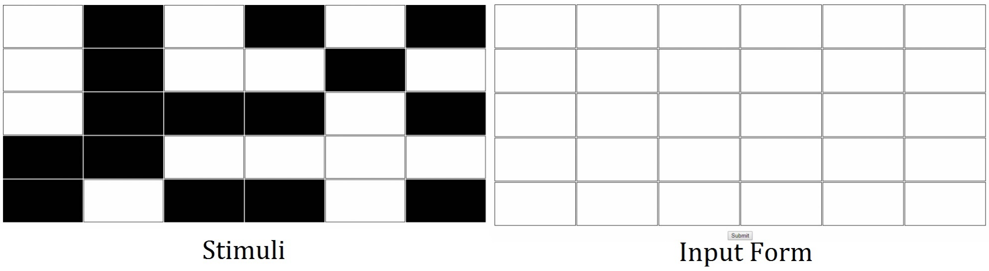


Figure 4.1 Spatial Task - Stimuli and Input form

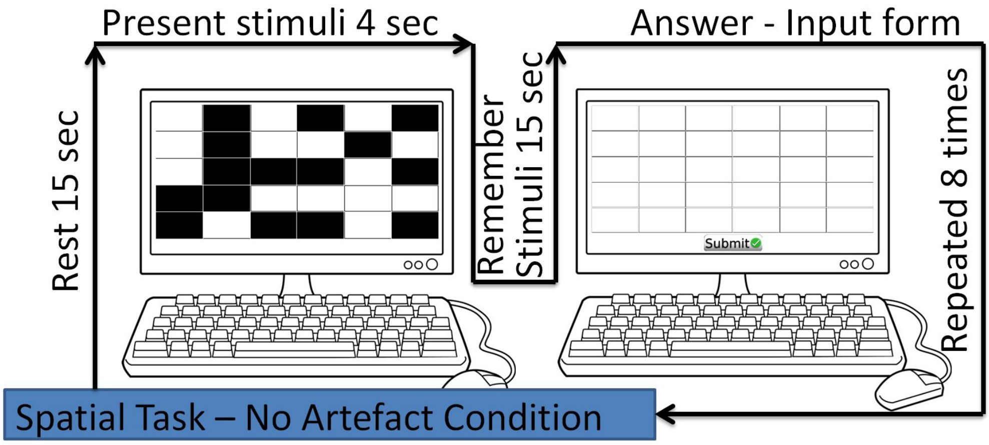


Figure 4.2 Experiment Procedure with Spatial Task

4.3.3 Study Protocol

The study procedure closely followed that of the original study by Solovey et. al.

Participants completed eight experiment parts; two task encodings x four Study Conditions (three Artefact conditions and one Control-No Artefact-condition), with each part composed by eight trials - like presented in Figure 4.2. Each trial started with 15s rest, followed by 4s presented stimuli (the s 7 digit number or the gird), 15s remembering the stimuli, and ended with an input form for answering the remembered stimuli. For the artefact conditions the 15s remembering period also included performing the specific artefact, and an additional 15s period of performing the artefact alone was performed after the task. Performing the artefact involved asking participants to move their head repeatedly, type random keys at the keyboard repeatedly, or verbalizing a previously decided word repeatedly, depending on the study condition.

4.3.4 Measurements, Data and Equipment

Two types of measures were collected in this study, namely brain activity using fNIRS and task performance.

fNIRS data was recorded using an fNIRS300 device and the associated COBI Studio recording software provided by Biopac Systems Inc. Using the Matlab Toolbox NIRS-SPM [150] we applied filtering algorithms to remove high-frequency noise, physiological artefacts such as heartbeats and motion derived artefacts. Finally, each trial was separated according to the condition under test (rest/ task/ artefact) considering the slow hemodynamic response [136], and averaged the data accordingly.

Task performance for both task types was calculated using two measures: Absolute performance - where an answer is simply correct or not, and Relative Performance - where answers were scored according to distance from the target answer (calculated with Levenshtein distance).

4.3.5 Study Conditions and Hypothesis

For each artefact, there were four steps tested by Solovey et. al. with the corresponding study research question, as described in Figure 4.3:

1 A baseline with no cognitive task or artefact;

2 The cognitive task alone with no artefact;

3 Artefact alone with no cognitive task;

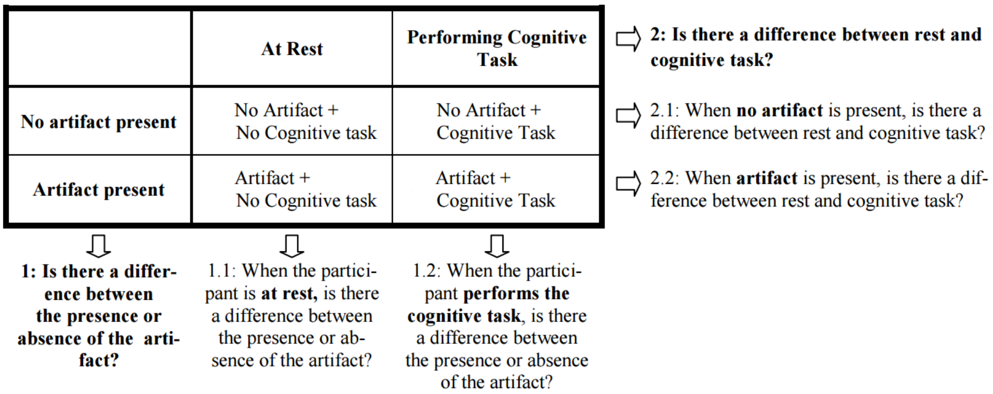
4 Task along with an artefact;

There were two major aims in this study, and we devised the following hypothesis.

A There will be significant differences in the fNIRS data between participants resting and participants performing the cognitive task.

B There will be significant differences in the fNIRS data between participants resting, participants performing the cognitive task in the presence of artefacts, and participants performing the artefacts alone.

We followed a similar approach in this study.



	At Rest	Performing Cognitive Task
No artifact present	No Artifact + No Cognitive task	No Artifact + Cognitive Task
Artifact present	Artifact + No Cognitive task	Artifact + Cognitive Task

2: Is there a difference between rest and cognitive task?

2.1: When **no artifact** is present, is there a difference between rest and cognitive task?

2.2: When **artifact** is present, is there a difference between rest and cognitive task?

1: Is there a difference between the presence or absence of the artifact?

1.1: When the participant is **at rest**, is there a difference between the presence or absence of the artifact?

1.2: When the participant **performs the cognitive task**, is there a difference between the presence or absence of the artifact?

Figure 4.3 Original Study Aims edited from [127]

4.3.6 Participants

Fifteen participants (11 male, 4 female) with an average age of 22.06 (SD = 2.31) were recruited to take part in the study. All participants had normal or corrected vision and reported no history of head trauma or brain damage. The study was approved by the Computer Science ethics committee. Participants provided informed consent and were compensated with gift vouchers.

4.4 Results

Table 4.2 summarizes the findings of this experiment in contrast to the original study by Solovey et. al. [127]:

		Control	Artefacts				
			Head Movement	Typing	Mouse Movement	Facial Movement	Verbalising
Task	Verbal	✓HbO	℧	✓HbO			✓HbO
	Spatial	✓HbO	✓Hb	℧			✓HbO
	▨Non-Replicated Conditions, ■Replicated Conditions, ▣Novel Conditions						

Table 4.2 Results and Contributions of the current and Solovey et. al. study. ✓HbO means fNIRS is fine to use in the presence of the investigated artefact, best measure to use HbO. ℧ means that the artefact needs to be avoided or filtered.

4.4.1 Performance data

No significant difference between conditions was reported by Solovey et. al. in task performance, where the number of correct (in-place) digits was used as the dependent variable. It was hypothesised that non-related verbalisation will negatively impact performance during the Verbal task, as demonstrated by Pike et. al. [108].

Based on Wickens MRM [144], no performance differences are expected under Spatial conditions as the resources are complementary. However, a within participants, one-way repeated measure ANOVA with LSD correction, revealed that participants performed significantly worse under the typing artefact compared to all other conditions during the Verbal task with $N = 15, df = 3, p < 0.025, F = 3.8$ (see Figure 4.4).

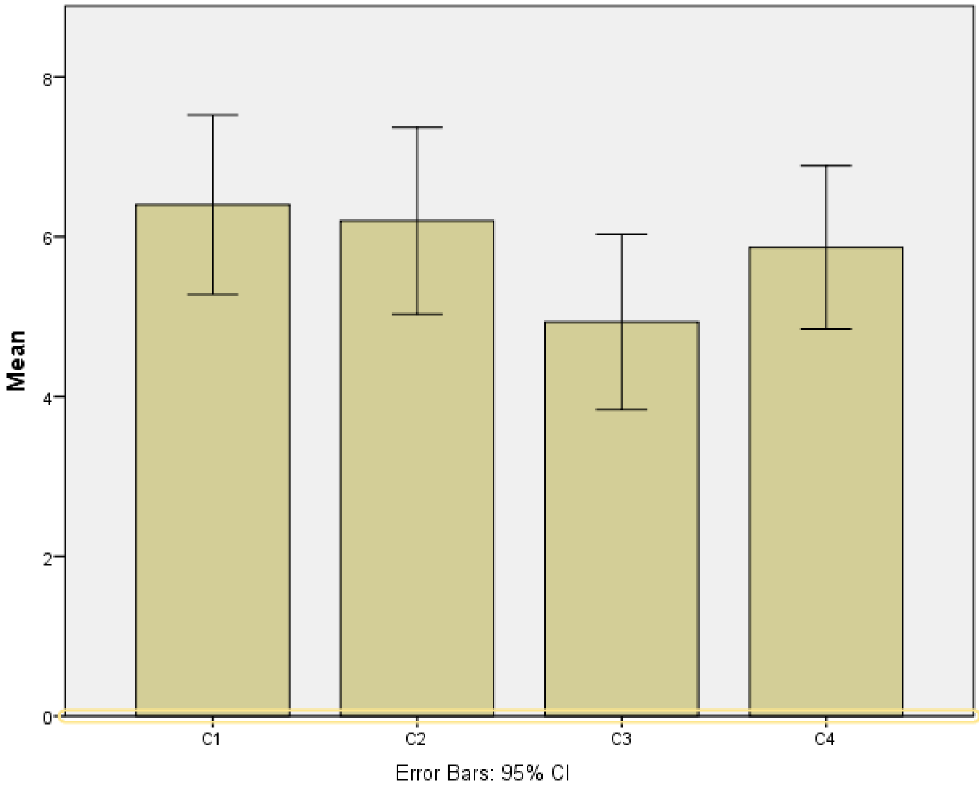


Figure 4.4 Performance outcomes during the verbal task

Participants also performed significantly worse in the Verbalisation artefact condition compared to the no artefact condition during the Spatial task with $N = 15, df = 3, p < 0.05, F = 3.8$ (see Figure 4.5).

The findings fail to prove the hypothesis, but do lead to an interesting discussion. For the Verbal task, the greatest interference was typing, which could be interpreted as being a Spatial input modality since the keys have a physical mapping. On the other hand, for the Spatial task, the verbalising artefact had the greatest interference providing a crossing of resource modalities, which is the opposite of our original hypothesis.

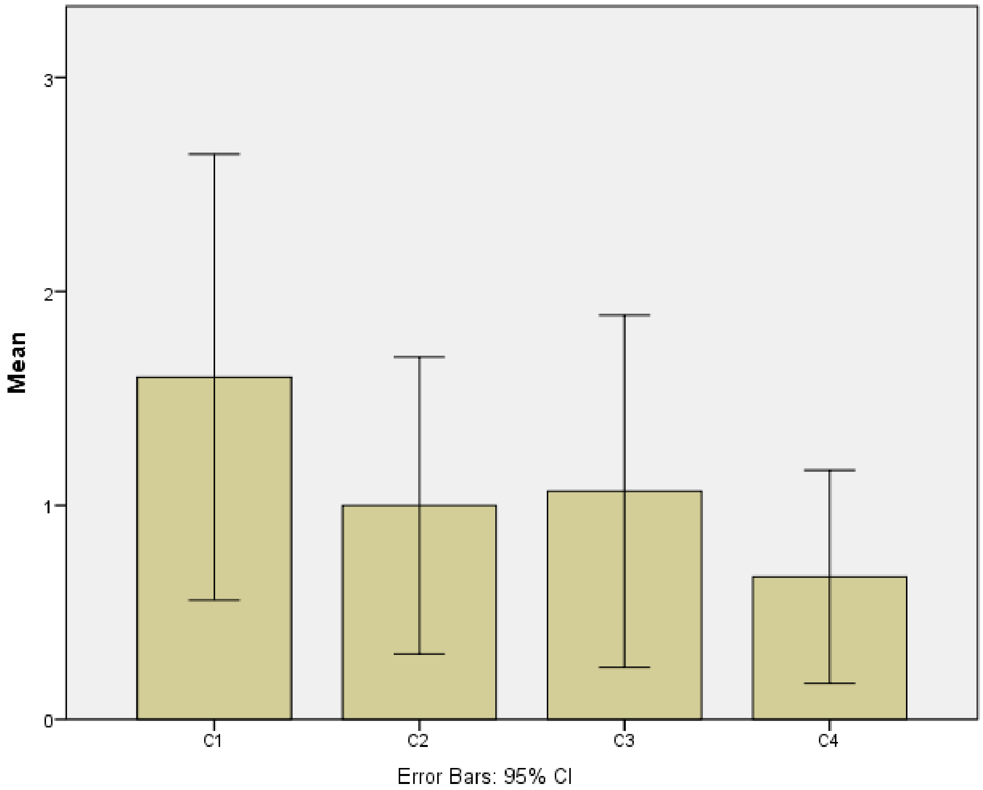


Figure 4.5 Performance outcomes during the spatial task

4.4.2 Experiments: No artefacts (C1)

One of the essential aims in this study was distinguishing between states of rest and cognition, a distinction described as “fundamental” in the original study. That is distinguishing, using fNIRS, between two conditions: participants performing the cognitive task and participants at rest. It was hypothesised that this would be the case in both, the Verbal task condition (as identified by Solovey et. al.), and also in the Spatial task condition. In both cases we hypothesised that fNIRS measurements will be significantly different when participants are resting, compared to when participants are performing either the Verbal or Spatial memory task.

A paired-sample t-test, within participants, revealed significant differences over multiple channels between rest periods and task periods, as hypothesised for both task conditions. Figure 4.8 shows the average HbO levels during the Rest VS Task conditions, across participants for all 16 channels of data from fNIRS. In both task conditions HbO was significantly higher compared to rest states, with $N = 15, p < 0.05, df = 14$ and t value ranging from $t = 2.3tot = 4.3$ for the significant comparisons. During the verbal task, the mean HbO value across all participants was 0.4 during the cognitive task and 0.13 during participants at rest (See Figure 4.6). During the spatial task, the mean HbO value across all participants was 0.42 during the cognitive task and 0.26 during participants at rest (See Figure 4.7).

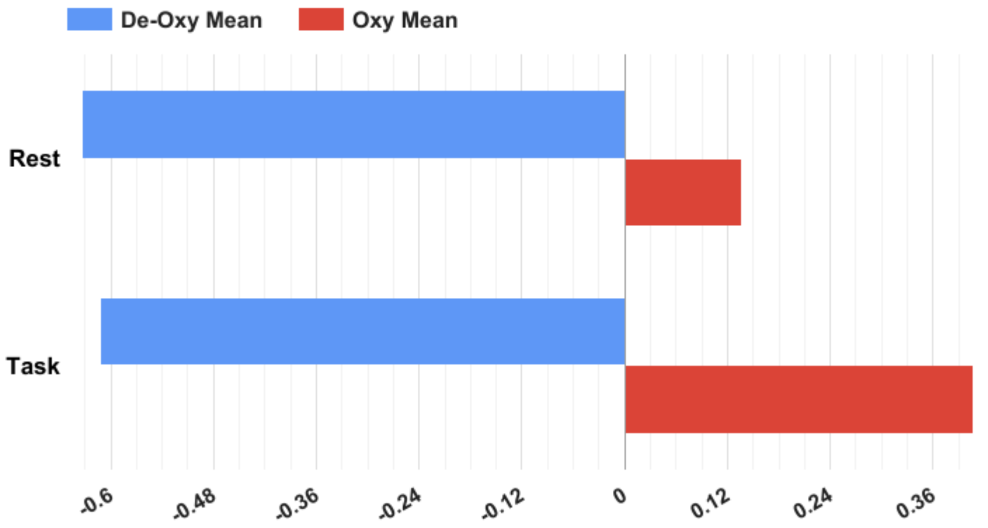


Figure 4.6 Mean Values for Hb and HbO for the Verbal Task in C1

.

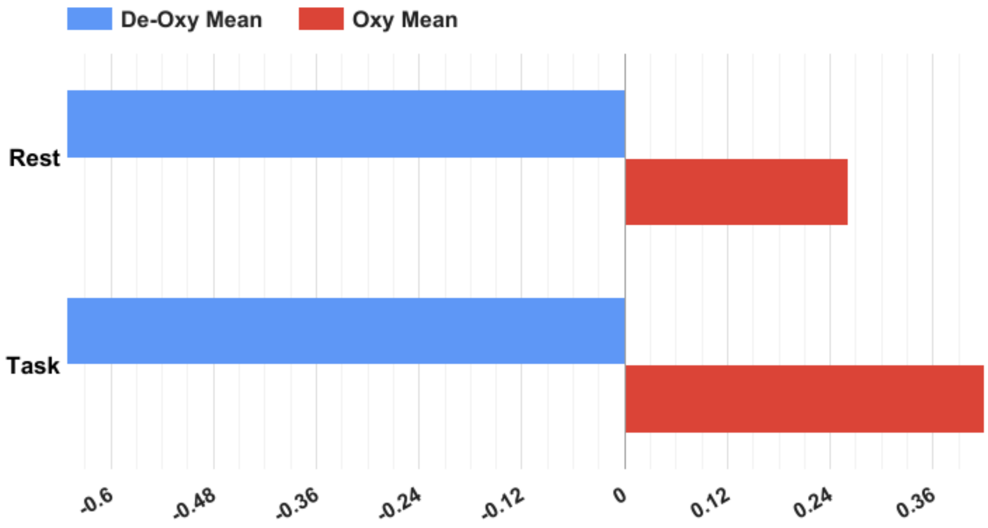


Figure 4.7 Mean Values for Hb and HbO for the Spatial Task in C1

.

Our results are in line with those identified by Solovey et. al. and with our hypothesis regarding the Spatial task. It is to note that the results favoured HbO over Hb in the detection of these states. Based on these findings we accept hypothesis A.

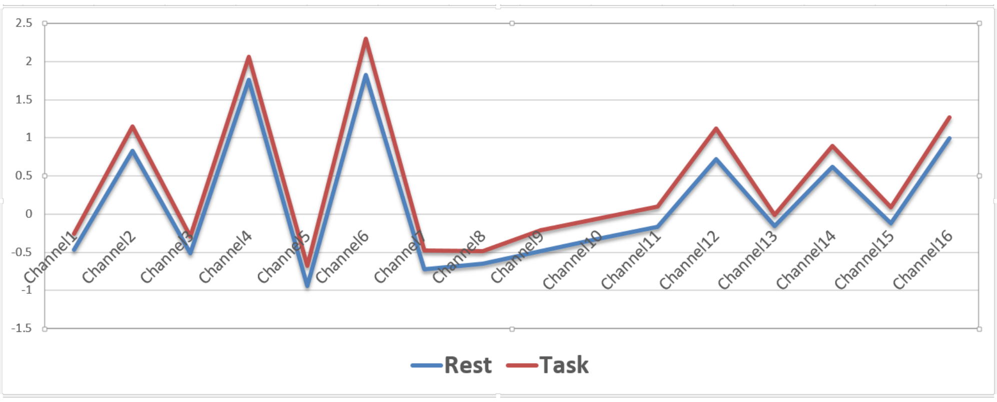


Figure 4.8 Average Oxygenated Hemoglobin (HbO) levels across participants during Rest VS Task task conditions. See how average HbO is higher during Task time compared to rest time across all 16 channels of data.

.

To provide a visual representation of fNIRS ability to distinguish between rest and cognitive states, Figure 4.9 visualises one participant fNIRS data (from one channel: ch.1) for the no artefact experiment (consisting of 8 trials hence the 8 peaks in HbO data).

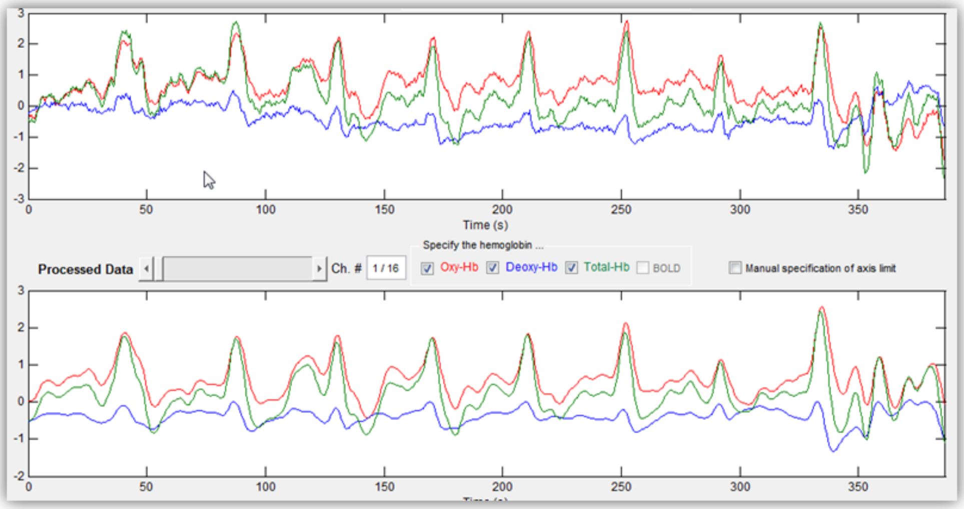


Figure 4.9 Oxygenation level peaks for 8 Verbal trials.

4.4.3 Experiments: With artefacts (C2, C3, C4)

The interest here lies in distinguishing cognition in the presence of artefacts (Table 4.2 provides the summary of our findings). To achieve this, a combination of the following three stages was necessary: rest periods, artefact (alone) periods, and cognitive task under artefact into paired comparisons. A series of one-way repeated measure ANOVAs within participant design with LSD correction were applied for each of the artefact conditions.

C2 - Head Movement Artefact

In the original study, as reported by Solovey et. al., it was not possible to significantly distinguish between participants at rest and participants performing the cognitive task in the presence of major head movement. However, a series of one-way repeated measures ANOVAs showed significant differences between conditions over multiple channels with $p < 0.05, df = 2, F = 3.261$. Our results showed significant difference between participant at rest and participants performing just the artefact ($Hb, p < 0.025$), indicating that head movement is detrimental to the fNIRS signal. Moreover, it was possible to distinguish between cognition in the presence of head movement and performing the artefact alone ($Hb, p < 0.01$), indicating the potential for filtering of this artefact in the future. Accordingly, it is advisable to account for major head movements during studies involving a Verbal memory task. For the same artefact, but under the Spatial task, the results suggest more relaxed restrictions. For the Spatial Task, the results showed significance in Hb for all comparisons (rest/task/artefact) ($p < 0.05, df =$) indicating that Spatial based tasks are less prone to head movement artefacts.

C3 - Keyboard Input Artefact

In the case of keyboard input artefact series of one-way repeated measure ANOVAs showed significant differences between conditions with $p < 0.025, df = 2, F = 4.7$. It was possible to distinguish between rest and task periods ($HbO, p < 0.05$) during the Verbal Task. However, the difference was no longer significant during the Spatial task. Potential for filtering exists again due to the significant difference between the remaining two comparisons (rest vs artefact and artefact vs cognitive task $HbO, p < 0.05$). The findings suggest that keyboard input does not affect the fNIRS signal during verbal tasks, however, it should be controlled for the spatial tasks.

C4 - Verbalisation Artefact

For the Verbalisation conditions again, a series of one-way repeated measure ANOVAs revealed significant differences between conditions with $p < 0.05, df = 2, F = 3.6$.

There were significant differences between rest and cognition periods for both Verbal and Spatial tasks ($HbO, p < 0.01$ Verbal Task and $HbO, p < 0.025$ Spatial Task) in the presence of verbalisation artefact. This finding implies that fNIRS could be reliably used in the presence of Verbalisation artefacts, confirming the findings of Pike et. al. [108]. The results also show that Verbalization artefact is the most compatible with fNIRS for typical evaluation settings.

In this study we discuss the use fNIRS in the presence of various artefacts during both verbal and spatial task encodings, and based on the findings above, we accept the second hypothesis (B).

4.5 Discussions

This study aimed to replicate and extend the work performed by Solovey et. al., investigating the effect of common human behaviours on fNIRS ability to distinguish states of cognition from other states. To do so, the investigated study tested the reliability of fNIRS as a measure during lab-based evaluation settings and extend our understanding of how various artefacts impact the fNIRS signal during different task types (verbal *and* spatial). The verbal memory task was a serial recall of 7-digit numbers and the spatial memory task was a serial recall of 6x6 shaped grids. There were four investigated conditions for both types of tasks; a baseline condition where no artefact was performed (C1) followed by three conditions where participants were asked to perform artefacts such as head movement (C2), random keyboard typing (C3), and repeatedly verbalising (C4).

Objective performance and physiological techniques were used in order to understand the effects of the artefacts on fNIRS.

The fundamental findings in this chapter confirmed that we are able to distinguish between cognitive and rest states during both Verbal (as confirmed by Solovey et. al.) and Spatial tasks.

The two types of tasks, however, were differently affected, according to the two key fNIRS measures, for each artefact. Our addition of a Spatial task, therefore, provided a greater understanding of fNIRS' ability to distinguish cognition under tasks using such encodings. Further, our inclusion of the verbalisation artefact also provided this greater understanding for an additional, but very common user study behaviour. These findings contribute towards a body of evidence to suggest that, for a typical evaluation context, fNIRS is indeed a valuable measure, and has the potential to be used with careful consideration. To provide further practical advice to other researchers about fNIRS reliability and portability, future work might examine other untested artefacts, such as: age of participants, interface familiarity, task expertise, but this is out of the scope for this thesis.

Additional to the original paper a new artefact was investigated - nonsense verbalisation, and a new type of task was introduced in the study design - a spatial memory task. fNIRS was found to be resilient to non-sense verbalisation artefact especially for the spatial task, and inducing a higher level of mental effort especially in the verbal condition (as hypothesised). It was also interesting to find the spatial and verbal task to be differently affected by various artefacts. In the presence of keyboard input artefact, for example, most of the significant differences for the verbal task were in the HbO data, whereas for the spatial task the significant channels of data were in Hb. This could be used for the future fNIRS analysis as an indication of what to expect.

Overall, fNIRS showed more resilience to artefacts in the presence of the spatial task.

Although against our hypothesis, the performance data showed an interesting finding, leading to an interesting discussion. For the Verbal task, the greatest interference artefact was typing, which could be interpreted as being a Spatial input modality since the keys have a physical mapping. On the other hand, for the Spatial task, the verbalising artefact had the greatest interference providing a crossing of resource modalities, which is the opposite of our original hypothesis.

Another way of looking at the two “interfering artefacts” used in this experiment, randomly typing at the keyboard and non-sense verbalisation while performing the two memory tasks is to treat them as the commonly used working memory secondary tasks of “tapping” and “articulatory suppression”. This way, our performance results show that a tapping task would have a higher interference on a verbal task, while articulatory suppression has a higher interference on a spatial task.

4.6 Chapter Summary

Based on the findings in this chapter, we addressed the aforementioned research questions, however, these will further leave space for evidence in the next chapters.

Now that we have discussed the challenges of using fNIRS during lab-based evaluation settings, we can start to understand the challenges of using it for real-time continuous assessment of workload. The next chapter explores the sensitivity and validity of the measure in the context of continuous measurement of workload.

Chapter 5

fNIRS Validity and Sensitivity - Moving towards continuous real time measure of workload

How can fNIRS be used as a sensitive and valid technique in the context of continuous, real-time use, to gain insights into mental workload during tasks?

5.1 Introduction

Now that we know from the previous chapter that fNIRS has the potential to be a useful measure during lab-based evaluation settings to provide an additional channel of information about the user during interaction with technology, we can discuss the sensitivity and validity of the measure in the context of real time use for continuously assessing workload.

Using the guidelines discussed in the previous chapter, we begin to investigate how fNIRS can be used as a sensitive measure of workload, and investigate how it can distinguish between various levels of workload in line with various levels of demand.

In this chapter we present a study in order to support the second research question RQ2, and present the potential of using fNIRS as a continuous measure of workload, moving away from block design analysis, towards more realistic applications (RQ2b). See the overview Table 5.1 below.

Table 5.1 Experiment 2 - Experimental Program Development

No	Research Questions and Aims	Main findings	What next?
RQ2(a)	Investigate the validity of fNIRS measure in contrast to the subjective techniques including NASA-TLX and the continuous Instantaneous Self Assessment technique (ISA). • Evaluate the Sensitivity of fNIRS: how can fNIRS distinguish between various levels of workload? • What is the impact of verbalization on fNIRS signal? • What is the relationship between fNIRS and subjective measurements (Validating fNIRS in relation to existing workload assessment technique NASA-TLX)?	Two conditions were designed (C1 and C2) for this experiment, eliciting different levels of workload (C2 involved a secondary task of non-sense verbalisation), in order to allow performance and workload measures to sense two levels of workload. Therefore, we expected significant difference between the two conditions in performance and workload measurements. • We found significant differences between conditions in the NASA-TLX scales: Mental Effort, Mental Demands, and Physical Demands. • We found a fNIRS to compliment the subjective technique NASA-TLX, and although fNIRS results were not directly conclusive, we found a close relationship between fNIRS and NASA-TLX. • We found no significant differences between the two conditions in task accuracy and performance, however, average time to complete the tasks in C1 was higher than the average time to complete the tasks in C2.	• Although NASA-TLX is a one off measurement technique, taken typically after the task has been completed, we tried to understand its relationship with fNIRS, which is a continuous technique that captures participants' workload during the task. We found significant evidence to show that fNIRS and NASA-TLX are complementary, and follow similar patterns. It is therefore essential to validate fNIRS technique with a continuous workload measurement, such as ISA (Instantaneous Self Assessment), a subjective technique that captures participants perception of workload at regular intervals during the task.
RQ2(b)	Investigate the implications of moving beyond block design, towards the real time-continuous measure of workload (using fNIRS). • How can we assess user's workload using fNIRS during interaction with technology? • Understanding the implications of moving beyond block design, propose the real time-continuous use of fNIRS to detect changes in operators' workload during tasks.	• The findings showed that fNIRS could be used as a reference to workload during tasks. Although this experiment was still following a block design analysis, the challenges of moving towards the continuous measure of workload using fNIRS were discussed.	• Now that we discussed the challenges of moving beyond block design, the next steps would be moving towards continuous, real-time assessment of workload using fNIRS. • Once we use fNIRS for assessing workload during tasks, we can then look into the understanding of how we can use the measure to provide users with real-time workload feedback during tasks.

Simultaneously, we are are trying to further validate fNIRS measure for the assessment of workload, and further contribute to the potential of using fNIRS in the presence of verbalization artefacts as discussed by Pike et. al. [108] and further contribute to 1b. We will do this by understanding the relationship between fNIRS and the subjective measure of workload - the well established workload questionnaire - NASA-TLX [55] (RQ2a).

Measurements such as primary task performance, secondary measures, and subjective ratings are commonly used methods of measuring workload. While performance measures are useful techniques that can reflect participants' workload throughout the task, they become harder to use for highly complex tasks, where metrics cannot be used to quantify performance directly. Subjective measurements are usually obtained after the task has been completed, commonly missing essential information about user's experiences during the task. Therefore, while subjective measures provide useful subjective information, an objective measure of workload such as the one based on fNIRS, could have the potential to provide continuous information about the user over long periods of time (granularity of data is presented in Figure 5.1). To address this issue, the use of a non-invasive, real time brain monitoring technique - fNIRS - is explored, to objectively measure and assess participants' physiological changes in the PFC region of brain related to mental workload during tasks.

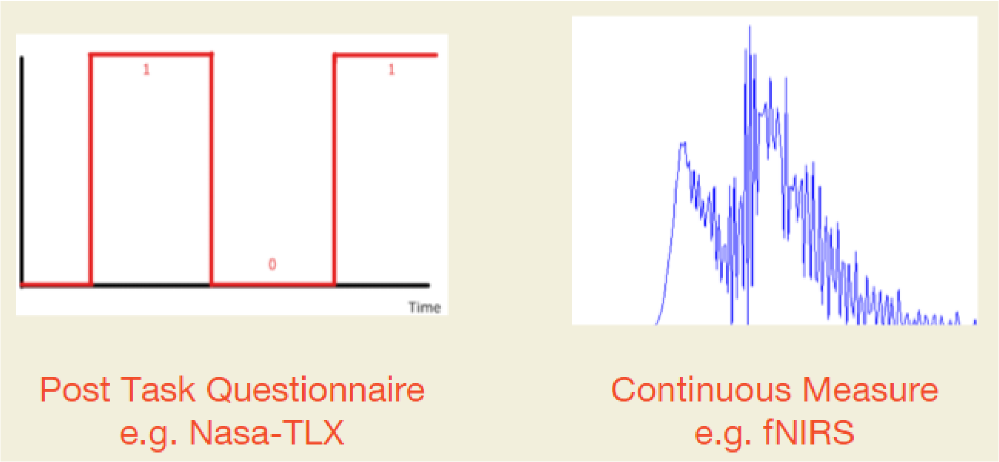


Figure 5.1 Granularity of the measures. Comparison between continuous measures of workload e.g. fNIRS - and - subjective questionnaires of workload e.g. NASA-TLX

Recent research has shown functional near-infrared spectroscopy (fNIRS) to be a highly suitable brain sensing technology for typical user studies, providing an objective, non-intrusive measure correlating to what is known as human Mental Workload. This was further confirmed in the previous chapter. In this chapter we further contribute to the findings in Chapter 4, where we have discussed and showed the reliability of fNIRS within lab-based evaluation settings.

The works presented in this Chapter are mainly based on two presented papers: "Continuous detection of workload overload: An fNIRS approach." presented *In Contemporary Ergonomics and Human Factors 2014: Proceedings of the international conference on Ergonomics & Human Factors 2014*, Southampton, UK, April 2014 and "Measuring the effect of Think Aloud Protocols on Workload using fNIRS" presented at CHI'14 ACM SIGCHI Conference on Human Factors in Computer Systems, Toronto, April-May 2014.

5.2 Experiment design

One aim of this study was to identify how fNIRS could be used as a continuous technique to assess workload as per research question RQ2, and how fNIRS can be sensitive to various task demands inducing different levels of workload on participants. The study also investigates the relationship between the fNIRS measure and the standard subjective workload questionnaire NASA-TLX, in an attempt to validate the technique for assessing workload, as proposed by [121]. The final aim, further contributes to the findings and aims in Chapter 4, and tests the reliability of fNIRS in the presence of verbalization artefact. As "talking" is part of typical "artefact" present when studying the interaction between people and technology, this study also investigates whether simply using your voice and simply verbalizing "Blah blah" during tasks creates an artefact in the fNIRS data.

We have summarized the following chapter research questions and aims:

1. How can we assess user's workload using fNIRS during interaction with technology?

2. Evaluate the Sensitivity of fNIRS: how can fNIRS distinguish between various levels of workload?

3. What is the impact of verbalization on fNIRS signal?

4. What is the relationship between fNIRS and subjective measurements (Validating fNIRS in relation to existing workload assessment technique NASA-TLX)?

5.2.1 Study Conditions

To answer the research questions and aims presented above, a study with two conditions was desired as follows:

- The conditions will have different levels of demand placed upon users, in order to allow fNIRS to sense the differences in users' workload between the two conditions.
- A baseline condition with users performing the task quietly, will be compared to a second condition where they will perform the task while also talking, in order to study the effect of talking on fNIRS.
- The task would be one with controllable difficulty, however, one that would allow a theoretical understanding of the underlying processes related to human cognition, mental workload and the interconnection between these.

A mathematical problem solving task was chosen, as described below, and a study with two conditions was designed as follows:

C1 Baseline Condition that required participants to simply solve the mathematical task, and

C2 Verbal Condition which introduced a nonsense verbal utterance ("Blah") that participants were required to repeatedly verbalise whilst solving the mathematical problem.

5.2.2 Study Task - Mathematical Problem Solving

Considering the desired task properties above, the task had to be chosen carefully, as verbalisation could potentially interrupt the task process. The first criterion, therefore, was devising a task that primarily uses the phonological loop, and thus be a verbally oriented task. It would also be desired that the task allows a secondary task in order to generate a higher demand in the second study condition, allowing fNIRS to distinguish

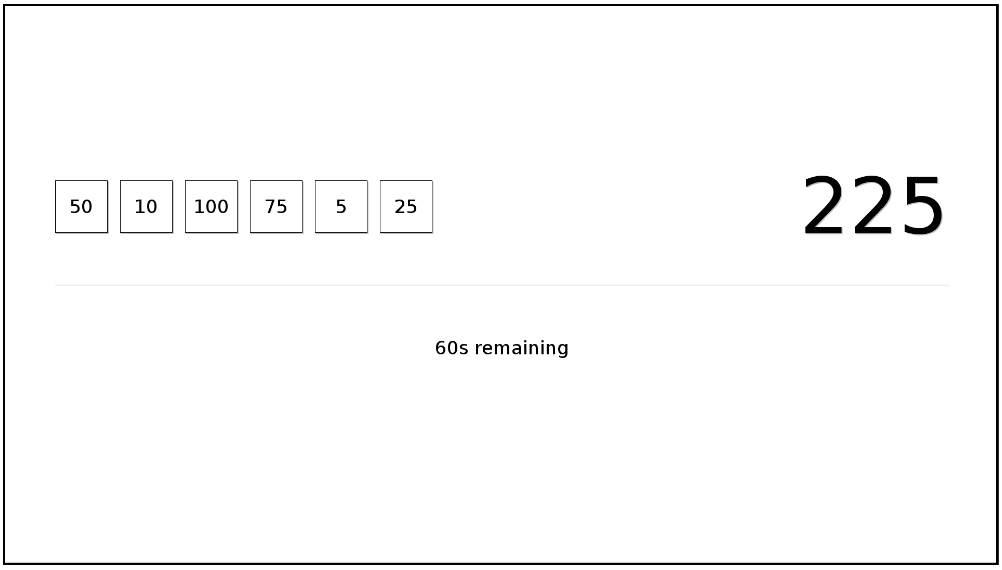


Figure 5.2 A screenshot of the task

between the two levels of demand. The task had to involve continuous use of the phonological loop, and so a simple and discrete memory task was not sufficient. Finally, the task also had to have various levels of difficulty, to enable control over the primary task mental demands; according to the Limited Resource Model [90] harder tasks would increase the demands and thus reducing participant's available resources for a potential secondary task (verbalization in our case). Finally, performance on the task had to be measurable in order to determine the effect of verbalisations, but also understand the implications and relationship between performance and workload. Based upon these four criteria, we decided on using a mathematics problem solving task. Participants were provided with a set of six numbers and had to get as close as possible to a target final number (See Figure 5.2). This problem is a variation on what is commonly known as the countdown problem[1]. Each number may be used only once (although there is no requirement to use every number), and participants have 60s to reach as close as possible to the target number by manipulating the six numbers using four operators: addition, subtraction, multiplication and division. In a simplified example, if the target number is 100 and the given numbers would be $1, 5, 21$, one solution would be $(21-1)*5$ to reach 100.

Sixteen versions of the task were generated at varying difficulties across the two conditions. To classify their difficulty, one researcher and two independent judges rated

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

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 [77]) 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.

5.2.3 Study Protocol

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 session 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 the two conditions which were counterbalanced using a Latin square rotation. Each condition began with a tutorial and practice session. The tutorial session was particularly important for condition C2, as it was used to train the participant on how to verbalise whilst performing the task. 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. To preserve the settings typically observed during evaluation studies, we made little changes to a typical study set-up, the participant was sat at a desk with a standard desktop computer and 20” monitor. 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.

5.2.4 Measurements, Data and Equipment

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

Task Accuracy - P

We measured task performance according to the distance from the targeted answer for each of the 16 problems across the two 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 16 problems performed across the two 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

The NASA-TLX questionnaire was used to capture participants’ subjective workload self-assessment [55], 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.

Table 5.2 fNIRS Measurements

Measure	**Channels**	**Description**
OxyLeft / DeOxyLeft / TotalLeft	3, 4, 5 ,6	The average of channels 3-6, which are located on the left hand side of the device.
OxyRight / DeOxyRigh / TotalRight	11,12,13,14	The average of channels 11-14, which are located on the right hand side of the device.
OxyOverall / DeOxy-Overall / OverallTotal	All Channels	The average of all channels on the device.

fNIRS data - C

fNIRS data was recorded using an fNIRS300 device and the associated COBI Studio recording software provided by Biopac Systems inc. During this experiment the device was placed on participants' PFC targeting the Brodmann area 10 (BA10).

Preprocessing was performed to transform raw data from the device into oxygenation values using the Modified Beer-Lambert law (MBLL) [136]. We also applied filtering algorithms to remove high-frequency noise, physiological artefacts such as heartbeats and motion derived artefacts. To perform this preprocessing step in this experiment we used the Matlab Toolbox, NIRS-SPM [150]. 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 [136] by omitting the first 10s of the trial when processing the data [105].

The Biopac fNIRS device used in this study provides 16 channels of brain data readings. 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. For the full left and right measures see Table 5.2. An overall measurement was produced by averaging the data from all 16 channels.

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 [97]. 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.

5.2.5 Study Hypothesis

From the research aims and questions presented, we propose the following hypotheses:

A There will be a significant difference in performance between conditions C1 and C2.

B There will be a difference in workload between conditions C1 and C2.

C There will be a relationship between the NASA-TLX ratings and fNIRS data.

5.2.6 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.

5.3 Results

The aim of this study is to identify whether fNIRS is suitable in the detection of workload, with the eventual aim of using the imaging technique to detect workload overload. Our hypothesis state that the workload measurement with fNIRS should correlate with the measures observed with the NASA-TLX scale, and thus show the relationship between fNIRS and workload. We report below results in order of the stated hypothesis.

5.3.1 Performance differences between conditions C1 and C2

The aim of hypothesis A was to identify if there was a significant difference in performance between conditions C1 and C2. In other words, trying to investigate whether or not participants felt overloaded during condition C2 compared to condition C1, and whether this had a significant impact on participants performance data (due to the additional demands on users through the continuous non-sense verbalisation).

A series of paired sample t-tests were conducted for all collected performance measures (including measures of task accuracy and response time). Against hypothesis A, the analysis showed no significant difference in task accuracy between conditions. There were no statistical significant differences in performance between the two conditions, and there was no difference in the number of tasks correctly calculated in each condition.

We can attribute this to the findings of Geddie et al [47], who states: “two systems with the same level of overall performance may impose quite different levels of workload on operators”. This was also discussed by Sharples and Megaw [121] in the Framework for mental workload definition and evaluation (see Figure 2.8).

Despite the non-significant results, it is interesting to observe that the average time to complete the tasks in C1 was higher than the average time to complete the tasks in C2 (see Figure 5.3). This is indeed surprising, however, this effect might have been caused by the frustrating secondary task (non-sense verbalization during a verbal task) rather than a lower workload in the first condition.

5.3.2 Workload differences between conditions C1 and C2.

The aim of Hypothesis B was to understand the impact of non-sense verbalization on participants' workload. As described in the previous section, we collected two types of workload data, one capturing participants' subjective workload levels using the NASA-TLX questionnaire, and one objective measure using fNIRS.

Subjective workload using the NASA TLX scale

In the case of NASA-TLX questionnaire, we have analysed both, the weighted score across all its sub-scale, as well as each individual sub-scale alone. In favour of hy-

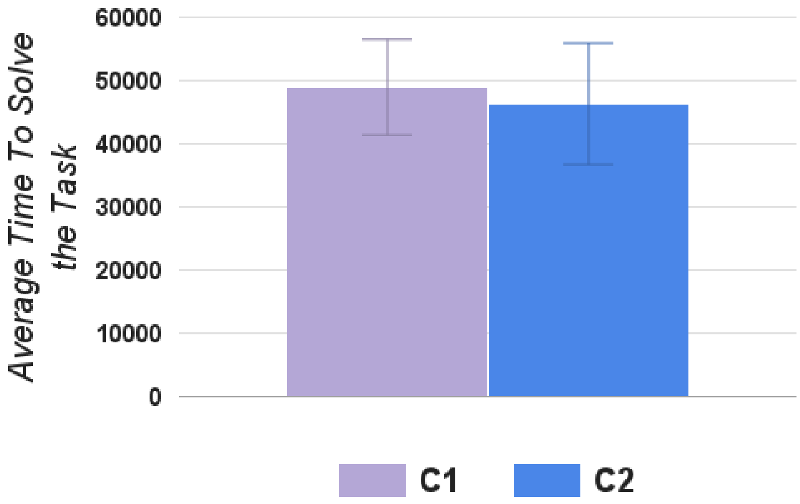


Figure 5.3 Average Time to complete the task in C1 and C2.

pothesis B, we found significant differences between conditions in the NASA-TLX scales: Mental Effort, Mental Demands, and Physical Demands. Twenty participants were recruited to understand whether non-sense verbalization elicited a higher perceived workload compared to the baseline condition. Statistical tests showed that B2 created significantly more mental effort than B1 ($Z = -2.058, p < 0.05$), and it required more mental demands ($Z = -2.292, p < 0.05$). A Wilcoxon signed-rank test showed that C2 created significantly more mental effort than C1 - a statistically significant median increase in the presence of non-sense verbalization for the Mental Effort subscale (mean rank $C2 = 8.68$ compared to the baseline of performing the task alone in $C1 = 7.67, z = 2.346, p < .019$), and it required more mental demand (mean rank $C2 = 7.5$ compared to the baseline condition in $C1 = 5.8, z = 2.292, p < .022$). Participants also rated C2 being more physically demanding in the Physical Demand Subscale (mean rank $C2 = 9.8$ compared to the baseline of performing the task alone in $C1 = 3.3, z = 2.263, p < .024$). This shows that participants found the additional utterance of the nonsense verbalization whilst solving the maths problems inducing a greater physical demand than the other condition.

fNIRS results

As shown in Figure 5.4, OverallHbO was, as expected, higher during C2 compared to C1. This follows a similar pattern to NASA-TLX measure, indicating a relationship between the two. As this effect was not statistically significant directly on fNIRS data, further analysis was performed. We found a statistical difference on the effects on of rest time at the end of each condition in fNIRS data: values at rest after C2 were significantly higher than values at rest after C1 ($p = 0.05$).

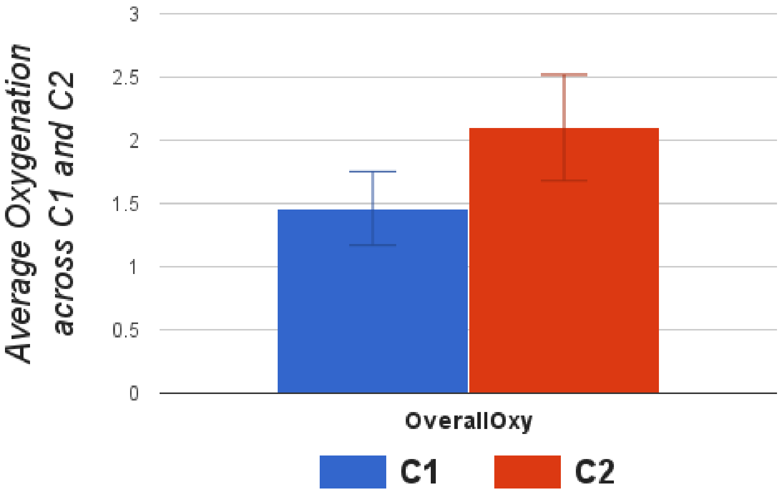


Figure 5.4 Overall Oxygenation comparison between C1 and C2.

Hypothesis B stated that a difference in workload should be observed between conditions C1 and C2 in the study. We found a significant amount of evidence to support this hypothesis, we can therefore reject the null hypothesis and accept hypothesis B.

5.3.3 Relationship between the NASA-TLX ratings and fNIRS data

In relation to workload, and specifically workload overload situations, it is not necessarily the case that our task was not demanding enough to elicit an overload state. Rather, since the performance measure used here were averages across all problems in each condition, some of the overload situations may be hidden through averaging. Hypothesis C states that there is a relationship between NASA-TLX and the measures obtained from the fNIRS device. Correlations between performance scales from unweighted NASA-TLX and performance data were found to support this hypothesis:

- A Pearson correlation ($r = -0.340, p = 0.03$) exists between overall deoxygenated hemoglobin and the mental effort subscale measure of NASA-TLX. This finding agrees with Peck et al [105] who found that decreases in deoxygenated hemoglobin correlated with increased mental effort in NASA-TLX.

- A Spearman test ($r = -0.352, p = 0.02$) identified a negative correlation between the Total oxygenation (HbO + Hb) and Mental Demand subscale from NASA-TLX questionnaire.

We believe the relationship between fNIRS and subjective measurements of workload could be stronger if compared with continuous subjective measurements rather than a retrospective technique (such as NASA-TLX). Further in this thesis, we investigate the relationship between fNIRS and ISA, a continuous real time subjective measure of workload, and we expect a stronger relationship between the two.

5.4 Discussion

The aim of this research was to investigate how fNIRS could be a suitable technique for detecting workload, and explore the sensitivity and validity of the measure. We devised a mathematical arithmetic task with varying difficulties to elicit different workload requirements from participants. Additionally, we introduced another condition which included a nonsense utterance (as a secondary task), requiring the use of additional, non-complementary resources.

The results of this study support the thesis research question RQ2, that aims to investigate how fNIRS can be used to assess workload, and indeed fNIRS could be used to indicate changes in workload between the compared conditions. The same study contributes to the thesis research question RQ1b, which investigates the reliability of fNIRS in presence of various artefacts. This one in particular was contributing to the understanding of fNIRS reliability in the presence of non-sense verbalization as an artefact. We found fNIRS and NASA-TLX technique complementary to each other, and the results showed a relationship between the two even though NASA-TLX is a one off measure, whereas fNIRS is a continuous one.

5.4.1 Articulatory suppression during the mathematical arithmetic task

As described above, in this study we used a mathematical arithmetic task that involved mental calculations. There were two investigated conditions, one where participants were asked to perform the task alone, and one where they also performed a verbal utterance simultaneously while performing the task in order to increase workload.

During mental calculations, the elements of working memory are know to be highly important. Logie et. al. [83] discuss the general implications for the role of working memory in arithmetic problem solving, and showed how the central executive component of working memory is necessary when performing the calculations required for mental addition but also essential in producing approximately correct answers. They additionally discuss how other elements such as the visuospatial resources are involved in approximations and the subvocal and rehearsal is providing means of maintaining accuracy.

Therefore, one could argue that the verbal utterance used in this study played the role of articulatory suppression during a mental arithmetic task. Our perspective is that the verbal utterance did play this role, overloading specific resources in working memory, however, the purpose of the study condition was indeed to create interference and increase workload (compared to the baseline). In this study we were more interested in how can fNIRS distinguish between various levels of workload, and less focused on what exactly was the cause of the workload differences.

5.4.2 Performance measures to reflect participants' workload.

It is known that participants' workload could be assessed using primary and secondary performance measures. This is based on the assumption that there is a direct relationship between task demands and performance outcomes; when task demands increase, there is a high chance of participants' performance to decline. In this study, we have collected participants performance throughout the conditions, and stated hypothesis accordingly. Therefore, a significant difference in the performance data was expected between conditions C1 and C2, due to the utterance and extra-demands placed upon participants during C2.

As there was no statistical significance in performance between the two conditions, and there was a statistical significance in the subjective workload measure, a discussion is raised on the relationship between performance measures and workload. Although some studies showed a direct relationship between workload and task performance, others showed dissociations between the two [103, 138], mostly in the context of subjective measures [50, 151]. The results in this study, add to the growing discussion and previous findings of Geddie et al [47], who states: "two systems with the same level of overall performance may impose quite different levels of workload on operators". This effect was also discussed by Sharples and Megaw [121] in the Framework for mental workload definition and evaluation (see Figure 2.8).

5.4.3 NASA-TLX questionnaire

In this study we have collected participants' perceived workload using the NASA-TLX questionnaire. There was an expected difference in workload between C1 and C2, due to the additional utterance and demands placed upon participants by the simultaneous non-sense verbalization task (of verbalizing "Blah Blah") while performing the verbal "countdown" problems.

The weighted score across all the sub-scales, as well as each individual subscale was analysed. Although the weighted score was not significant, the Mental Demand, Physical Demand and Mental Effort sub-scales were conclusive. The results suggest, as expected, increased mental effort, physical effort and mental demand in the presence of the non-sense verbalization task, as hypothesised (Hypothesis B).

5.4.4 fNIRS - continuously assessing workload

Activations in the left side of the pre-frontal cortex are known to occur during semantic, relative to non-semantic, tasks that relate or involve "the generation of words to semantic cues or the classification of words or pictures into semantic categories" [45]. 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.

fNIRS is picking up an indicator related to mental workload and the fNIRS data showed a higher workload during C2 compared to C1. This could be explained based on the non-compatibility and non-complementarity resources used during B2 with the mathematical reasoning task.

Despite not finding any significant differences with fNIRS data directly, further analysis showed how fNIRS results found an indicator related to mental workload and that C2 induces more workload, and how fNIRS measure is complementary to the existing measure, NASA-TLX.

One of the reasons for not having significance on the conditions data with fNIRS could have been the power of the study; increasing the number of participants would increase power, reduce type II error and positively impact our findings; hence the after effect we found in the rest data. Also we note that overload situations might be hidden through averaging the fNIRS data over conditions. Data that cannot be identified with the NASA-TLX questionnaire might be detected using fNIRS.

This experiment followed a block design analysis, having two conditions with an expected varying demand. Therefore, it was quite simple to draw hypothesis and interpret results. One of the future challenges, is moving beyond block design, towards continuous assessment of workload during task, in less controlled settings. One solution can be having tasks that are easily modelled in terms of their demand placed upon users. A different approach could include a second continuous workload measure, having it as a reference in future fNIRS experiments.

5.4.5 Relationship between subjective and objective measures: insights into using a continuous subjective measure

Workload is a construct that sits in the intersection of multiple contributing factors (as presented by Sharples and Megaw [121] in the framework for mental workload definition and measurement - Figure 2.8), that could be categorized depending on the measurement technique. For example, one will consider capturing the different elements that influence demand on an individual, or the elements that have an impact on the individual's performance, or how the individual perceives different changes in task demands.

In this study, we attempted to capture participants' workload by:

- evaluating participants' performance during the study,
- capturing participants' perception of their load using the self-assessment NASA-TLX questionnaire,
- and, using fNIRS to capture participants physiological (brain) responses to different levels of task demands.

One of the aims of this thesis (and research question RQ2a) was validating the fNIRS technique for assessing workload. This is one of the criteria when selecting and establishing a workload measurement techniques, as proposed by previous authors (e.g. [101] and [62]).

We have therefore attempted to understand the relationship between the different techniques used in this study. The relationship between fNIRS and NASA-TLX provide insight into fNIRS ability to detect different levels of workload. Coinciding with the findings from [105] and [37], this study suggested that fNIRS is in fact capable of detecting workload continuously, and the technique is able to sense between different levels of workload.

This study was an example of how a combination of complementary measures (NASA-TLX, fNIRS and other measures of workload) can provide greater insight into human mental workload. fNIRS property of being a continuous measure enables the detection of workload states that are not observable in NASA-TLX data alone.

An interesting discussion when comparing fNIRS measure with NASA-TLX is the different nature of the two measures. Whilst NASA-TLX is a one off questionnaire at the end of a condition or experiment, fNIRS is a continuous measure that captures insights throughout the experiment. Therefore, the comparison between the two measures is difficult. However, one solution can be comparing fNIRS to a continuous subjective measure, such as ISA (the Instantaneous Self Assessment). The following chapter (Chapter 6) presents one study where fNIRS is compared to ISA.

This works and the ones presented in the next chapter contribute towards research question RQ2a.

5.4.6 fNIRS - use in evaluation settings

One of the aims in this study was evaluating the potential of using fNIRS as a sensitive measure for evaluating participants' workload in lab-based evaluation settings. fNIRS was chosen for its non invasive nature, portability and relative resilience to motion artefacts as presented in the previous chapter (Chapter 4). This chapter further investigated the suitability of the technology in lab-based evaluation settings in the presence of verbalization, a typical artefact in such settings. This was in line with the thesis research questions RQ1b. We found the device to be suitable ecologically for such settings, providing rich data about the interaction in the presence of verbalization, with minimal distraction and interference. Furthermore, at the end of the study, participants were informally questioned about their comfort and experience with regards to 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. Therefore, it is advisable that studies utilising fNIRS, and the particular headband used in this study, should aim to keep study sessions below one hour in a single sitting.

In line with the research question RQ1b we found fNIRS well suited to typical evaluation and usability testing settings.

5.5 Chapter Summary

In this thesis so far we have investigated the reliability of fNIRS in lab-based evaluation settings. We discussed the consideration for its use for Human Computer Interaction research, and measured the impact of artefacts typical for this sort of study settings. This chapter further validated the technique, by comparing it to the widely used subjective technique NASA-TLX. Although fNIRS proved to follow similar patterns in terms of sensitivity to workload, considerations were discussed whether a continuous subjective technique is more appropriate for comparison to fNIRS, and ISA was proposed. This chapter also explored the challenges of moving beyond block design, and future studies were proposed where fNIRS should be used to assess workload continuously. The next chapter will further investigate the validity and sensitivity of the measure by looking at other correlations between fNIRS and continuous subjective techniques for assessing workload. However, fNIRS proved to be a useful measure during Human Computer Interaction, and its use in real time for continuous assessment and feedback of workload is proposed for the next chapter.

Chapter 6

Workload Alert. Using Physiological Data to Assess and Feedback Workload in real-time

6.1 Introduction

How can a real time, continuous version of fNIRS be used to give workload feedback to the user during tasks?

Up to this point, the works presented in this thesis showed how various techniques can be applied with fNIRS in order to assess users mental workload during human computer interaction studies. Chapter 4 presented the reliability of fNIRS in such settings, the following chapter showed the sensitivity of the measure and its ability to distinguish different levels of workload that can be used in studies in order to evaluate interactions.

This chapter presents the final study in this thesis, which explores how a real time, continuous version of fNIRS can be used to give workload feedback to the users during tasks. Therefore, the first part of the chapter includes the literature review on feedback. It will address RQ3 with the associated RQ3a and RQ3b, exploring the impact of workload feedback on both, task performance and participants' subjective ratings. This chapter is based on the "Workload Alerts - Using Physiological Measures of Mental Workload to Provide Feedback during Tasks" paper which was submitted for publication at the ACM Transactions on Computer-Human Interaction (TOCHI). See the overview Table 6.1 below.

Table 6.1 Experiment 3 - Experimental Program Development

No	Research Questions and Aims	Main findings	What next?
RQ3	How can a real time, continuous version of fNIRS be used to give workload feedback to the user?	• A rather simple approach to continuously assessing workload using fNIRS was presented. Objective measurements of brain activity were monitored using fNIRS and classified based on individual physiological responses to task demand into two states of interest: High and Low workload states. These were further used to provide real-time feedback to participants during tasks using an office dynamic lighting environment that had programmable office lamps with capabilities of changing their colour. • Participants were affected in different ways by the feedback. This was identified in multiple data streams, both, objective and subjective, collected during the experiment Some of the participants described feedback as a good indicator of "how much" is going on during the task. Others found the feedback intrusive.	Although this study was not focused on measuring in-task behaviour change, qualitative anecdotes imply that people did reflect on their mental workload and considered their current status. It would be highly interesting in future work to more directly study whether or not there are behavioural markers for when participants take action based on their feedback.
RQ3(a,b)	• Explore the impact of workload feedback on task performance. • Explore the impact of workload feedback on on subjective ratings.	• Both subjective and objective measurements of workload used in this study, provide evidence that participants' mental workload was associated with task demand. • The performance and perceived performance results suggested a negative impact in the presence of the concurrent subjective workload measurement technique used, ISA, most likely due to the additional required resources. We found that this was not the case with providing objective feedback of workload. • Although not significant within our sample, feedback appeared to slightly improve actual performance and participants perceived that they performed slightly better.	• One of the future directions of this work could be moving towards the assessment of workload of everyday task, moving beyond controlled lab-settings, even moving away from brain based sensors, using less invasive physiological techniques (e.g. HR, EDA, BHP). • In this study we provided participants with objective based feedback of their workload during task using a binary feedback type for two states of interest: high and low workload. Future work could also examine other feedback types, including more granular types, as noted qualitatively by participants during the post-experiment interview.

Sharples and Megaw [121] described the effect of workload as “the relationship between primary task performance and the resources demanded by the primary task”. They illustrate two causes of decrease in performance: 1) underload and 2) overload conditions, where task performance drops as mental workload increases beyond an individuals capacity. One concern in this thesis is understanding users’ capabilities and limitations in terms of their Mental Workload during interaction with technology.

The study presented in this chapter explores and investigates whether giving users real time unobtrusive feedback based upon an objective assessment of their mental workload can help them understand and manage it during tasks. We compare this feedback to traditional methods of asking users to self-assess and report their own mental workload during tasks.

An individual self-assessment of workload, would potentially allow them to regulate their resource allocation to the primary task, therefore not reaching the described 2 conditions. There are a variety of subjective and objective methods used for measuring mental workload including primary and secondary task analysis [91], physiological or psychophysiological techniques [59, 76, 108, 127], as well as user opinions using subjective techniques [55, 72]. The most common methods, however, involve asking people to self report their level of mental workload simultaneously with the task in hand, but this has been shown to negatively impact workload and task performance itself [133].

In the following sections of this chapter, we present fNIRS as a real-time vs post experiment tool to assess workload as used in the analysis for this particular study. The chapter continues by describing a user study comparing the impact of mental workload feedback to the traditional method of asking users to self-assess and report their own mental workload, ISA (Instantaneous Self Assessment). We then present the results of the study, discuss the findings in terms of what we can learn about feedback of mental workload in general, and give recommendations for further work in this research direction.

6.2 Feedback

Feedback allows us to review, reflect, and improve our performance. In this section we review the related works and literature on feedback. In line with the final aim of this

PhD, we will explore later in this chapter how feedback of workload could be used to support operators during tasks by allowing them to reflect over their state.

Reflection

The word “reflection” originates from the Latin verb “reflectere” which means bend or turn ("flectere") backwards or back ("re") [17], and it was used to describe the reflection of light against some reflective surface (e.g. water, mirror). When the word “reflection” is used to refer to thinking, meditation, cogitation and similar intellectual activities, it means that some phenomenon is subjected to thorough consideration, and involves focusing for a longer period of time on an object in order to get a better and deeper understanding of it [17]. When the object of reflection is one’s own activity or character, then we can refer to it as self-reflection.

To explain reflection, Schon discusses the “feeling” when doing something right or wrong [118]. When one notices doing something good, this will encourage him/her to repeat the exact thing that he/she did before that proved successful.

Schon describes that “studying the winning habits”, makes us think about the “know-how” that enabled us to win. This process of understanding and thinking about various patterns of actions, while we perform various tasks or immediately after, can be refereed to as reflecting on action and, in some cases, reflecting in action. In this study, we aim to understand what “feels” appropriate for people when presented with mental workload feedback. We will be looking to understand if people notice any patterns of actions when reflecting on their mental workload.

> “Reflecting-in-action. If common sense recognizes knowing-in-action, it also recognizes that we sometimes think about what we are doing. Phrases like “thinking on your feet”, “keeping your wits about you,” and “learning by doing” suggest not only that we can think about doing but that we cant think about doing something while doing it. Some of the most interesting examples of this process occur in the midst of a performance”. [118]

Starting from knowledge, Schon et. al. [118] presents the following properties of “knowing”:

- There are actions, recognitions and judgements which we do not have to think about during or prior their performance. So knowing, we will carry these out spontaneously.
- We are often not aware of learning new things, “we simply find ourselves doing them”.
- Whether we are/were aware or not of the understandings for certain actions, we are usually unable to describe the knowing which our action reveals.

Moreover, even in the context of workload and performance, Sharples and Megaw discuss feedback as one of the five relationships in the framework for mental workload measurements. They present both the unconscious and explicit feedback of performance as something that operators monitor during tasks to self-judge and cognitively think about their state. Similar to the unconscious and explicit feedback of performance, the aim of this chapter is to allow operators to reflect on their workload during task completion.

6.2.1 Biofeedback

Biofeedback is the process of presenting a person or user with feedback of their physiological information such as blood pressure and heart rate. “The motivation for early biofeedback research was to explore whether displaying real-time physiological information ... would be sufficient to condition physiological processes” [131]. As presented in [131], initial work in the field was much focused on using biofeedback towards the treatment of chronic illnesses such as migraine headaches and hypertension, but this was soon disregarded due to unsuccessful validation of the technique [114].

With a different scope, biofeedback was adopted by neuroscience and BCI research. Neural biofeedback is the biofeedback of brain and neural activity and it plays an essential role in training and control for the BCI system. As presented in [131], a typical BCI user learns to control a particular brain signal (such as increasing the amplitude of motor cortical signals, [149]) or to indicate the relative status of a brain signal (also explored using fNIRS [99]).

In this chapter we will explore neural biofeedback in the context of providing the user with workload feedback based on their brain activity measurements using fNIRS.

Metacognition and Mental Workload

Metacognition is the state where one reflects upon one's thoughts i.e. *"thinking about thinking"*. Fletcher [44] showed that when in meta-cognitive states, users can monitor their performance, task cues and other states, and therefore potentially assess their mental workload throughout the task and 'act' accordingly, if given mental workload feedback. One consideration of continuous subjective measures (ISA) during tasks is that rating your own mental workload subjectively, will also make you aware of this information, potentially having an impact on your meta-cognitive state. The interest of this study in meta-cognition, is the potential for supporting such acts by presenting relevant mental workload feedback, but in a way that does not affect task performance such as continuous subjective methods do.

6.3 Implementation

While Brain-Computer Interfaces (BCIs) were traditionally focused on users with disabilities, providing them direct control or interface with the outside world, current advances investigate the use of brain as an additional channel of information about healthy users interacting with technology. This "passive" rather than "active" channel, sometimes called implicit [126], can act as a complementary source of information about users' state, that can be combined with traditional methods, or used as an input to system, task, or interface. In this section, we present a Workload Feedback System (WFS - see Figure 6.1) that uses passive BCI to 1) measure, 2) detect, and 3) feedback users' workload during tasks.

6.3.1 Monitoring workload using fNIRS

Measures of brain activity were recorded using an fNIRS300 device and the associated Cognitive Optical Brain Imaging (COBI) Studio hardware integrated software platform provided by Biopac Systems Inc.

For real time use, COBI Studio requires collecting baseline signal levels before the start of the study conditions that are used in order to calculate oxygenation in real time, via the Modified Beer Lambert Law (MBLL) [136]. The resulting data was used as an input for the WFS classification step described later in this section.

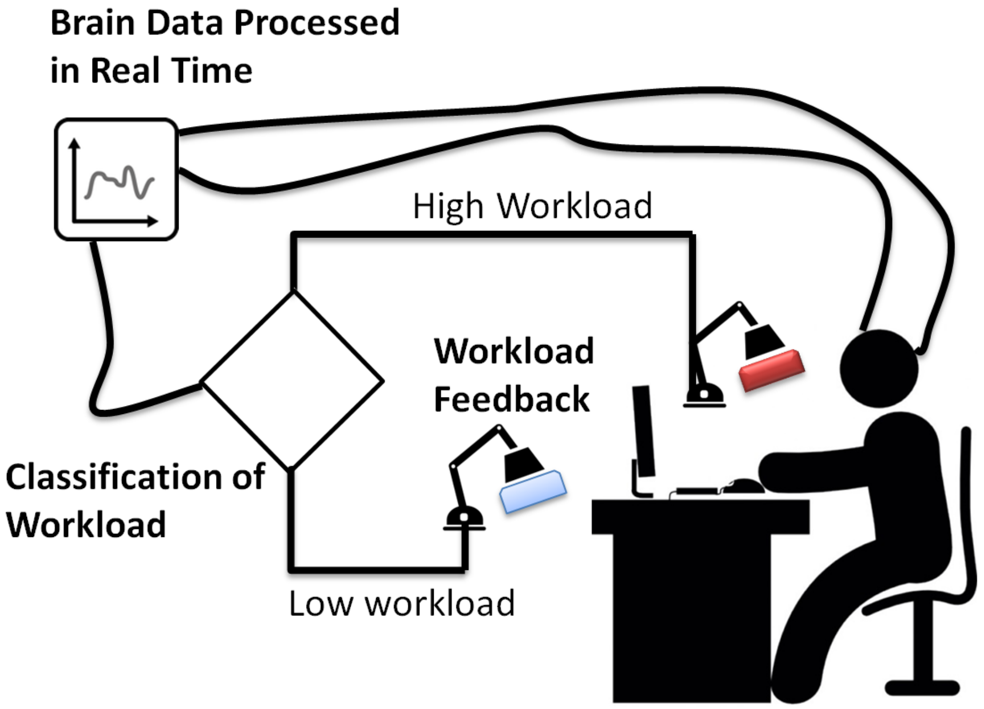


Figure 6.1 The Workload Feedback System WFS

fNIRS data can be further processed and analysed post-experiment, with the advantage of using additional filters and feature extraction techniques discussed below. For the post-experiment fNIRS analysis the data was processed using fnirSoft, the Comprehensive Signal Processing, Analysis and Visualization Platform for Optical Brain Imaging [8]. A low pass filter with cut off frequencies of 0.2 Hz was used in order to remove high-frequency noise, physiological artefacts such as heartbeats and motion derived artefacts. Additionally, the Correlation Based Signal Improvement (CBSI) method was applied [35], a technique designed for fNIRS technology in order to improve detection of workload. In the analysis, the delay associated with the hemodynamic response [136] was taken into account using various techniques including: averages across blocks of data, omitting the first few seconds of the trials when processing, or simply delaying the trial data by a few seconds [105, 108].

For each participant in the study, the device was placed on the PFC targeting the Brodmann area 10 (BA10) (see Fig. 6.2). The Biopac fNIRS device provides 16 channels of brain data reading. Each channel is defined by the relationship between a IR source and a near by IR detector pair. After the MBLL is calculated, COBI studio provides a file with two measures HbO and Hb, and for the post experiment analysis the TotalHb can be calculated (Hb+HbO), and one additional measure of oxygenation

(OXY) can be obtained from fnirSoft, that is oxygenation (OXY).

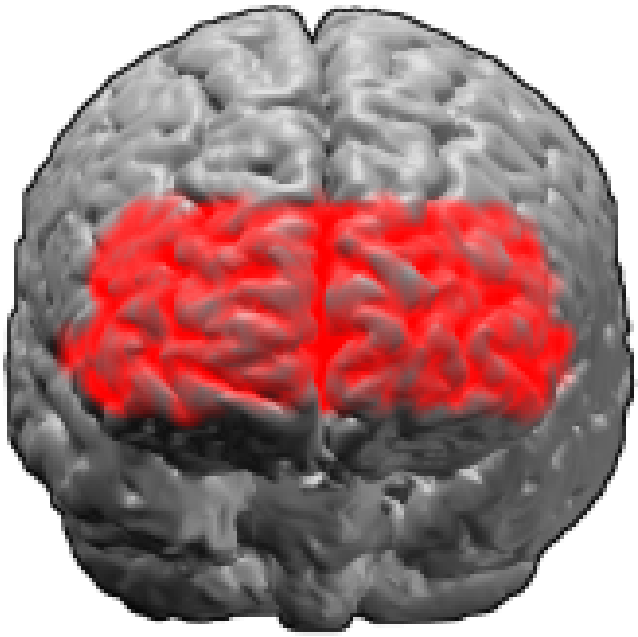


Figure 6.2 Brodmann area 10 (BA10) - targeted using fNIRS

6.3.2 Detection of state

As our fNIRS device provides 16 locational channels of data with two readings per second, an important step before the study could began was identifying localised Hb/HbO changes for each participant. There are three reasons for this step: 1) fNIRS data is highly sensitive to individual differences between participants, 2) the physical placement of the 16 channels varies between participants (based on the shape and size of the forehead), and 3) different forms of workload create changes in Hb/HbO in different regions of the forehead [87]. We used training tasks to identify the most sensitive region, and identified the most valuable channel for the WFS system to focus on; the post-task evaluation, however, utilised recordings from all 16 channels.

Configuration task

With workload having so many different aspects related to the operator performing a specific task, instead of using e.g. N-Back tasks that are well known for eliciting increased levels of workload, we used variations of a task intended to be more representative of the complexity of a real world task, with manipulations corresponding to increased variation of difficulty. This way, we observed the responses associated with increased demand to our actual task for all the 16 channels and two measures (Hb and HbO). The study task, described further below, consisted of an Air Traffic Control game, where participants had to coordinate the landing and departure of aeroplanes. The calibration phase, therefore, included a 30 seconds resting state of relaxing and not performing the task, followed by two 30 second variations of increased demands: low-normal load (3-5 aeroplanes to control), and normal-high load (>7 aeroplanes to control). Averages of HbO and Hb values were used to calculate range thresholds for the three periods (rest, low, and high), that were used later on in detecting significant increases and decreases of workload.

State Tracking

We were particularly interested in monitoring, detecting and feeding back two states of interest: participants reaching a "high" workload state, as well as going back to a "low" workload state. Therefore, using the most sensitive channel and using a running window of 30 seconds, we continuously calculated a rolling average based on the previous 30 seconds worth of readings. The WFS monitored significant increases and decreases in Hb/HbO (of the selected channel) by comparing each new real-time value with the rolling average. A high workload state would be detected if HbO/Hb increase/decrease value was higher than the threshold set during the calibration stage. A low workload state would be detected in the opposite conditions.

6.3.3 Feedback choice

Once the participant state was detected, the WFS produces a binary integer that can be used for changing the state of a feedback mechanism. For the purpose of this study, we specifically designed feedback to be noticeable, but at the same time transparent and in the background of the task, such that a minimum of resources would be used by operators.

For our study, the output of the WFS was used to invoke changes in the desk lighting around the participant, using Philips Hue Bulbs (programmable light bulbs) in desk lamps. Initially, the lighting was set as normal white lighting, which would turn red when participants entered states detected by the fNIRS measurements to have High Workload and return to white as participants returned to lower levels of Mental Workload. We discuss this colour choice further in the section below.

6.4 Experiment Design

As stated above, our aim was to investigate whether providing people with real-time feedback on their mental workload, now that we can objectively and reliably measure it with fNIRS, could facilitate a form of Reflection-in-Action during tasks: that participants, in knowing their Mental Workload levels, can take action to manage their task or workload. Sharples & Megaw 2015, said "*Operator workload or effort is not simply a function of task demands, but is influenced by how the task is perceived by the operator...*". In this case, we are making the Operator Workload explicit in the model, and examining the impact on both performance, and the demands of the task.

Our primary aim, therefore, was to evaluate the effectiveness of using the Workload Feedback System (WFS) to aid an individual's self awareness of current workload, such that they could a) be more aware of their mental workload, and b) achieve good performance outcomes. As a secondary aim, we wanted to examine these outcomes against one of the widely used techniques for keeping people aware of their workload during tasks: Instantaneous Self Assessments (ISAs); ISA, described further below, requires people to self-report their workload at intervals in order to keep them self-aware of it.

6.4.1 Study Conditions

Based on the aims of the study, our two primary independent variables were: 1) the use, or not, of the WFS and 2) the use, or not, of ISA reporting. This created four within-subject repeated-measure conditions, as shown in Table 6.2. Initially, however, we designed the WFS lights to turn red (from normal white) when participants were experiencing high workload. However, midway during the study we noted that multiple participants reported in interviews that the red colour acted as a stressor to their experience. We decided to identify the participants thus far as **Phase 1**, and introduced a 3rd between-subjects independent variable to create a **Phase 2** with the colours reversed: turning white from red when participants experience high workload. In both lighting phases, the lights returned to their base colour when workload reduced, and thus could change back and forth multiple times during each task. We include the colour-based independent variable in the results, and examine the implications of colour choice in the discussion.

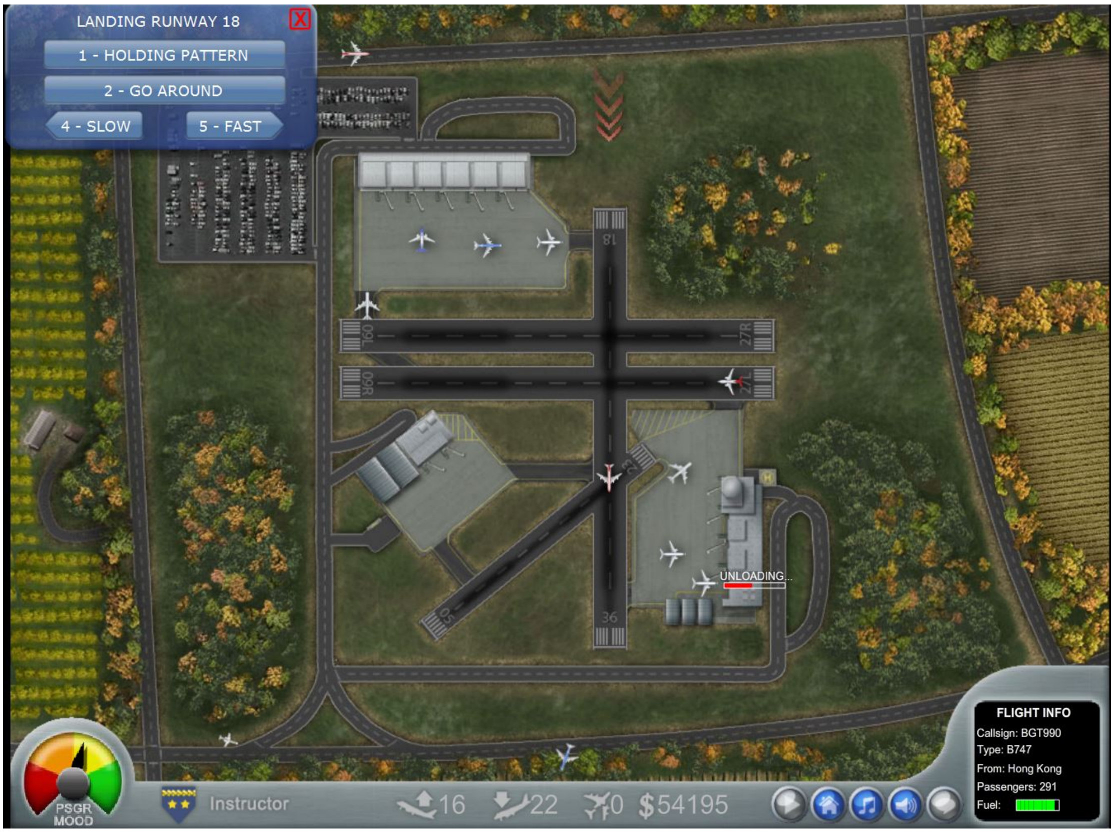
Table 6.2 Four main conditions in the study

Id	**Condition**	**Includes WFS**	**Includes ISA**
C1	Task Only	No	No
C2	Feedback	Yes	No
C3	Feedback + ISA	Yes	Yes
C4	ISA	No	Yes

6.4.2 Air Traffic Control (ATC) Task

For the experiment, we required a task that a) increased in difficulty, and b) could be managed by participants taking action in response to feedback. We selected an ATC task, using the commercially available Airport Madness 4 Game[1], shown in Figure 6.3, in all task conditions. Participants had to coordinate the landing and departing of as many aeroplanes as possible, without causing incidents (e.g. collision between aeroplanes); the number of aeroplanes increased over time, thus increasing the demand as the task progressed. Planes are managed by clicking on the desired plane icon and selecting an appropriate action - ‘Land at runway X’, ‘Go Around’, ‘Increase/Decrease speed’, ‘change direction’. Similar options existed for planes requiring take off e.g. ‘line up’, ‘immediate take off’. These controls allowed participants to use various strategies to reduce their mental workload during moments of high demands by e.g. sending aeroplanes around, managing all landings on one lane and departures on other. The task interface also presented participants with direct measures of performance (seen in Figure 6.3), such as the number of landed/departed aeroplanes.

[1] More information about the study task and a free trial version of the game can be found here: https://www.bigfatsimulations.com/game/airportmadness4



Callsign: BGT990
Type: B747
From: Hong Kong
Passengers: 291
Fuel:

Figure 6.3 Airport Madness 4 - Screenshot of participant managing the landing of an aeroplane.

6.4.3 Participants and Study Protocol

A total of 32 participants were recruited to take part in the experiment. Fifteen participants (9 male, 6 female) with an average age of 25.3 (SD = 2.31) experienced the white-to-red lighting in Phase 1, and seventeen (9 male, 8 female) with an average age of 25.5 (SD = 8.05) experienced the red-to-white lighting in Phase 2. All participants had normal or corrected vision and reported no history of head trauma or brain damage. Participants were given a £10 voucher as a thank you and remuneration for their contribution to the project. The study protocol below was approved by the School of Computer Science ethics committee.

After gaining informed consent, participants began with a task familiarisation tutorial. All participants watched the same recorded video that introduced all the interactions with the video, and then were given the opportunity to practice the task until they felt confident in the game play; participants determined the time when they were ready to begin the experiment. The WFS was then calibrated for each participant, as described in Section 3.

Participants completed each of the four study conditions, which were counterbalanced using Latin-square design to account for learning effects. In each condition, they were required to perform the study task from scratch for a period of seven minutes. If they were to cause more than three major incidents within a condition, the game would automatically stop and the study condition would end (this however was not common). Seven minutes was enough to see numerous workload changes in the lights, but keep the full length of participation, including training, calibration, four tasks and between-task rest periods, to approximately one hour. After each condition, participants filled in a questionnaire to record perceived performance, before moving onto the next condition. After all four conditions, the study finished with a short interview, where participants had the chance to discuss the study experience and the way they perceived the WFS.

6.4.4 Measures of Dependent Variables

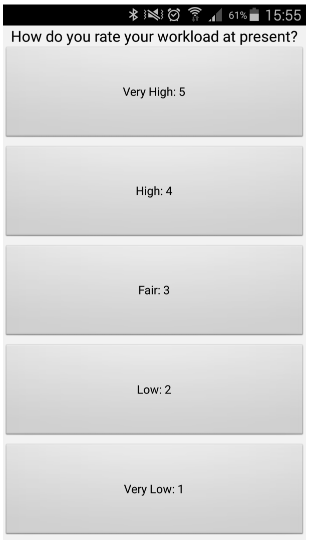
We collected three forms of data from each condition in the study: fNIRS data, ISA data, and Performance Data. We also recorded debriefing interviews to gain insights into participants responses to the conditions that were not otherwise observable in data.

fNIRS data

Although the WFS was only used during the Feedback conditions, participants wore the fNIRS sensor in all conditions. While the WFS system used a single channel in the most sensitive region to monitor workload during tasks, comprehensive fNIRS data (HbO and HO) was recorded from all channels for the duration of all conditions. fNIRS data was further processed for post-experiment analysis using additional filters and feature extraction techniques. fNIRS data was processed using fnirSoft, the Comprehensive Signal Processing, Analysis and Visualization Platform for Optical Brain Imaging [8]. A low pass filter with cut off frequencies of 0.2 Hz was used in order to remove high-frequency noise, physiological artefacts such as heartbeats and motion derived artefacts. Additionally, we applied the Correlation Based Signal Improvement (CBSI) filter [35], a technique designed for fNIRS technology in order to improve detection of workload. We named the resulting data OXY. In analysing the OXY data, we also considered the delay associated with the hemodynamic response [136], using various techniques including: averages across blocks of data, omitting the first few seconds of the trials when processing, and simply delaying the trial data [105, 108].

ISA data

During ISA conditions, participants had to respond to the 5-point rating scale on a mobile device (Figure 6.4) every 30 seconds, prompted by an audible notification; 1 means very low and 5 very high experienced workload. Although the question ‘how do you rate your workload at present?’ had a small font on the device, all participants were instructed and familiar with the tool, and had the chance to play with the mobile app before the start of the experiment. The ISA scores were recorded, as was time-to-respond to the prompt. It is common for secondary task performance, as with reporting ISA scores, to drop during periods of extremely high workload. If participants did not respond to the ISA prompt during a 30 second period, the score was recorded as a 5 (high workload); time-to-respond was calculated from the most recent prompt.



How do you rate your workload at present?

Figure 6.4 Instantaneous Self Assessment (ISA) Recorder App

Performance data

The task was screen captured and recorded for subsequent analysis. Actual performance was analysed in two ways: 1) the performance outcomes (number of planes landed, number of take-offs) at the end of each condition, and 2) demand levels at moments during the task either a) at each ISA interval (number of planes in the air and on the ground), or b) demand levels when WFS lighting changed (number of planes in the air and on the ground). Informally, we were also able to examine the actions and timing of actions taken by the user after key events such as plane accidents, WFS changes, and after ISA responses (go around, increase speed, decrease speed, change direction). After each condition, *perceived performance* scores were collected using a five point rating scale (1 - poor, 5 - excellent performance).

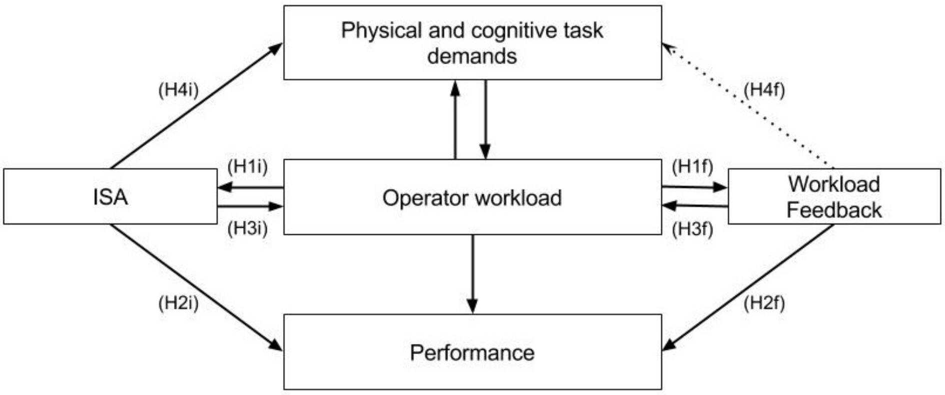


Figure 6.5 Framework for Mental Workload Measurement: the relationship between MWL Feedback, ISA, performance and workload. (Adapted from [121])

6.4.5 Study Hypotheses

To better understand the relationship between Feedback, ISA, workload, and performance, we state our hypotheses based on an adapted version of Sharples & Megaw's Framework for Mental Workload Measurement 2015, where we controlled the External Influences. Essentially, as shown in Figure 6.5, our two primary independent variables are shown as alternatives to the External Influences boxes.

- H1 - *Variation in task demand will create measurable differences in Mental Workload.*

  As a baseline, as it is generally expected that increased task demands will generate increased levels of workload, we therefore expect that changes in task demands should be observable in both the ISA ratings and the objective measures of workload. We expect ISA ratings (H1i in Figure 6.5) to increase with increased task demand, and fNIRS measures to correlate (H2f), either positively or negatively, with task demand.

- H2 - *Participants' performance will be affected, positively or negatively, when made aware of their mental workload.*

  Ideally participants may perform better because they are more aware of their workload without having to self-report using ISA, but may also have decreased performance if the feedback affects their ability to focus. We expect a lower performance in the presence of ISA (H2i) because of the activity involved in self-reporting. However, we do not expect a negative impact on performance from the WFS lighting changes (H2f).

- H3 - *Participants' perception of performance will be affected positively or negatively, when made aware of their mental workload.*

  As operators monitor their own performance, unconsciously or explicitly, having their workload levels presented during tasks should allow operators better reflect on performance. Aside from actual performance, participant's perception of their performance might be affected as they are made aware of their workload levels - increased workload could create a sense of poorer performance or higher performance for both independent variables (H3i and H3f).

- H4 - *Participants' perception and management of the task demands will be affected, positively or negatively, when made aware of their mental workload.*

  We expect that, given feedback-in-action, participants will think about their state whilst performing the task. Participants may also then take action to manage and manipulate the demands of the task, in order to maintain their workload levels to a particular point. In the case of ISA, H4i in Figure 6.5 highlights that there is a direct connection from ISA to the task demands, as participants have to do extra work to report their workload levels. As the WFS does not require additional effort from participants during tasks, we expect that their task demands will not change, however, being presented with feedback of their workload levels more explicitly during tasks, participants perception of the task, and the decisions during the task may be affected (H4f).

6.5 Results

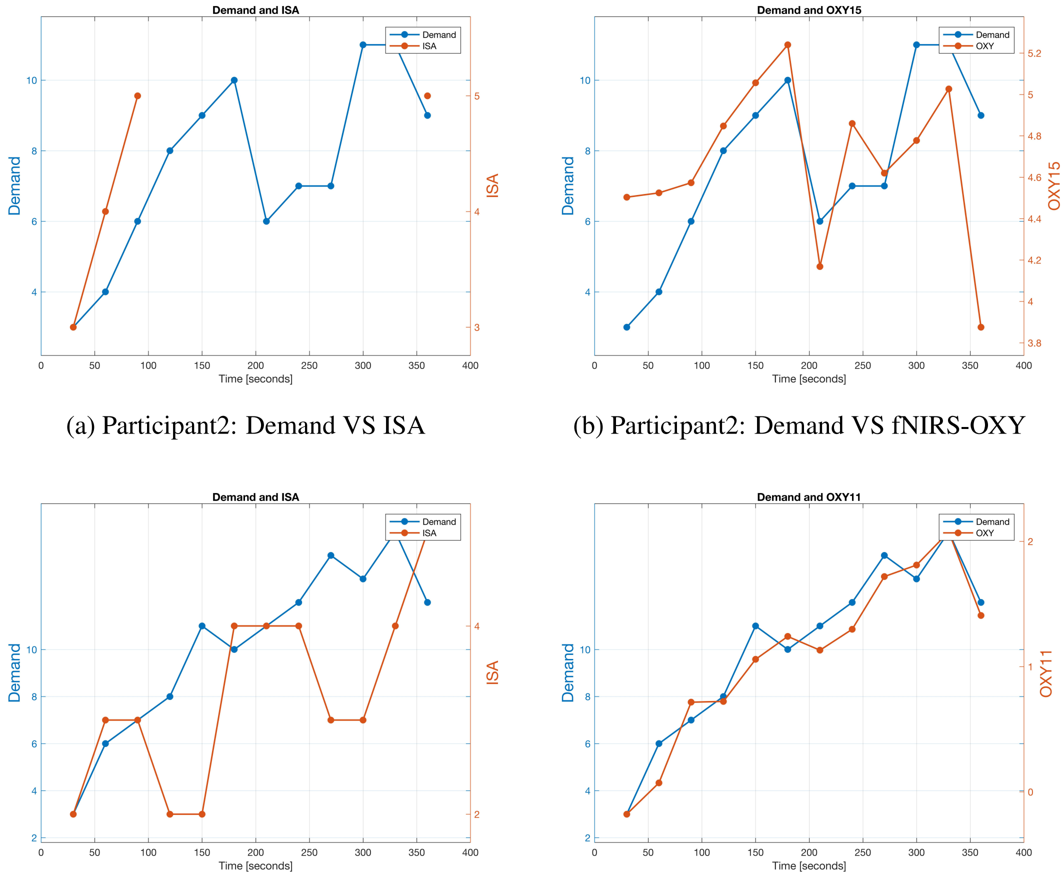
Below, we address each of the four hypotheses in subsections; statistical tests were conducted across both phases to determine whether feedback or ISA conditions had any impact on task performance and workload. Additional between-phase tests were used, when relevant, to examine whether there was an effect created by the tertiary variable of feedback colour (Phase1 vs Phase2).

6.5.1 H1) Variation in task demand will create measurable differences in Mental Workload

To begin our analysis, we first sought to confirm that our measures of participant mental workload were affected by and related to the task demand. To do this, task demand was quantified as the total number of aeroplanes participant was monitoring every 30 seconds (the frequency at which ISA scores were collected). Below we analyse how both our subjective ratings (ISA) and objective measures (fNIRS) correlate with these task demands over time.

Subjective ratings from ISA

We found strong correlations between demand and ISA measures, with examples shown in Figure 6.6, further showing the hypothesis H1i in Figure 6.5. The average correlation value across all participants, between ISA and demand (measured every 30 seconds) was $r = .68$ with the maximum value of $r = .899(p = .006)$ for Participant4 in Phase 2. This correlation was strong for some participants, where P1's ISA correlation with task demand in Phase 2 during the ISA Condition, for example, was $r = .808, p = .003$, and $r = .751, p = .003$ in the Feedback+ISA condition. There were, however, several cases across participants, where ISA did not reflect well its relationship to task demand, such as when participants were either too busy or to focused on the task and thus did not respond to the ISA questions (Figure 6.6a). This range of correlations highlights one of the known limitations of using mid-task self assessment scales, as they rely on operators rating their workload during tasks.



(a) Participant2: Demand VS ISA

(b) Participant2: Demand VS fNIRS-OXY

(c) Participant6: Demand VS ISA

(d) Participant6: Demand VS fNIRS-OXY

Figure 6.6 Emphasizing the value and limitations of ISA (when participants fail to self-report values of their workload (a) it becomes hard to understand what happened during the task).

Objective measures from fNIRS

As described in the previous subsection, ISA was not always able to reflect the user's state (See Figure 6.6a and Figure 6.6c), and mainly because it relies upon users subjectively reporting how they *feel*. For the same conditions and the same participants, Figure 6.6b and Figure 6.6d show how OXY correlates more objectively with the task demand. The average correlation value across participants, between fNIRS OXY and demand (measured every 30 seconds) was $r = .81$ with the maximum value of $r = .973(p = .001)$ for Participant6 in Phase 2. This shows how fNIRS could be used to assess workload without relying on a subjective measure such as ISA.

Summary of H1 Results

Based on these results, we are able reject the null hypothesis and accept H1, as both subjective and objective measures provide evidence that participants' mental workload was associated with task demand. We conclude, however, that our objective measure (hypothesis H1f in Figure 6.5) was able to provide stronger and more consistent evidence of increased workload than subjective ratings (hypothesis H1i).

6.5.2 H2) Participants' performance will be affected, positively or negatively, when made aware of their mental workload

Considering all participants across both phases, a series of two-way repeated measure ANOVAs showed no statistical significance in the three performance measures (Total Departures, Total Landings and Total Performance) between the four conditions. Total Performance, shown in Figure 6.7 for example, was different between the two phases. Similarly, the number of departures[2] and landings varied between phases, especially during the presence of both Feedback + ISA (see Figure 6.8). Participants performed slightly worse in Task Only condition in Phase 2 (the mean Total Departures was 10.2 for Phase 1 and 11.13 for Phase 2, and the mean Total Landings was 18.15 for Phase 1 and 17.36 for Phase 2), and slightly better in Feedback+ISA condition in the same study, compared to Phase 1 (the mean Total Departures was 11.1 for Phase 1 and 9.93 for Phase 2, and the mean Total Landings was 16.46 for Phase 1 and 18.43 for Phase 2).

Phase 1 performance data

To consider the performance in Phase 1 (white light changed to red in periods of high workload as detected via fNIRS), we examine the Total Landings and Total Departures data over the four conditions in **Phase 1**, shown in the blue bars of Figure 6.8. For all the cases, the Total Departures, the Total Landings and the Total Performance measures, performance appeared to decrease in the presence of ISA, suggesting that ISA might have a negative effect over the average participants' performance, and therefore affecting hypothesis H2i in Figure 6.5. This is not the case in the presence of Feed-

[2] There was one outlier in the Total Departures and the Total Performance data, which had a studentized residual value greater than ±3. The outlier was removed from the analysis.

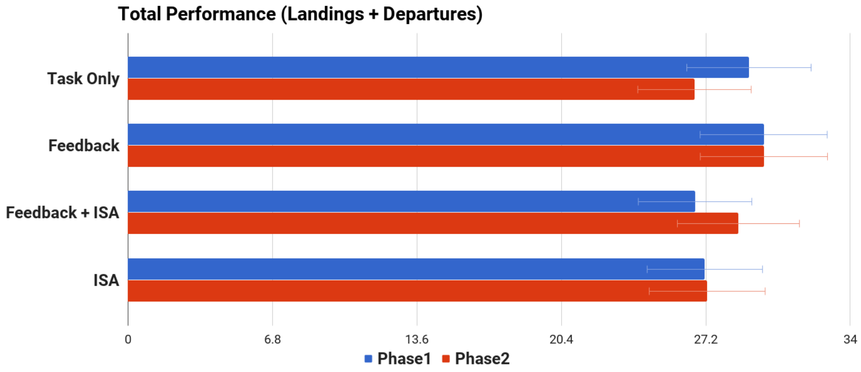
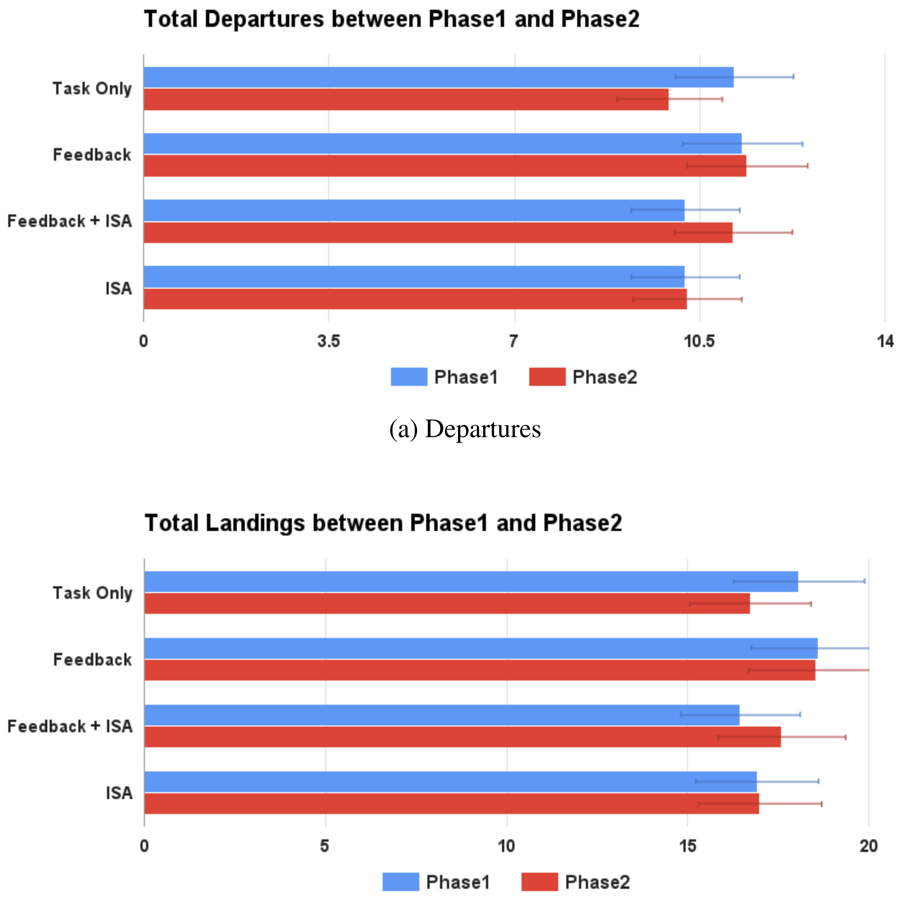


Figure 6.7 Total performance difference between Phase 1 and Phase 2 - mean scores across conditions

back, therefore, hypothesis H2f presents workload feedback having no explicit negative effects on performance.

There was no statistical significance found in the Total Departures with feedback impact $F(1,12) = .055, p = .819$, and ISA impact $F(1,12) = 2.476, p = .142$ as assessed by a two-way repeated measure ANOVA. There was also no statistically significant two-way interaction either between Feedback and ISA effect, $F(1,12) = .014, p = .907$. For the Total Landings, the presence of feedback showed no significant impact $F(1,12) = 1.147, p = .305$, however ISA significantly reduced performance $F(1,12) = 5.637, p = .035$ as assessed by a two-way repeated measure ANOVA. These results indicate that participants who responded to the ISA scale during task performance performed less well on the task, hence, hypothesis H2i in Figure 6.5 suggests a negative impact of ISA use on performance. There was no statistically significant two-way interaction between Feedback and ISA effect, $F(1,12) = .014, p = .907$. A two way repeated measure ANOVA additionally showed the impact of ISA on the Total Performance measure $F(1,12) = 5.368, p = .039$, and no effect of the presence of feedback was found $F(1,12) = .675, p = .427$, nor the two way interaction between the two, $F(1,12) = .007, p = .937$. From this we conclude that objective feedback provided via the change in lighting colour had no explicit negative impact on performance scores, and thus all the significant differences are caused by the deployment of ISA.



(a) Departures

(b) Landings

Figure 6.8 Performance data across the 2 phases

Phase 2 performance data

In **Phase 2** (red light changed to white in periods of high workload as detected via fNIRS), looking at the red bars in Figure 6.8, the negative effect of ISA was not found to be significant. Instead, the graph suggests a higher average performance during the presence of Feedback, after the colour change via lighting. This may suggest that workload feedback, and hypothesis H2f in Figure 6.5 may have a positive impact on performance.

Feedback showed no significant impact on performance with Total Departures $F(1,14) = .008, p = .932$, Total Landings $F(1,13) = 0.127, p = .727$, and Total Performance $F(1,13) = 0.072, p = .793$. ISA had no longer significant impact on performance, with Total Departures $F(1,14) = .229, p = .639$, Total Landings $F(1,13) = .011, p = .919$, and Total Performance $F(1,13) = 0.064, p = .804$ There was a statistically significant two-way interaction between ISA and Feedback effect on all performance measures;

Total Departures $F(1,14) = 7.565, p = .015$, Total Landings $F(1,13) = 6.475, p = .024$, and Total Performance $F(1,13) = 9.388, p = .009$. The average performance across participants was lower in the absence of Feedback and ISA, compared to all other conditions as shown in Figure 6.8.

Summary of H2 Results

Based on the results above, we reject the null hypothesis and accept H2 for ISA only, and not for our WFS. Overall, we found that performance was negatively impacted by ISA - an effect that was exaggerated when also being given feedback by our WFS - but overall we did not see performance being impacted by the WFS alone.

6.5.3 H3) Participants' perception of their performance and workload will be affected, positively or negatively, when made aware of their mental workload

Perceived Performance Scores

A five point rating scale was used to capture participants' subjective perception of performance after each condition. Across both phases, a Friedman test was conducted to understand the within-subjects effects between all levels of the two factors: Feedback and ISA on *perceived* performance scores with Feedback-NoFeedback x ISA-NoISA conditions. Results showed a statistically significant difference between conditions, $N = 30, X^{(2)} = 9.072, p = .05$ (Figure 6.9). Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons. There was a significant difference in the perceived performance scores between Feedback+ISA and Task alone $p = 0.05$, and between Feedback+ISA and Feedback condition $p = 0.05$. Figure 6.9 shows how participants' perceived performance was significantly lower when ISA present. These results show how ISA significantly reduces participants' perceived performance (hypotheses H2i and H3i), while the presence of feedback has no negative impact (hypotheses H2f and H3f).

For each of the two phases, the subjective performance scale generally showed lower perception of performance during the presence of ISA as showed in Figure 6.10, this being somewhat expected. This is directly related to hypotheses H4i, H2i, and H3i

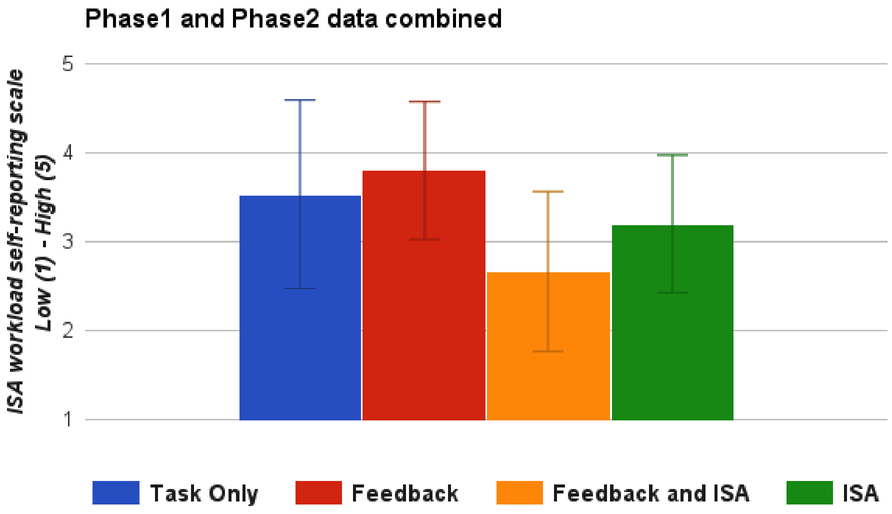


Figure 6.9 Combined Phase 1 and Phase 2 mean perceived performance scores across conditions

in Figure 6.5. It is interesting to observe how the perception of performance increased during the presence of ISA for Phase 2 compared to Phase 1, this effect being significant in Feedback+ISA condition. This suggests an impact caused by the feedback type. To investigate the within-subjects effects between all levels of the two factors Feedback and ISA on *perceived* performance scores, a Friedman test was conducted with Feedback-NoFeedback x ISA-NoISA conditions for each phase separately.

Phase 1 perceived performance

In **Phase 1**, the test showed statistically significant difference between conditions, $X^{(2)} = 12.756, p = .005, df = 3$. Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons. This shows that ISA presence significantly impacted the perceived performance scores in the presence of feedback $p = .003$, but also compared to the baseline condition $p = .013$. However, the presence vs absence of workload feedback had no impact on the perceived performance indicating that the presence of ISA negatively affected perceived performance. From these results we concluded that, in contrast to feedback and the hypotheses H2f and H3f, ISA and the hypotheses H2i and H3i significantly reduced task performance as well as perceived task performance; this was further significantly exaggerated by objective feedback.

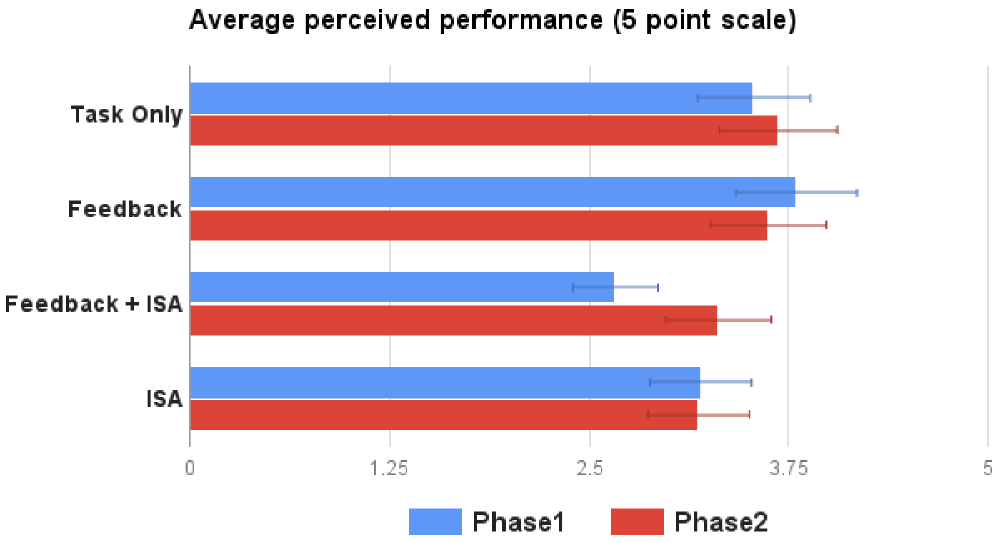


Figure 6.10 Mean perceived performance scores across conditions

Phase 2 perceived performance

Similar to the performance data, the subjective perception of performance data increased for the Feedback + ISA condition in **Phase 2** (Figure 6.10). The Friedman test revealed no more significance in the perceived performance data between conditions, the presents of ISA having no longer significant impact $X^{(2)} = 1.38, p = .71, df = 3$.

Comparing Phases

A Mann-Whitney U test was conducted for each of the four conditions between the two studies to determine if there were any significant differences in perceived performance scores between Phase 1 and Phase 2. The test showed statistical significance between the two studies in the presence of both workload feedback and ISA. The distribution of perceived performance scores between Phase 1 and Phase 2 was different as assessed by visual inspection. Mean rank value for Phase 1 was 12.47 (N=15) and for Phase 2 18.53 (N=15). The test showed significant statistical difference in perceived performance scores between the two studies in the Feedback + ISA condition, with $N = 30, U = 158$, and Asymptotic Sig. (2-sided test) $p = 0.05$. This finding suggests there was a difference between the feedback type used, with Phase 2 type having higher perceived performance scores.

Summary of H3 Results

We found that the deployment of ISA significantly reduced participants' perceived performance (hypotheses H2i and H3i), while the presence of feedback had no negative impact (hypotheses H2f and H3f in 6.5). However, as with actual performance, it is interesting to note that the mean ISA score (across participants) revealed higher perceived workload in the presence of our WFS (see Figure 6.10). We again conclude that the impact of deploying ISA was exaggerated by additional demands, since participants' perception of workload was affected by presence of ISA, but not by WFS alone. Overall, we have enough evidence to reject the null hypothesis and accept H3 for ISA, but not in the case of the WFS.

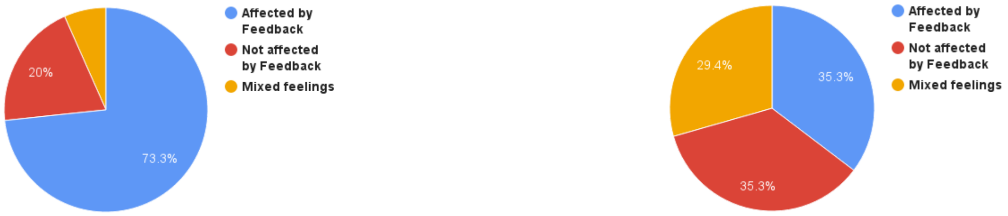
6.5.4 H4). Participants' perception and management of the task demand will be affected, positively or negatively, when made aware of their mental workload

Each participant took part in a short post-experiment interview about their experience during the study. The focus of this was to collect opinions related to perception of mental workload feedback, ISA, the way they foresee feedback of workload in their every day lives and their views of its use in a critical jobs scenario, similar to the task they performed.

Impact of Feedback on participants

In Phase 1, 11/15 of participants reported feedback affected them, 3/15 reported feedback did not affect them in any way and just 1/15 was not sure about it (See Figure 6.11a). In Phase 2, only 6/17 reported feedback affected them, 6/17 reported feedback did not affect them, and 4/17 were not sure about it (See Figure 6.11b). This finding suggests that one particular feedback type (Phase 1) had more of an effect on participants, or the case where feedback in Phase 2 was more transparent, hence, not directly affecting participants' perception. However, it does not reflect the type of effects (negative or positive) the feedback had.

Participants were affected in different ways by the feedback. Some participants described feedback as a good indicator of "how much" is going on during the task; P11



(a) Phase 1

(b) Phase 2

Figure 6.11 The impact of different feedback on participants

(Phase 1) gave an example of how he/she used the feedback during the task: *"helpful in that you knew there was a lot going on and you are concentrating, which meant you had to pay extra attention to the details. When the lights went red, it meant that, yeah, I am probably concentrating a lot, that means I am probably miss some of the smaller aircrafts, and I would try and go back and have a look around to see if I was missing any small planes"*. P9 (Phase 1) had similar feelings *"They did help me to figure out how I was feeling, and how I was going. Sometimes when it went bright red, I thought to myself, I need to be calm, and think for a second what was going on in the game again. It made me take a second and just relax, that is all"*.

In other cases participants felt that feedback was stressing them even more and making them feel anxious: P1 (Phase 1) said *"So whenever it went red, it was kind of stressful ... and I felt like why are you showing me red if I am stressed? ... It was NOTICEABLE! So in the sense that if I was doing a thing and I was stressed as I was playing that game, and I was trying to focus on the game, then I am being told that I need to focus more on the game, than that does not really help me, does it?"*, and P7 (Phase 1) said *"It does not really inform me about my next decision, because I did not stop even when the lights were red. It actually added to the stress. I tried to be calm..."* Similar feelings were found in participants' opinions for Phase 2. P17 (Phase 2) used the feedback to 'relax' during stressful moments *"I enjoyed the experience and I think the feedback is very... indicative. Though not very specific... But for the person interacting with the system, is like.... oh now I am tensed... maybe I should relax"*.

Even though some participants noticed and used the feedback, others ignored it, and better focused on the task, P22 (Phase 2) said *"I did notice it. I did not pay attention to it."* and P28 (Phase 2) *"I did not worry too much because sometimes you forget about* *feedback”*.

Validity, Accuracy, and Delay

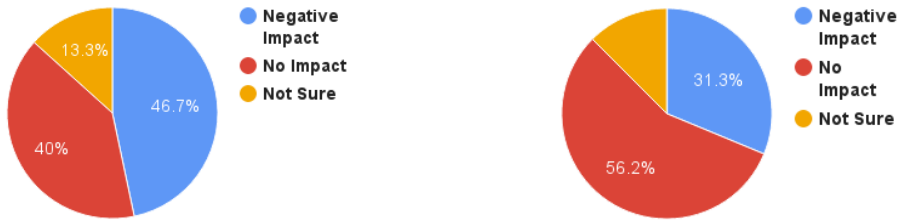
Participants also questioned the validity and correctness of the feedback. There were mixed views on what the feedback was informing, and how it worked. P29 (Phase 2) said *“I noticed that when I would get calmed they would change to red. But sometimes I thought that I was working and they would still be red.”*, P2 (Phase 1) said *“...most of the times it was accurate...”*, and P4 (Phase 1) said *“I was impressed though, they seemed what was not immediately responsive, but accurate. They seemed to change red when I was under a higher workload. When I was under more stress I said: crap...another plane is coming in, they will crash!”. It did seem to pick up on that quite a lot which was cool”*. These comments imply that, despite being able to perceive the delay caused by the hemodynamic response [136], participants found the WFS to be mostly accurate to their current perceived workload.

Granularity and other limitations

One of the limitations identified by participants during the study was the granularity of the WFS feedback; P1 (Phase 1) said *“I was a bit annoyed in a way, because the changes were not gradual, it is like uh, it is now white and it is suddenly turning red ekhhhhh (electroshock noises) ... I think if I were to align it with what I felt my workload to be at that point in time, then the colour changes would be much more frequent”*, and P13 (Phase 1) also suggested a more transparent modality of communicating the feedback *“...they seem to switch from an extreme to another and apart from them being in my eyes and bothering me while I was looking at the screen I didn’t really pay attention to them”*. P14 (Phase 1) added to this *“If it was in the background more in the background it would have been nicer. Now it was straight in your face”*.

Impact of ISA on participants

During the interview, participants were also asked thoughts about ISA, its use, whether they think ISA had any impact on their performance and whether it was ambient or distracting by nature. We found 7/15 participants in Phase 1 and 5/17 in Phase 2 believed their performance was worse because of it, 6/15 in Phase 1 and 9/17 in Phase 2 believed



(a) Phase 1

(b) Phase 2

Figure 6.12 The impact of ISA on participants perception.

it had no effects and the rest were not sure about it (see Figure 6.12a and Figure 6.12b). One participant reported that ISA had no impact on them, however, in the presence of both, feedback and ISA, it made him/her think: P16 (Phase 2) saying: *"No. It did not bother me that much. But at times it made me think, during the condition with feedback as well, especially when there was a discrepancy between the two"*. The majority of participants who reported ISA having a negative impact on them, also reported losing focus when having to answer ISA questionnaire during the task. P26 (Phase 2) said *"It definitely made it worse ... because it takes you out of the action, and then it takes a little while to figure out where you were"*. P31 (Phase 2) reported that ISA had a continuous negative impact, *"Sometimes I would hear the notification in my subconscious, and I did not pay attention because I was very concentrated on the task"*.

Others said they ignored answering the questions when concentrating on the task: P7 (Phase 1) reported *"I do not think I paid much of attention of answering the questions. I know I missed some, and for some questions I did not really think about the question, I just answered it"*, and similar view was found with P9 (Phase 1) *"I forgot about the questionnaire as the task demand went higher. I completely blanked out. I focused on it when I was relaxed. When I got busy it went down my mind. Ignored it sometimes..."*. Some participants, however, considered ISA easy and fast, and ISA presence improving their performance in some cases: P8 (Phase 1) reported *"...it is very easy it took less than a second, and the buttons were really big. I think it was a normal performance, even better than normal"*, and P13 (Phase 1) had similar feelings *"just another button I had to press, I do not think it has an impact on my performance"*; P21 (Phase 2) described ISA as an *"automatic move"*.

Summary of H4 Results

This hypothesis was mostly examined through the post-experiment interview and participants' opinions, since we did not have any objective measurements of how people reflected on the feedback mid-task. It is clear that both ISA and feedback had an impact on participants' management of the task. Although the opinions were divided, and participants were "affected" differently by the feedback and ISA, the findings were very constructive and informative. We found insights to confirm that participants noticed the feedback and considered it at a meta-cognitive level during the task, and therefore have enough evidence to reject the null hypothesis in case of H4. In comparison to ISA, P24 (Phase 2) said *"I think the lights are more effective, because the cellphone app (ISA) just made me feel more busy. They lights show me when I am busy, where ISA made me feel busy"*. In the discussion section below, we consider what future work may do to investigate this finding in more detail.

6.6 Discussion and summary

This chapter describes an alternative way to raise self-awareness of mental workload, through providing workload feedback based upon a concurrent objective measure, and our results showed that it did so without negatively affecting performance (as with ISA). We expected that if users are alerted that they are approaching a drop/dip in performance because of high/low workload, then they might be able to take action to avoid it. Table 6.3 summarizes the results of the study that relate to the relationships presented in the adapted version of Sharples & Megaw's Framework for Mental Workload Measurement [121] (shown earlier in Figure 6.5).

6.6.1 Impact of ISA

In line with the findings of previous work [81, 133], we also found that self-reporting mental workload through ISA had a significant impact on both actual and perceived performance. In the task itself, participants landed significantly fewer planes than in other conditions. As is also typically expected with ISA, we saw many participants miss ISA entries when under high workload, and were often surprised when we showed them gaps in their self-reporting. P13, who missed several ISA responses, said: *"I did not* *know if there was a time limit I had to answer. I do not think I missed any"*, but said ISA was *"just another button I had to press... it became a mechanical task"*. Conversely, P14 said *"It was annoying! I was OK, go away, go away... It is like an alarm in the morning"* and P15 said *"Annoying..your phone app workload questionnaire is really annoying... I did not notice the phone sometimes when I was concentrating on the task, so completely ignoring ISA. Sometimes people get easily distracted and for this kind of task it can be dangerous..."*. Even though the general feeling was against ISA, some participants' perception of ISA was not that bad; P3 said *"I do not think that ISA had an impact on performance..."*. These sentiments were generally observable in the data, and so our findings match the consensus of prior research into both its validity for measuring mental workload (since it had strong correlation scores) and the interference it has on the primary task.

6.6.2 Impact of Objective Feedback

The aim of this study was to investigate whether presenting users with real time mental workload feedback, would make them aware of their load without notably reducing either actual or perceived performance. This was presented in contrast to ISA measure, which requires the user to reflect upon their mental workload and take action to report it. In our results, feedback did not affect actual and perceived performance in a negative way. Although not significant within our sample, feedback appeared to slightly improve actual performance and participants perceived that they performed slightly better (Feedback Condition - See Figure 6.10). The findings alone, however, do not tell us whether participants noticed the feedback, understood their mental workload, and took action to reduce them.

In interviews, some participants indicated that they did take note of the feedback: *"When the lights become red, it works as a reminder to take a big breath and relax [...] it is like a warning ..."* (P15 Phase1). For some, this was positive, with P14 (Phase 1) saying *"I really liked it! The whole experience ... If I would get another chance I would do it again"*. Some, however, were frustrated that they couldn't do much about it: *"It is actually affecting me. When the feedback is red, I try to relax. To try to make it white. But it did not work, because I felt even more concentrated ... because I was looking at the planes and to the lights as well, so it added up really to my concentration"* (P14).

Other participants felt that feedback had no use as they already know when they are busy and when not: P1 (Phase 1) said *"I felt like: why are you showing me red if I already know that I am busy?"* even though he said later that *"I am usually really bad at judging my own workload"*. These insights confirm that participants noticed the feedback and considered it at a meta-cognitive level during the task. For some participants the feedback, however, perhaps increased the sense of anxiety (especially in the first phase), when participants were not able to take action to change it. Because of this, P15 (Phase1) went further to suggest that although the objective measure of their mental workload was useful, they would have preferred to see it afterwards, rather than during the task: *"I would like to add that it would be much more interesting for me to have a feedback to reflect on but not a concurrent one. So maybe record it and reflect on it later on."*. This may be an interesting area of future work, as a mental workload parallel to life-logging and tracking daily fitness activity - a form of *Mental Workload Fitness* tracker. Further more participants suggested various levels of feedback would be much more useful *"rather than a cut off point ... a gradual transition in a way"* (P1 Phase 1).

Overall, the results show strong support for a) helping people to reflect, in action, about their current mental workload but without negatively impacting performance or indeed their Mental Workload. We did not, however, manage to observe improved performance, nor changes in behaviour because of the feedback they received. We discuss these more below.

Does this mean making a single evaluation at the end of each experimental task or trial? Cr making measurements from moment to moment during each trial. I suggest making this clearer by stating what the time granularity for measurements was and perhaps elaborating on the difficulties of interpreting more continuous measurements.

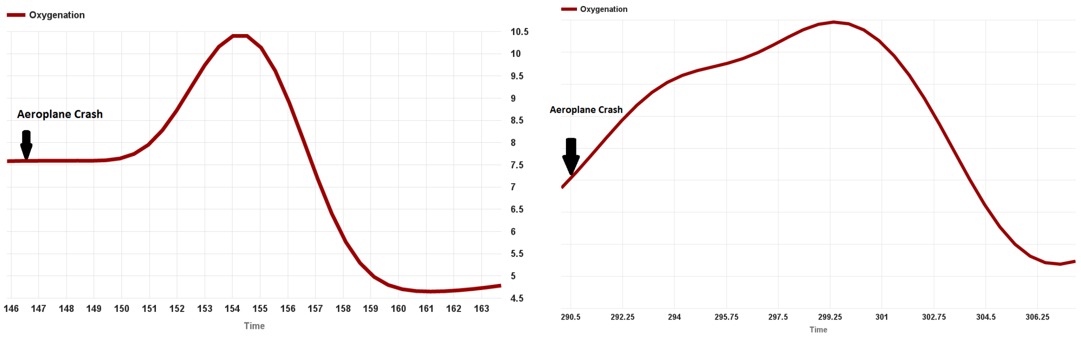
6.6.3 Continuously assessing mental workload

Continuous assessment of workload, as presented in this thesis, refers to the assessment of the operators' workload during tasks, and it is presented as opposed to post-task analysis - that is making a single evaluation at the end of each experimental task. This enables us to better understand what has happened during the task, having a grater granularity of data. In case of real time measurements, having a continuous technique will further enable direct intervention in case of workload overload/underload. In practice, however, there is much effort invested in filtering and processing the data after the experiment has been completed, as these steps require extra time and resources. Hence using a technique in real time might constrain the amount of data processing, therefore the quality of the results.

Beyond the challenges of real time processing of data, there are a number of interesting events that can occur when continuously assessing mental workload - we noted that mental workload fluctuated noticeably when aeroplanes crashed, and informally participants noted feeling stressed. It is interesting to consider what participants do in these situations, and what this might look like in mental workload data.

The Impact of Fail on Physiology.

We wanted to investigate what happened when participants failed to monitor and control all aeroplanes on the screen, and two or more aeroplanes ended up colliding. Figure 6.13a and Figure 6.13b show, for example, measurable changes in fNIRS OXY signal after such a fail, and its impact on Feedback. It is clearly important to consider whether this is cause or affect, but based on informal secondary analyses of our data, we saw many of these dramatic shifts in mental workload around fail scenarios.



(a) Example 1

(b) Example 2

Figure 6.13 Participant Oxygenation Levels measured with fNIRS after a “crash” event

In future work, we would like to more directly evaluate and, accommodate these reactions to events, whilst still giving people reliable feedback about their mental workload. Such future work may also focus on using other measures of mental workload, as emotional reactions are typically more observable through other physical reactions. We used an fNIRS device, which has been shown to be suitable for HCI user study evaluations, but more commercially available devices like the NeuroSky[3] EEG device might be more suitable for day to day feedback. Similarly, even less invasive measures of mental workload could be taken from Heart Rate Variability [59] through smart watches, remotely detected by pupil dilation [76] or facial skin temperature [128] with cameras. Many of these other measures also better detect emotional responses, and perhaps concepts like stress and anxiety, and so might better serve future work on recording both Mental Workload and emotional response.

The implications of using different feedback types

In general, the binary feedback of workload was alerting users of a high workload when a sudden increase in Oxygenation was detected using fNIRS. In the same way, a sudden decrease would cause a low workload alert after. The changes were visible to participants, such that they could monitor and use their workload feedback presented by the WFS. However, future work could first examine more granular forms of feedback, as noted qualitatively by participants. It was interesting, however, to first informally observe, and then analytically find differences between the choice of lighting feedback in two phases in the study. This post-hoc independent variable in our analysis revealed interesting results that confirm Sharples & Megaw's description that mental workload is closely affected by the way in which participants experience that workload. In Phase1, red colour was used for feeding back high workload states, and white colour for low workload states, and in Phase2 the colours were swapped. Having white light to alert of high workload made some participants feel "right" being on the white colour rather than red making them feel they are not working enough; P29 (Phase 2) *"...when they were red, I thought I am not working enough. When they were white it felt right, it felt that I was paying a lot of attention, it was in the right track"*. On the other side, having the red colour to alert of high workload generated pressure when "being" on red, P9 (Phase 1) reported that *"When the lights become red, it works as a reminder to take a big breath and relax"*. It would be extremely interesting in future work to artificially manipulate changes in feedback, and to observe changes in mental workload in a similar way to when participants experienced crashes (like in Figure 6.13). Such an analysis would help us to separately examine the impact of mental workload created by feedback and

[3] http://neurosky.com/

mental workload created by task demand. Future work should also, therefore, explore the design and mode of feedback, as well as the granularity of feedback.

Behaviour change

One large research area is behaviour change, and this study was not designed to measure and observe it. Although this study was not focused on measuring in-task behaviour change, qualitative anecdotes imply that people did reflect on their mental workload and considered their current status. It would be highly interesting in future work to more directly study whether or not there are behavioural markers for when participants take action based on their feedback. Such work would need more accommodating task conditions that allow people to manage, delay or even share their workload with others. We consider this avenue of research to be a very interesting direction for the future.

6.7 Summary of Chapter

In this chapter the last study of this thesis was presented attempting to understand whether brain sensing techniques, which are increasingly becoming commercially available, and in particular fNIRS, could be used to give people concurrent feedback about their Mental Workload levels. Although existing techniques, like the Instantaneous Self Assessment (ISA) tool, are designed to help people to report and reflect on their current Mental Workload levels, they also often have a negative impact on the primary task at hand. We hoped that, with objectively measuring and providing concurrent feedback during tasks, participants would be able to reflect on the mental workload levels, without the associated performance drops.

In order to capture and understand these effects, we have adapted the Framework for mental workload evaluation (presented in Figure 2.8). We “controlled” the “external factors” presented in the relationship 5 of the framework (in our case the presence of ISA and the presence of Feedback). Our results first confirmed both approaches to measuring Mental Workload during tasks, accurately correlating the measures with task demands, this further contributing to the findings in previous chapters, validating and understanding the sensitivity of fNIRS measure of workload. We then confirmed prior research findings that self-reporting techniques had an impact on both actual and perceived performance, as well as increasing the task demands on the participants. Our results, however, showed no such drops in performance were found with our Mental Workload Feedback System. Using the framework, we confirmed the existing relationships between the physical and cognitive task demands, and the operator workload (see relationship 1 in Figure 2.8), as well as the direct connection between operator workload and performance (see relationship 2 in Figure 2.8). Further, our interviews confirmed that feedback led participants to think metacognitively during tasks, but that the choice of feedback (using red lights to warn them of high Mental Workload) created a negative stressor to their experience. This effect was removed after changing the choice of colour in our feedback mechanism. The results suggested that participants do use the feedback of workload, therefore showing the relevance of relationship 3 in the same framework. Relationship 4 was not directly studied, however, the future work section presented discusses an interest in associating various physiological reactions to events (e.g such as task failure).

We conclude that objectively measured concurrent feedback of Mental Workload can help people to understand and actively manage their behaviour during tasks, but without the negative affects on performance created by self-reporting techniques. Such personal insight would be important for safety critical tasks like Air Traffic Control, but has the potential for a much wider impact, helping the general population to understand and manage their own mental workload across the many tasks that fill our lives.

The future direction of this work could be assessing the workload of everyday tasks, outside controlled lab-settings, moving away from brain based sensors towards using less invasive physiological techniques (e.g. HR, EDA, BHP). Another direction of research could be studying reflection of workload extensively, and understanding how people could actually use workload feedback to reflect on the work/break patterns throughout the day.

The following chapter presents the discussions and conclusions of this thesis.

Table 6.3 Summary of key findings, by hypothesis

Hypothese	Expected effects	Results
H1	Participants' workload generated by the task demands would have measurable effects with ISA and fNIRS	We found both fNIRS and ISA measures sensitive to task demands (in our case the number of aeroplanes to control during an ATC game). Overall, we found high correlation coefficients between fNIRS and demand, and we showed how it can be used to assess workload without relying on participants' ability to self-report during the task (See Figure 6.6).
H2 & H3	ISA would have a negative impact on performance. Feedback would have no explicit negative impact on performance. Workload and perception of performance would increase or decrease in the presence of both ISA and our WFS.	We presented evidence supporting ISA's negative impact on both performance and perceived performance measures. In contrast to ISA, we found mental workload feedback having no explicit negative impact. Instead, Figure 6.10 suggests similar or slightly better performance with our WFS. We found no significant evidence of our WFS increasing performance or participant's perception of performance.
H4	In contrast to our WFS, ISA will create additional physical and cognitive task demands.	Although we found no direct evidence, the performance and perceived performance results suggested a negative impact in the presence of ISA, most likely due to the additional required resources. This was not the case with our WFS. During the interview, participants had mixed feelings about the impact of ISA; views were divided into participants affected by ISA, participants who considered ISA having no negative impact (describing it "easy" and "fast"), and participants who ignored ISA when concentrating on the task.

Chapter 7

General Discussions and Thesis Conclusions

In order to design interactive systems (such as tools for the digital economy) to be used by people, one should take into account users capabilities and limitation (e.g. their cognition). Cognition refers to memory, attention, the amount of information we can "handle" in a given time, and the ways we solve problems and make decisions. It is really important to understand how we use and deal with information, so we can design systems that support, rather than mitigate users during interaction. For example, if we are designing an interactive system for a car, we need to consider the existing demands on the users (drivers). Drivers already have a high visual load during driving, so it may not be appropriate to design a visual display of a high complexity that demands the driver's attention for a long period of time. Instead, a different interaction modality could be considered (see Wickens MRM [143, 144]).

This thesis was largely focused on this matter, exploring physiological methods to learn about user's state during interaction, and in particular the assessment and feedback of users' mental workload during tasks. The first part of the thesis was focused on understanding the suitability of fNIRS during Human Computer Interaction. We investigated how various artefacts typical for lab-based evaluation settings impact the fNIRS signal and provided guidance on how to use the technology in such settings. The next part presented the suitability of fNIRS for assessing workload, we therefore investigated the replicability, reliability, sensitivity and validity of the measure. This was continuously investigated throughout the thesis. We showed that there is a relationship between fNIRS and subjective techniques, including NASA-TLX, however the relationship was much stronger in contrast to a continuous subjective measure, such as ISA, as proposed in Chapter 5. The last part of the thesis investigated the use of a continuous, real time version of fNIRS technique to assess and feedback workload in real time. We found that people could self-reflect when given workload feedback during tasks, without side effects on their performance, and this area of research could be further explored.

7.1 Summary and contributions of the research to theory and practice

The contributions of this thesis are:

- This thesis contributes to the measurement and assessment of workload using fNIRS. The reliability of the measure was tested within lab-based evaluation settings, and we extended the understanding of its use during both verbal and spatial tasks. This was presented in Chapter 4.
- This thesis further contributes by testing the sensitivity and validity of the technique, and extended our understanding of workload in relation to performance measurements, subjective techniques and physiological methods using fNIRS. We have also presented the challenges of using fNIRS continuously during tasks. This was presented in both Chapter 4 and Chapter 6.
- The last contribution of this thesis, presented in Chapter 6, is exploring the potential impact of presenting users with concurrent feedback of their workload during tasks. We investigated how feedback of mental workload (based on real time measurements during tasks using fNIRS) could be useful to people, and we showed how people think metacognitively about their state during tasks.

To preserve naturalistic interaction settings, the methods and sensors used to collect useful data about the users during interaction should ideally allow a normal, unrestricted interaction, with minimal controlled settings. Solovey et. al. investigated the potential of using fNIRS [127] in such settings, and reproduced artefacts normal to a typical evaluation study settings, to measure their effects on the fNIRS signal. Chapter 4 presented a study focused on investigating the the sensitivity of fNIRS measure to workload, but also the reliability and replicability of the results but reproducing and confirming some of the Solovey et. al. [127] study settings and results. This is particularly important in the context of using fNIRS as a technique to assess workload. Sharples and Megaw present the property of sensitivity and reliability as key aspects when establishing a workload assessment technique. Solovey et. al. results were confirmed and new insights were drawn by extending the original study, and exploring the reliability of the fNIRS measure in the presence of movement artefacts during a spatial memory task of remembering a 6x6 shaped grid (the original task was a verbal memory task of remembering a 7 digit number). Chapter 4 addressed in particular the first research question RQ1 with the associated subquestions RQ1a, RQ1b, and RQ1c.

As one of the aims in this thesis was to use fNIRS to provide real time mental workload feedback to users during tasks, the second research question RQ2 with the related subquestions explored the capabilities of using fNIRS as a sensitive measure to continuously assess workload, and the challenges of using the measure in real time. Chapter 5 addressed some of the above mentioned issued, but also contributed to RQ1b, and further validated the measure by understanding its relationship to the widely used subjective technique, the NASA-TLX questionnaire. The results of the study presented confirmed previous findings such as the results from Peck et. al. [105], and a few discussions are presented in the next sections of this chapter.

The third and the last aim with associated research question (RQ3) explored how a continuous, real-time version of fNIRS can be used to assess peoples' workload in real time, and provide them with workload feedback during tasks. This system was called the Workload Feedback System WFS (see Figure 6.1), and used a passive BCI to measure, classify, and feedback users' workload in real time. The WFS was presented in Chapter 6, however, the research was focused on exploring how WFS can be used to allow people to understand, manage, and reflect over their workload during tasks, by presenting them with workload feedback based on measures of brain activity.

7.2 Using fNIRS in typical evaluation and user testing settings

To be suitable for studying human interaction with technology, the sensors and techniques used to collect useful information about users during interactions should ideally be as transparent as possible - allowing a naturalistic interaction - therefore not heavily restricting users from their normal way of interacting with technology or performing various tasks. The study presented in Chapter 4 aimed to investigating the effects of common human behaviours on fNIRS ability to distinguish states of cognition from other states, replicating and extending the work of Solovey et. al. [127]. Solovey identified and studied using a verbal memory task, a number of four typical artefacts in such stetting and their impact on fNIRS signal, including: head movement, keyboard input, mouse input, facial movement, and further investigated a control condition in the presence of no artefacts. Table 7.1 presents the conditions investigated in the original study, but also shows the novel contributions investigated in this thesis. Chapter 4 presents a study replicating three of the original artefact conditions, and additionally identified and investigated the impact of verbalization as an artefact on fNIRS signal. Further more, the study extended our understanding of artefacts' effects on fNIRS during a Spatial task as opposed to Verbal alone, with the addition of a new spacial task of remembering a 6 x 6 shaped grid.

The fundamental finding confirmed in this study is that fNIRS can be used to distinguish between cognitive and rest states in both Verbal (as confirmed by Solovey et. al.) and Spatial tasks. Table 4.2 presented how artefacts affected the two task types, such that the significance between rest and task was sensed using different fNIRS measures for different artefact and task type. Therefore, the addition of a Spatial task, provided a greater understanding of fNIRS' ability to distinguish cognition under tasks using such encodings, and further stressing that both fNIRS key measures, Hb and HbO need to be considered during experiments.

Testing the reliability of fNIRS in the presence of artefacts, and further confirming fNIRS' ability to distinguish rest vs task times in the presence of the investigated artefacts regardless of task type (Verbal *and* Spatial) was at the base of this thesis. This further answered the first research question RQ1, allowing the next research questions

Table 7.1 Reliability of using fNIRS in typical evaluation settings

	Artefact Conditions			
	Control No-Artefact	**Head Movement**	**Keyboard Input**	**Verbalizing**
Verbal Task	Showed fNIRS Reliability in presence of various Artefacts during a Verbal task - as Investigated by Solovey et. al. [127].			Extended original study further investigating the impact of verbalizing -as an artefact on fNIRS signal.
Spatial Task	Extended original study further investigating the Reliability of fNIRS during a Spatial Memory task.			

to be investigated (research questions in this thesis were dependent on previous research questions - falling out of each other).

7.3 fNIRS as a continuous measure to assess workload during tasks

Perhaps the most important property of fNIRS in relation to workload assessment, and one shared amongst various other physiological techniques, is the continuous capabilities of reflecting users' workload during tasks. Compared to other techniques that restrict the workload assessment to a one-off measure, typically taken after the task has been completed, the physiological techniques provide granularity of data, further facilitating cause and effect analysis with data that reflects participants' experience during the task.

As physiological techniques, including fNIRS, do not involve any extra work to be done by the users during tasks, they have the potential to be used in settings otherwise hard to study using other techniques e.g. subjective techniques. Exploring the use of physiological techniques is therefore essential for the study of operators working within safety-critical systems such as ATC.

One of the aims in this thesis was to explore the use of fNIRS for continuous monitoring of workload during tasks. Chapter 5 explored and discussed the challenges of using fNIRS for continuous assessment of workload.

Chapter 6 explored how the measure could be used for the assessment of workload in real time, and how feedback of workload based on fNIRS objective assessment could help users during tasks. The contribution of this thesis in relation to fNIRS is not focused on the way the signal was analysed, nor exploring new data processing techniques; instead, it was much more focused on exploring the potential of using fNIRS as a useful measure during lab-based evaluation settings based on the existing knowledge of data processing and analysis.

Workload is a construct that sits in the intersection of multiple contributing factors (as presented by Sharples and Megaw [121] in the framework for mental workload definition and measurement - Figure 2.8), that could be categorized depending on the measurement technique. Because various techniques reflect different workload components, the studies presented in this thesis were designed such that a combination of performance data, subjective measures and physiological techniques (mainly fNIRS) are captured to better understand the relationship between them, and in particular, to understand the relationship between fNIRS and other measures of workload.

With this approach, fNIRS was found to be a complementary measure, providing useful information about users in a continuous manner during interaction.

7.4 Real time Mental Workload Feedback

The last study presented in this thesis explored how a real time, continuous version of fNIRS could be used to give workload feedback to the users during tasks. As this was not previously explored, we compared the workload feedback to the individual self-assessment technique where users would reflect and report their workload during tasks. We believed the two might have similar effects, potentially allowing users to self-reflect over their state and regulate their resources allocation to the primary task when reaching a high workload state.

Between the subjective self-assessment techniques, ISA was chosen due to its robustness and non-invasive nature, allowing self-reflection continuously during tasks, the same way feedback of workload would potentially allow it without the additional effort of self-reporting.

We have implemented the Workload Feedback System (WFS), that uses measurements of brain activity in the prefrontal cortex in order to assess, classify and feedback workload to users during tasks. Two states of interest were used for feedback in the study, high and low workload. The feedback and the WFS was specifically designed to be noticeable, but at the same time transparent and in the background of the task, such that a minimum of resources would be used by operators to perceive the feedback during tasks. The office desk lights were the means of providing the feedback. A dynamic lighting environment (DLE), designed to aid in an individuals self awareness when completing a task, was programmed to turn red (from normal white) when participants were experiencing high workload, and white (from red) when experiencing low workload. Midway through the study, the colours were reversed: turning white from red when participants experience high workload.

During an air traffic control simulator game participants were presented with (close) to real-time feedback of their workload. We found high correlations between task demand and the objective workload measure from fNIRS and we concluded that fNIRS has the potential to be a more reliable measure for detecting periods of high workload compared to ISA subjective ratings. We further found that performance was negatively impacted by ISA - an effect that was exaggerated when also being given feedback by our WFS - but overall we did not see performance being negatively impacted by the WFS alone. This effect was also significant in the case of perceived performance scores, we found ISA significantly reducing participants' perceived performance, while the presence of feedback had no negative impact.

In the post experiment interview, the opinions were divided, and participants were "affected" differently by the feedback and ISA, the findings were very constructive and informative. We found insights to confirm that participants noticed the feedback and considered it at a metacognitive level during the task, thus opening interesting research directions.

7.5 Specific instance of future research to initially test out our ideas

In this thesis, fNIRS was presented as a useful tool for assessing mental workload during tasks. Being a continuous measure, fNIRS allows users' workload assessment at every stage of interaction, and does not rely on operator reflecting and recalling their experienced workload during tasks or retrospectively. Being cheap and with a quick setup, this research showed how fNIRS can be used for both real-time use, and for post-experiments analysis, to get insights into human mental workload. Moreover, fNIRS was showed to be suitable and useful during typical evaluation study settings, without further restricting users during interaction.

Despite these key properties that makes fNIRS suitable for these types of study settings, there are other scenarios associated with limitations to the technology. One could consider measuring workload of everyday life, that does not happen in a controlled lab settings. In fact, this is one of the future works discussed in Chapter 6, where we used fNIRS to assess and feedback workload within controlled lab-settings during a game which simulated an ATC job. However, brain scanners are not commonly worn by the general public, on the streets or in the office, and alternatives should be considered for monitoring the workload of everyday life. If the works in this PhD are focused on using fNIRS as a technique to establish a consistent reference baseline for mental workload, the future research could explore the use of other physiological techniques, that are better to be used outside the lab settings. In order to do this, fNIRS could be used as an objective reference to mental workload, and other methods and techniques could be used in relation to fNIRS and subjective techniques.

7.5.1 Future work. Pilot Experiment.

The works in this thesis established a baseline of knowledge in the field of physiological measurement and feedback of workload. As a first step towards more complex applications of the present work we present below a pilot experiment investigating the relationship between mental workload, variation of performance and other objective physiological parameters in comparison to fNIRS. The aim of this pilot study is therefore to explore how other physiological methods that are less invasive compared to fNIRS could be used to reflect a similar assessment of workload outside laboratory settings, where fNIRS could be hardly used.

In order to control the demand placed on the participant, a specific computer based task was designed, that would impose different levels of experienced mental workload. The task consisted of a computer game played in two different versions. Using a Joystick controller, and a 50 inch TV approx 2 meters away from the joystick, the participant is presented with moving coloured balls on a black background. The movement of the balls gives the impression that they are falling from the top of the screen and the task is to aim and shoot the target balls using the joystick. See Figure 7.1 for a description picture of the task.

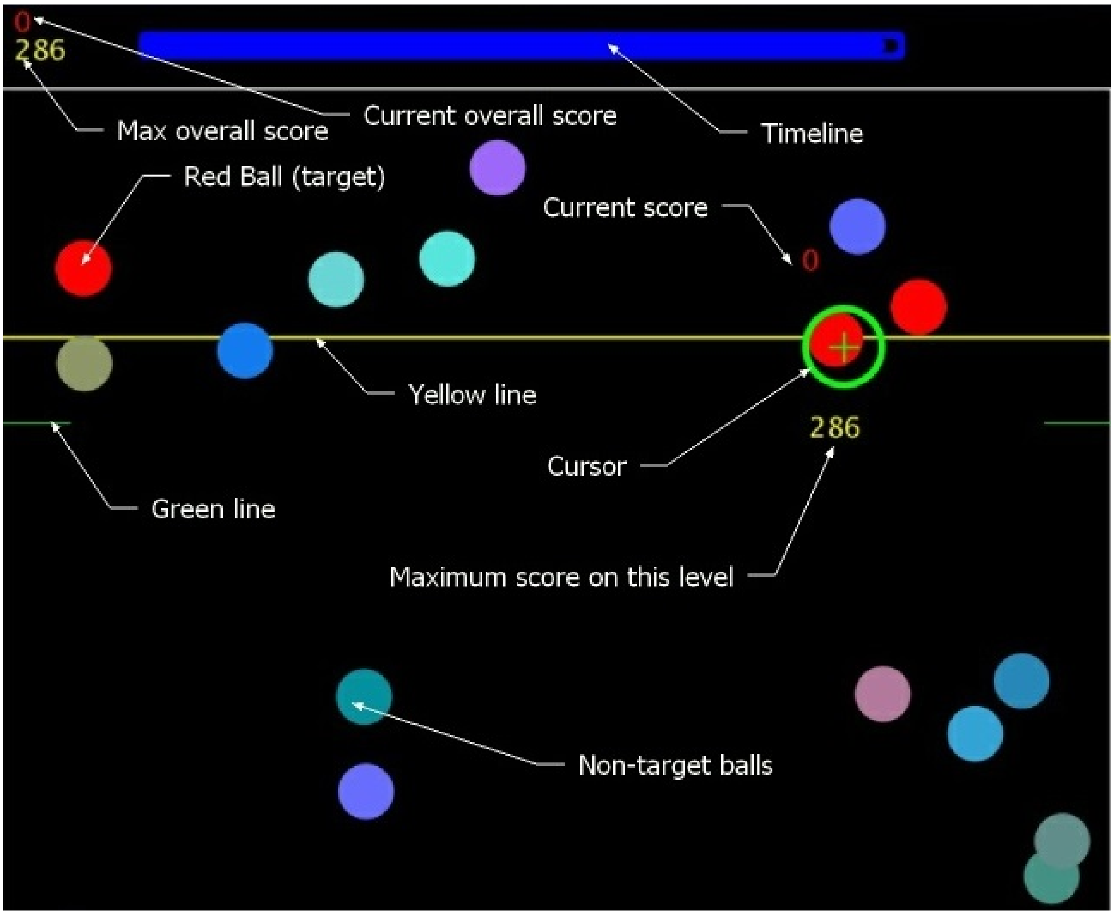


Figure 7.1 Pilot Study Task of shooting red balls.

There were three study conditions:

- In the first condition, participant is asked to only shoot the red balls as targets.
- In the second condition, balls have numbers on them, and participants are asked to shoot odd numbered ball regardless of colour, introducing an additional cognitive element with the intent of increasing mental demand.
- The third condition is identical to the first.

Each condition consisted of 13 stages (45 seconds each) of varying difficulty [120]. The number of target balls was varied in order to control the level of demand, ranging from 3 target balls to 15, and then back to 3. A yellow line which started at the top of the screen was dragged down with every miss shot, or with every target ball reaching it (see Figure 7.1). Therefore the position of the yellow line on the screen would reflect participants performance in relation to the success rate and demand at every point during the task.

Similar to the approach in Chapter 6, a 5-point version of ISA was used, with the participant self-rating subjective workload on a scale from 1 (low) to 5 (high). This version required the participant to verbalize the number, rather than use the mobile phone app due to the nature of the task. The scores were recorded on paper, and further digitalized and used for analysis.

The participant was presented with an information sheet and consent form prior the study, and was required to have no pre-existing heart/brain related condition and have no skin conditions or allergies that could prevent them from wearing the physiological sensors. The study was approved by the Faculty Research Ethics Committee at University of Nottingham.

We collected a variety of physiological measures including:

- heart R-R inter-beat intervals;
- breathing rate data;
- pupil diameter for both left and right eyes;
- fNIRS OXY measure (based on the CBSI filter that combines both OXY and de-OXY haemoglobin);
- skin temperature from E4;
- EDA from E4;
- BVP from E4;

The Zephyr BioHarness 3 chest strap was used for measuring posture, heart and breathing activity. The device outputs raw ECG data at a sampling rate of 1000 Hz and also a processed version of the raw signal including the R-R intervals and heart rate (Medtronic, Annapolis USA).

For eye-tracking, the RED 250 eye tracker was used in stand-alone configuration, measuring pupil diameter and gaze data at 60 Hz (SensoMotric Instruments, Teltow-Germany).

The Empatica E4 [46] is a hand wearable wireless multisensor device for real-time computerized biofeedback and data acquisition. In this study, the E4 band was used to capture electrodermal activity (EDA), skin temperature variations, accelerometers data, Photoplethysmography Data (PPG) - that measures Blood Volume Pulse (BVP), from which heart rate, heart rate variability (HRV), and other cardiovascular features may be derived. The main aim for including the E4 was to investigate the accuracy of such user-friendly device (compared to the Zwphyr chest strap) that efficiently combines 4 sensors into the wristband, replacing traditional multiple sources (e.g., heart rate chest strap, finger-placed EDA sensor, wrist worn accelerometers and temperature). Unlike traditional physiological acquisition devices, the E4 wristband can be worn during daily activities; therefore, the wristband is less likely to interfere with everyday activities, and could useful for research outside of the lab for monitoring the workload of everyday task.

Measures of brain activity were recorded using an fNIRS300 device and the associated Cognitive Optical Brain Imaging (COBI) Studio hardware integrated software platform provided by Biopac Systems Inc. A similar approach to the one presented in Chapter 6 was used for analysing fNIRS data.

There were 13 instances of the ISA questionnaire during each of the three study conditions to capture the participants' perceived workload while performing the task.

The aims of this pilot study were investigating whether there is a close relationship between demand, task performance and the different workload techniques. The results were promising and a full study is to be conducted and submitted for publication at ACM CHI2018 conference.

We expected to find:

- a negative correlation between perceived workload (ISA) and task performance (yellow line on the screen);
- a positive correlation between fNIRS OXY and ISA measure, as they are both continuous measures reflecting participants' workload during the task, and
- a negative correlation between fNIRS OXY and performance (the yellow line).

The findings in the pilot study showed a strong negative Spearman's correlation between the ISA scores and the performance measure $r = .921$ and $p < .001$. Figure

7.2 shows the close relationship between fNIRS and both performance data and ISA. There was a Pearson Correlation between fNIRS OXY and Performance data (yellow line), with $r = .716$ and $p < .01$ as well as a Spearman's correlation between fNIRS OXY and ISA $r = .725$ and $p < .01$. The results further confirm fNIRS as a potential baseline measure for workload assessment.

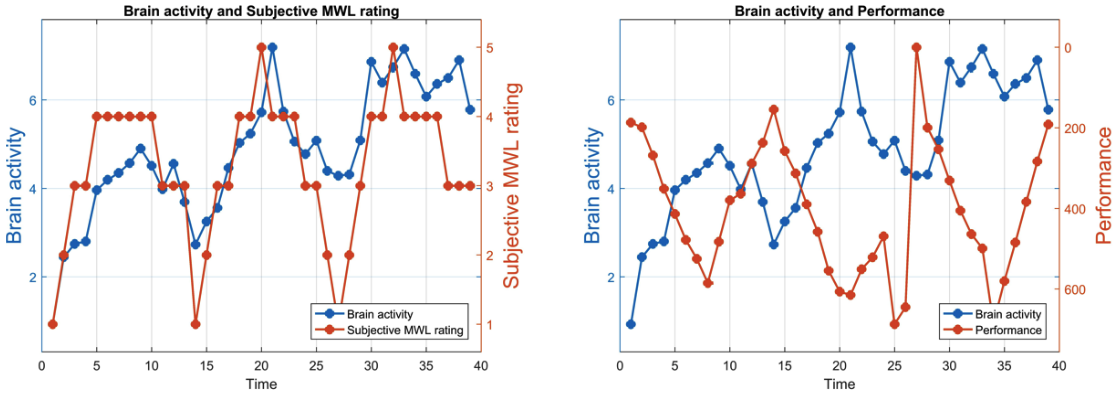


Figure 7.2 Relationship between fNIRS OXY measure, ISA, and Performance.

We then explored the relationship between workload, demand and performance using the other physiological techniques.

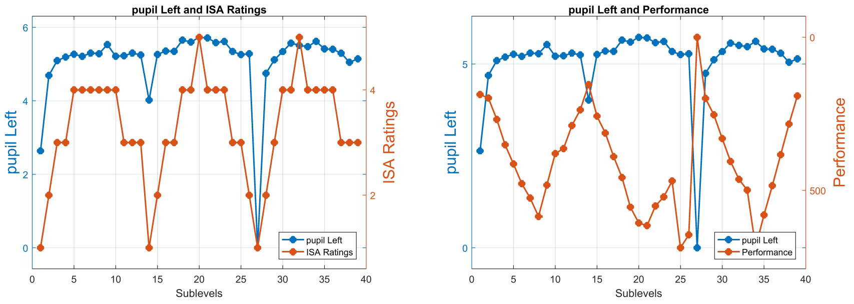


Figure 7.3 Pupil Left VS ISA scores

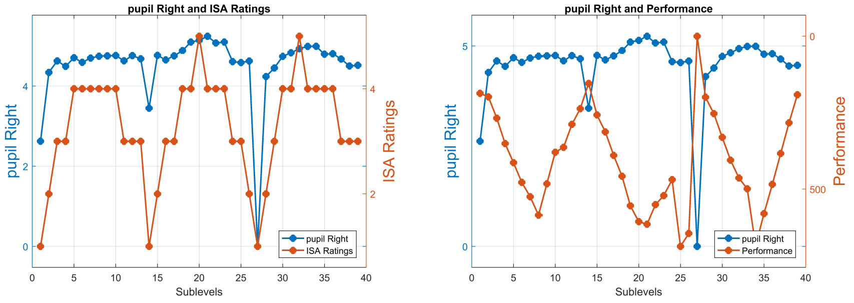


Figure 7.4 Pupil Right VS ISA scores VS Performance

Figure 7.3 and Figure 7.4 shows the close relationship between the continuous subjective measure of workload ISA and the physiological response of pupil size for both the left and right eyes. The results showed a Spearman's correlation between ISA and both, Left and Right pupil $r = .659, p < .05$. There was however a stronger correlation between fNIRS's OXY measure and the left and right pupil response to workload $r = .868, p < .001$ and $r = .854, p < .001$ respectively.

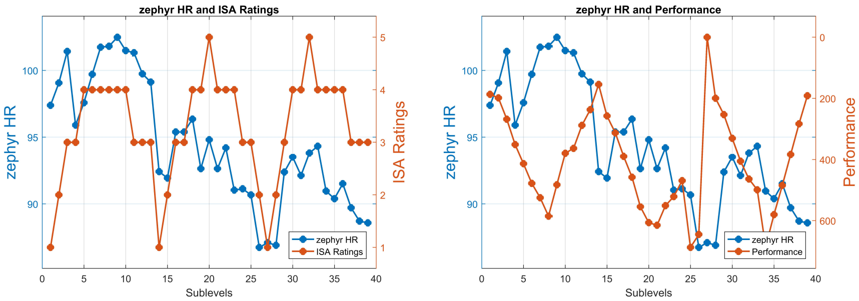


Figure 7.5 Heart Rate VS ISA scores VS Performance

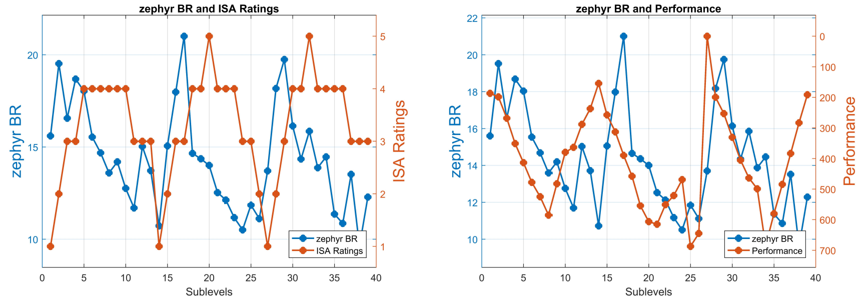


Figure 7.6 Breathing Rate VS ISA scores VS Performance

Similar patterns could be observed in the heart rate HR signal Figure 7.5 and breathing rate BR from the Zephyr chest band. However the correlations between HR, BR, and ISA were not statistically significant.

Another aim was to evaluate the quality and limitations of using the E4 wrist band in comparison with the dedicated chest band for measuring the heart activity. There was a Pearson correlation between the two sensors $r = .614, p = .026$, and generally there was a similar pattern between the two (Figure 7.7), however, a stronger correlation was expected.

There was a close relationship between the E4 heart rate, EDA, and skin temperature and fNIRS signal. There was a Pearson correlation between E4 HR and fNIRS Oxy measures $r = .838, p < 0.001$, E4 EDA and fNIRS Oxy measures $r = .7, p < 0.01$, E4

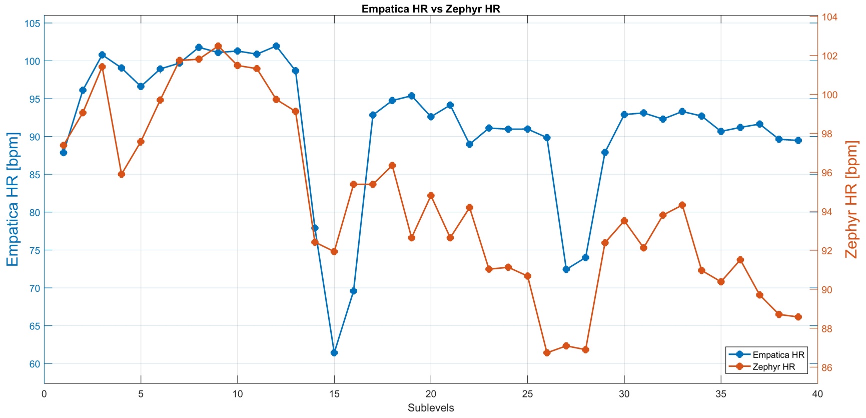


Figure 7.7 HR using E4 VS Zephyr

Skin Temperature and fNIRS Oxy measures $r = .56, p < 0.05$. There was no direct correlation between the E4 measures and ISA.

A similar discussion was presented by Nagasawa and Hagiwara [98].

7.5.2 Limitation of physiological techniques outside the lab settings

fNIRS was shown to provide a useful reference baseline for assessing mental workload and changes in demand during tasks. However, when comes to using fNIRS, one is limited to lab-based evaluation settings. Even-though wireless fNIRS alternative exists, it is not yet common to have a wearable techniques to measure mental brain activity throughout every day task. The aim of this pilot study was to investigate the limitations of other physiological techniques that tend to be more suitable when comes to studying mental workload outside of the lab.

Even-though there was no direct relationship the subjective ISA measure of workload and the other physiological sensors, there was a close relationship between the physiological data and fNIRS, indicating the same patterns. It is therefore to suggest that there is an indication of workload in the data from the E4 band, and there is scope to use fNIRS to better understand how to combine physiological data in order to assess workload of everyday life.

7.6 Considerations and Guidance for researchers new to fNIRS workload experiments

Depending on the interests and settings of the research, fNIRS technique could be useful in different ways for Human Computer Interaction and Human Factors research. In this section we present a practical guidance to future researchers that have not used fNIRS before, who want to adopt this technique for their research.

The fNIR technology is designed to allow you to track relative changes in oxygen consumption as well as changes in blood volume in various parts of the brain, however, for workload experiments the aim is the prefrontal cortex, the area behind the forehead for a typical healthy adult.

7.6.1 Operating procedure

To operate a typical fNIR System you will need:

- an fNIRS probe (the actual sensors you place on subjects) that could be wired or wireless,
- an fNIRS control box, connected to the probe via wired/wireless technology,
- and a laptop or personal computer.

The probe is collecting the actual physiological signals using the IR sensors, the signal is further sent for pre-processing to the control box. The control box is further connected for signal acquisition and processing to the personal computer.

After obtaining informed consent, participants are ready to be connected to the fNIRS probes. Applying the sensor probes correctly is the most important step in getting good data. Each individual forehead has a different size and shape, however, the typical targeted place is between the subject's hair and eyebrows. To place the sensor on the forehead, the subject's hair must be pulled back using one hand (for long hair a hair band could be used), and the other is used to place the sensor. It is important that all sensors are in contact directly with the skin. Finally, the sensors should ideally be covered using a bandage or a headband, to keep out the extraneous light (e.g. sunlight).

A baseline step is required with each individual subject, in which usually participants are asked to stay quiet and rest, trying not to perform any mental activities. Beer–Lambert Law, is at the heart of the fNIRS technology for calculating the oxygen concentrations in various regions in the brain, and these are indicative of brain activity. Using the properties of the law, which could be further explored in [136], it is possible to calculate the OXY-haemoglobin and de-OXY hemoglobin levels in the target medium in relation to the levels at baseline, where participants are asked to rest. A few other measurements including total oxygenation could be obtained by further combining the two.

7.6.2 Stimuli and study tasks

Using various stimuli of interest, including interactions with technology (e.g. testing various interfaces), typically presented using a different laptop or personal computer, participants' physiological changes in the brain could be monitored continuously using fNIRS. Time markers are used to track various points during interaction, or to mark the beginning and the end of each study condition.

As fNIRS measurements are not absolute values, the comparison between participants is not ideal, and the within participants design is more appropriate. This means that the oxygenation measurements from the fNIRS device are relative to the individual, and not to the rest of the participants.

A good starting point could be the use of fNIRS for a typical block design analysis, for say having a study comparing three variations of a user interface. fNIRS could be useful to reflect which of the three variations of the interface generates a higher average workload during interaction, however it could be also used to detect period of high workload during the conditions.

fNIRS measurements are complimentary to existing workload techniques, and measures of performance subjective techniques and other physiological methods could be used to better understand and interpret results. As fNIRS is a continuous technique, the use of continuous subjective measures such as ISA rather than one-off measures such as NASA-TLX, would be better off in contrast to fNIRS.

The design of the experiment tasks should allow a good monitoring of task performance, but also a understanding of the demands placed upon the subjects during the experiments. Combined with subjective techniques, one could study the relationship between task performance, perceived workload, and physiological (more objective workload) using fNIRS, but also understand where during interaction, high and low periods of workload have an impact on task performance and perceived workload.

7.6.3 Signal processing

A typical fNIRS experiment involves various levels of pre-processing and processing, and depending on the technology and software you get with it (manufacturing company and software package), you have some of the processing done automatically. For example, fNIRS picks up artefacts related to respiration and heart beats [33]; some fNIRS technologies clear these automatically, however some of the challenges of filtering techniques are mentioned by [127] and explored in [89].

For the studies presented in this thesis, measurements of brain activity were recorded using an fNIRS300 device and the associated Cognitive Optical Brain Imaging (COBI) Studio hardware integrated software platform provided by Biopac Systems Inc. The processing and filtering was performed using two software packages:

- the Matlab Toolbox NIRS-SPM [150],
- and fnirSoft, the Comprehensive Signal Processing, Analysis and Visualization Platform for Optical Brain Imaging [8].

A low pass filter with cut off frequencies of 0.2 Hz can be used in order to remove high-frequency noise, physiological artefacts such as heartbeats and motion derived artefacts. In relation to workload feature extractions, the Correlation Based Signal Improvement (CBSI) method can be applied [35], a technique designed for fNIRS technology in order to improve detection of workload (based on the expected negative correlation between changes in Oxy and De-Oxy hemoglobin).

7.6.4 Things to watch out for

fNIRS is particularly sensitive to skin colour, and it was found to be less reliable in particular for individuals with dark skin [139]. This effect was mainly observed in the medical domain, where patients identified as black had a significantly increased chance of not being able to have a NIRS reading [104], and the interpretation of results were interpreted with caution. From my experience with the technology, this effect is caused by the change in the level of IR light absorption by the skin itself, darker skin absorbing more of the IR light and therefore not reaching the right depth in the brain. Some of the fNIRS technologies today allow further control over the light intensities, further allowing various intensities for different skin colours.

Small foreheads tend to be another problem with fNIRS experiments, mostly when using a pre-defined sensor layout, which is designed for a typical/average adult. In most of these cases some of the sensors are either reaching the hair, or sitting on the eyebrows, in both cases we tend to ignore the channels of data coming from the sensors not sitting right on the skin.

The delay associated with the hemodynamic response [136] can be taken into account using various techniques including: averages across blocks of data, omitting the first few seconds of the trials when processing, or simply delaying the trial data by a few seconds [105, 108]. Depending on the experiments, we use different techniques, however for the continuous use and cause and effect analysis, we tend to delay the whole blocks of trials before we triangulate the data with events.

Another common problem shared between most of the BCI experiments is participants sleepiness during the experiments. Extensive periods of time in the lab, wearing brain based sensors tend to make participants feel sleepy, we found this in a few experiments, in one case in particular the participant fell asleep during one of the study conditions. We therefore advise on having a targeted 1 to 1 and a half hour per experiment, bur a no longer than 2 hours of continuous data collection is advisable.

7.7 Thesis Strengths and Limitations. Future Work.

This thesis explored the use of fNIRS, as an emerging technique for studying the interaction between people and technology. The first part of the thesis investigated the feasibility of using fNIRS in typical lab-based evaluation settings, we then looked into its properties in relation to workload (sensitivity, reliability, validity), and the thesis finished by exploring its use for the continuous evaluation and feedback of workload during tasks.

We found the technique relatively suitable and complementary for typical evaluation settings, providing useful information about the user during interaction without further restricting users from their interaction with a computer based system. However, there are major drawbacks for using such technology in the wild, for the continuous monitoring of the workload of everyday task. This is one future direction of this research, and therefore, looking into other, more portable techniques, including wireless fNIRS techniques and other physiological sensors is essential. Heart, skin, eye activity, facial thermography, have the potential to reflect a similar understanding of workload, however, with the benefit of being less invasive and more portable. The specific instance of future research presented above was exploring how these less invasive and more portable physiological techniques could be used to measure workload; the first step into monitoring the workload of everyday task.

If the above research will be possible, the feedback of workload of everyday task could be explored. This future research would explore how people could reflect when given feedback of their workload throughout the day. Similar to physical activity trackers and based on a combination of physiological data, the feedback will reflect the amount of mental activity and workload a person is going through during daily activities.

Another future direction could be exploring feedback of workload in the context of critical settings such as Air Traffic Control, Train Driver, but also in in-car-settings. A further understanding of what feedback of workload means, and how it can be used, and studying the way and modality people prefer to receive the feedback could also be explored.

One other important aspect of future work is moving towards machine learning algorithms to automatically detect users' state. It was not in the scope of this thesis to explore this area of research, however, this is an important step for future work, where the continuous assessment of workload outside laboratory settings - using a data triangulation between multiple physiological data streams - is required.

If workload is a concept discussed particularly when comes to tasks and jobs with high demands where the operators are under a high mental workload, with a similar importance but at the opposite end of the spectrum is the issue of boredom. Certain jobs and tasks that place a low demand on operators, and/or jobs that are repetitive over long periods of time (e.g. train drivers) may lead to operator simply not having enough work to do causing underload or boredom. During this state, the operators are again prone to mistakes. In the previous chapter we discussed the importance of providing both feedback of low and high workload. One particular path of future work could explore in detail the use of physiological techniques to particularly assess boredom.

The works presented in this thesis established a baseline of knowledge in the field of physiological measurement and feedback of workload. It requires further evaluation to understand its true feasibility and contribution in a load of contexts, however, this work developed an extensive understanding of how physiological techniques - such as fNIRS - can be used and evaluated in terms of mental workload measurement in the context of Human Factors and Human-Computer Interaction research.

Bibliography

[1] Afergan, D., Peck, E. M., Solovey, E. T., Jenkins, A., Hincks, S. W., Brown, E. T., Chang, R., and Jacob, R. J. (2014). Dynamic difficulty using brain metrics of workload. In *Proc. SIGCHI*, pages 3797–3806. ACM.

[2] Alloway, T. P., Kerr, I., and Langheinrich, T. (2010). The effect of articulatory suppression and manual tapping on serial recall. *European Journal of Cognitive Psychology*, 22(2):297–305.

[3] Alm, H. and Nilsson, L. (1995). The effects of a mobile telephone task on driver behaviour in a car following situation. *Accident Analysis & Prevention*, 27(5):707–715.

[4] Aron, A. R., Robbins, T. W., and Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in cognitive sciences*, 8(4):170–177.

[5] Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. *The psychology of learning and motivation*, 2:89–195.

[6] Ayaz, H. and Onaral, B. (2005). *Analytical software and stimulus-presentation platform to utilize, visualize and analyze near-infrared spectroscopy measures*. Drexel University.

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

[8] Ayaz, H., Shewokis, P. A., Bunce, S. C., and Onaral, B. (2012b). Functional near infrared spectrocopy based brain computer interface. US Patent App. 14/007,203.

[9] Baddeley, A. (1992). Working memory. *Science*, 255(5044):556.

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

[11] Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature reviews neuroscience*, 4(10):829–839.

[12] Baddeley, A. D. (2002). Is working memory still working? *European psychologist*, 7(2):85–97.

[13] Baddeley, A. D. and Hitch, G. (1974). Working memory. *The psychology of learning and motivation*, 8:47–89.

[14] Battiste, V. and Bortolussi, M. (1988). Transport pilot workload: A comparison of two subjective techniques. In *Proceedings of the Human Factors Society Annual Meeting*, volume 32, pages 150–154. SAGE Publications Sage CA: Los Angeles, CA.

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

[16] Becker, A. B., Warm, J. S., Dember, W. N., and Hancock, P. A. (1991). Effects of feedback on perceived workload in vigilance performance. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 35, pages 1491–1494. SAGE Publications.

[17] Bengtsson, J. (1995). What is reflection? on reflection in the teaching profession and teacher education. *Teachers and Teaching: theory and practice*, 1(1):23–32.

[18] Billman, G. E. (2011). Heart rate variability–a historical perspective. *Frontiers in physiology*, 2.

[19] Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., and Babiloni, F. (2014). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75.

[20] Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., and Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage*, 5(1):49–62.

[21] Brennan, S. (1992). An experimental report on rating scale descriptor sets for the instantaneous self assessment (isa) recorder. *Portsmouth: DRA Maritime Command and Control Division. DRA Technical Memorandum (CAD5)*, 92017.

[22] Broadbent, D. E. (1958). *Perception and communication*. New York: Pergamon Press;.

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

[24] Bunce, S. C., Izzetoglu, K., Ayaz, H., Shewokis, P., Izzetoglu, M., Pourrezaei, K., and Onaral, B. (2011). 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.

[25] Bunce, S. C., Izzetoglu, M., Izzetoglu, K., Onaral, B., and Pourrezaei, K. (2006). Functional near-infrared spectroscopy. *IEEE engineering in medicine and biology magazine*, 25(4):54–62.

[26] Cain, B. (2007). A review of the mental workload literature. Technical report, DTIC Document.

[27] Casali, J. G. and Wierwille, W. W. (1984). On the measurement of pilot perceptual workload: a comparison of assessment techniques addressing sensitivity and intrusion issues. *Ergonomics*, 27(10):1033–1050.

[28] Chance, B., Anday, E., Nioka, S., Zhou, S., Hong, L., Worden, K., Li, C., Murray, T., Ovetsky, Y., Pidikiti, D., et al. (1998). A novel method for fast imaging of brain function, non-invasively, with light. *Optics express*, 2(10):411–423.

[29] Chance, B., Zhuang, Z., UnAh, C., Alter, C., and Lipton, L. (1993). Cognition-activated low-frequency modulation of light absorption in human brain. *Proceedings of the National Academy of Sciences*, 90(8):3770–3774.

[30] Colibazzi, T., Posner, J., Wang, Z., Gorman, D., Gerber, A., Yu, S., Zhu, H., Kangarlu, A., Duan, Y., Russell, J. A., et al. (2010). Neural systems subserving valence and arousal during the experience of induced emotions. *Emotion*, 10(3):377.

[31] Collet, C., Salvia, E., and Petit-Boulanger, C. (2014). Measuring workload with electrodermal activity during common braking actions. *Ergonomics*, 57(6):886–896.

[32] Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338.

[33] Coyle, S., Ward, T., and Markham, C. (2004). Physiological noise in near-infrared spectroscopy: implications for optical brain computer interfacing. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 2, pages 4540–4543. IEEE.

[34] Coyle, S. M., Ward, T. E., and Markham, C. M. (2007). Brain–computer interface using a simplified functional near-infrared spectroscopy system. *Journal of neural engineering*, 4(3):219.

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

[36] D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S., and Grossman, M. (1995). The neural basis of the central executive system of working memory. *Nature*, 378(6554):279–281.

[37] D'Esposito, M., Zarahn, E., and Aguirre, G. K. (1999). Event-related functional mri: implications for cognitive psychology. *Psychological bulletin*, 125(1):155.

[38] Durantin, G., Gagnon, J.-F., Tremblay, S., and Dehais, F. (2014). Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural brain research*, 259:16–23.

[39] Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., and Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly*, 6(1):1–20.

[40] Duvinage, M., Castermans, T., Dutoit, T., Petieau, M., Hoellinger, T., Saedeleer, C., Seetharaman, K., and Cheron, G. (2012). A P300-based quantitative comparison between the Emotiv Epoc headset and a medical EEG device. In *IASTED*.

[41] Eady, M. J. and Lockyer, L. (2013). What hemodynamic (fnirs), electrophysiological (eeg) and autonomic integrated measures can tell us about emotional processing. *Learning to Teach in the Primary School*, page pp. 71.

[42] Eggemeier, F. T., Wilson, G. F., Kramer, A. F., and Damos, D. L. (1991). Workload assessment in multi-task environments. *Multiple-task performance*, pages 207–216.

[43] Fadiga, L., Craighero, L., and D'Ausilio, A. (2009). Broca's area in language, action, and music. *Annals of the New York Academy of Sciences*, 1169(1):448–458.

[44] Fletcher, K. (2015). The dynamic effects of task demands on resource availability, resource allocation and metacognitive states.

[45] Gabrieli, J. D., Poldrack, R. A., and Desmond, J. E. (1998). The role of left prefrontal cortex in language and memory. *Proceedings of the national Academy of Sciences*, 95(3):906–913.

[46] Garbarino, M., Lai, M., Bender, D., Picard, R. W., and Tognetti, S. (2014). Empatica e3—a wearable wireless multi-sensor device for real-time computerized biofeedback and data acquisition. In *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*, pages 39–42. IEEE.

[47] Geddie, J. C., Boer, L., Edwards, R., Enderwick, T., and Graff, N. (2001). Nato guidelines on human engineering testing and evaluation. Technical report, DTIC Document.

[48] Girouard, A., Solovey, E. T., Hirshfield, L. M., Peck, E. M., Chauncey, K., Sassaroli, A., Fantini, S., and Jacob, R. J. (2010). From brain signals to adaptive interfaces: using fnirs in hci. In *Brain-Computer Interfaces*, pages 221–237. Springer.

[49] Gopher, D. and Donchin, E. (1986). Workload: An examination of the concept.

[50] Guastello, S. J., Shircel, A., Malon, M., and Timm, P. (2015). Individual differences in the experience of cognitive workload. *Theoretical Issues in Ergonomics Science*, 16(1):20–52.

[51] Haapalainen, E., Kim, S., Forlizzi, J. F., and Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 301–310. ACM.

[52] Hancock, P. A. and Meshkati, N. E. (1988). *Human mental workload.* North-Holland.

[53] Harbluk, J. L., Noy, Y. I., Trbovich, P. L., and Eizenman, M. (2007). An on-road assessment of cognitive distraction: Impacts on drivers' visual behavior and braking performance. *Accident Analysis & Prevention*, 39(2):372–379.

[54] Hart, S. G. (2006). 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 Sage CA: Los Angeles, CA.

[55] Hart, S. G. and Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload.*

[56] Hart, S. G. and Wickens, C. D. (1990). Workload assessment and prediction. In *Manprint*, pages 257–296. Springer.

[57] Heilman, R. M., Crişan, L. G., Houser, D., Miclea, M., and Miu, A. C. (2010). Emotion regulation and decision making under risk and uncertainty. *Emotion*, 10(2):257.

[58] Heine, T., Lenis, G., Reichensperger, P., Beran, T., Doessel, O., and Deml, B. (2017). Electrocardiographic features for the measurement of drivers' mental workload. *Applied Ergonomics*, 61:31–43.

[59] Hernandez, J., McDuff, D., and Picard, R. W. (2015). Biowatch: estimation of heart and breathing rates from wrist motions. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*, pages 169–176. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

[60] Hirshfield, L. M., Solovey, E. T., Girouard, A., Kebinger, J., Jacob, R. J., Sassaroli, A., and Fantini, S. (2009). Brain measurement for usability testing and adaptive interfaces: an example of uncovering syntactic workload with functional near infrared spectroscopy. In *Proc. CHI*, pages 2185–2194. ACM.

[61] Hirth, C., Obrig, H., Villringer, K., Thiel, A., Bernarding, J., Mühlnickel, W., Flor, H., Dirnagl, U., and Villringer, A. (1996). Non-invasive functional mapping of the human motor cortex using near-infrared spectroscopy. *Neuroreport*, 7(12):1977–1981.

[62] Hollands, J. G. and Wickens, C. D. (1999). Engineering psychology and human performance. *Journal of surgical oncology*.

[63] Huang, F.-H., Hwang, S.-L., Yenn, T.-C., Yu, Y.-C., Hsu, C.-C., and Huang, H.-W. (2006). Evaluation and comparison of alarm reset modes in advanced control room of nuclear power plants. *Safety science*, 44(10):935–946.

[64] Huey, B. M., Wickens, C. D., et al. (1993). *Workload transition: Implications for individual and team performance*. National Academies Press.

[65] Huppert, T., Hoge, R., Diamond, S., Franceschini, M. A., and Boas, D. A. (2006). A temporal comparison of bold, asl, and nirs hemodynamic responses to motor stimuli in adult humans. *Neuroimage*, 29(2):368–382.

[66] Hwang, S.-L., Yau, Y.-J., Lin, Y.-T., Chen, J.-H., Huang, T.-H., Yenn, T.-C., and Hsu, C.-C. (2008). Predicting work performance in nuclear power plants. *Safety science*, 46(7):1115–1124.

[67] Iqbal, S. T., Zheng, X. S., and Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. In *CHI'04 extended abstracts on Human factors in computing systems*, pages 1477–1480. ACM.

[68] Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., and Chance, B. (2004). Functional optical brain imaging using near-infrared during cognitive tasks. *IJHCI*, 17(2):211–227.

[69] Izzetoglu, M., Bunce, S. C., Izzetoglu, K., Onaral, B., and Pourrezaei, K. (2007). Functional brain imaging using near-infrared technology. *IEEE Engineering in Medicine and Biology Magazine*, 26(4):38.

[70] Jobsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198(4323):1264–1267.

[71] Jones, D., Farrand, P., Stuart, G., and Morris, N. (1995). Functional equivalence of verbal and spatial information in serial short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(4):1008.

[72] Jordan and Brennen (1992). Instantaneous self-assessment of workload technique (isa).

[73] Jordan, C. (1992). Experimental study of the effects of an instantaneous self assessment workload recorder on task performance. *Report No. DRA/TM (CAD5)/92011. Farnborough: Defence Evaluation & Research Agency.*

[74] Kahneman, D. (1973). *Attention and effort*, volume 1063. Prentice-Hall Englewood Cliffs, NJ.

[75] Kane, M. J. and Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic bulletin & review*, 9(4):637–671.

[76] Klingner, J., Kumar, R., and Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 symposium on Eye tracking research & applications*, pages 69–72. ACM.

[77] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

[78] Larsen, J. D. and Baddeley, A. (2003). Disruption of verbal stm by irrelevant speech, articulatory suppression, and manual tapping: Do they have a common source? *Quarterly Journal of Experimental Psychology Section A*, 56(8):1249–1268.

[79] Leanne, M. and Robert, J. (2009). Using brain measurement to evaluate reality based interactions. *Challenges in the Evaluation of Usability and User Experience in Reality Based Interaction*, 5:19–20.

[80] Lee, Y.-H. and Liu, B.-S. (2003). Inflight workload assessment: Comparison of subjective and physiological measurements. *Aviation, space, and environmental medicine*, 74(10):1078–1084.

[81] Leggatt, A. (2005). Validation of the isa (instantaneous self assessment) subjective workload tool. In *Contemporary Ergonomics 2005: Proceedings of the International Conference on Contemporary Ergonomics (CE2005), 5-7 April 2005, Hatfield, UK*, page 74. CRC Press.

[82] Li, X., Xu, L., Yao, L., and Zhao, X. (2013). A novel HCI system based on real-time fmri using motor imagery interaction. In *Foundations of Augmented Cognition*, pages 703–708. Springer.

[83] Logie, R. H., Gilhooly, K. J., and Wynn, V. (1994). Counting on working memory in arithmetic problem solving. *Memory & cognition*, 22(4):395–410.

[84] Lohse, M., Rothuis, R., Gallego-Pérez, J., Karreman, D. E., and Evers, V. (2014). Robot gestures make difficult tasks easier: the impact of gestures on perceived workload and task performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1459–1466. ACM.

[85] Lukanov, K., Maior, H. A., and Wilson, M. L. (2016). Using fnirs in usability testing: understanding the effect of web form layout on mental workload. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4011–4016. ACM.

[86] Macken, W. J. and Jones, D. M. (1995). Functional characteristics of the inner voice and the inner ear: Single or double agency? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(2):436.

[87] Maior, H. A., Pike, M., Sharples, S., and Wilson, M. L. (2015). Examining the reliability of using fnirs in realistic hci settings for spatial and verbal tasks. In *Proceedings of CHI*, volume 15, pages 3807–3816.

[88] Mandrick, K., Peysakhovich, V., Rémy, F., Lepron, E., and Causse, M. (2016). Neural and psychophysiological correlates of human performance under stress and high mental workload. *Biological psychology*, 121:62–73.

[89] Matthews, F., Pearlmutter, B. A., Wards, T. E., Soraghan, C., and Markham, C. (2008). Hemodynamics for brain-computer interfaces. *IEEE Signal Processing Magazine*, 25(1):87–94.

[90] Megaw, T. (2005). The definition and measurement of mental workload. *Evaluation of human work, Eds. Esmond N. Corlett, and John R. Wilson*, pages 525–551.

[91] Merat, N., Jamson, A. H., Lai, F. C., and Carsten, O. (2012). Highly automated driving, secondary task performance, and driver state. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54(5):762–771.

[92] Meshkati N, Hancock P, and Rahimi M (1992). Techniques in mental workload assessment. In Wilson John R and Corlett Nigel E, editors, *Evaluation of human work*. Taylor & Francis London.

[93] Metzger, U. and Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors*, 47(1):35–49.

[94] Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202.

[95] Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The psychological review*, 63:81–97.

[96] Mu, Z., Hu, J., and Yin, J. (2016). Driving fatigue detecting based on eeg signals of forehead area. *International Journal of Pattern Recognition and Artificial Intelligence*, page 1750011.

[97] Mueller, S. (2012). Pebl: The psychology experiment building language (version 0.10).[computer experiment programming language]. *Retrieved Nov*.

[98] Nagasawa, T. and Hagiwara, H. (2016). Workload induces changes in hemodynamics, respiratory rate and heart rate variability. In *Bioinformatics and Bioengineering (BIBE), 2016 IEEE 16th International Conference on*, pages 176–181. IEEE.

[99] Naito, M., Michioka, Y., Ozawa, K., Yoshitoshi, I., Kiguchi, M., and Kanazawa, T. (2007). A communication means for totally locked-in als patients based on changes in cerebral blood volume measured with near-infrared light. *IEICE transactions on information and systems*, (7):1028–1037.

[100] Nickel, P. and Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1-hz component of heart rate variability as an indicator of mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(4):575–590.

[101] O'Donnell, R. D. and Eggemeier, F. T. (1986). Workload assessment methodology.

[102] Or, C. K. and Duffy, V. G. (2007). Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. *Occupational Ergonomics*, 7(2):83–94.

[103] Oron-Gilad, T., Szalma, J. L., Stafford, S. C., and Hancock, P. A. (2008). The workload and performance relationship in the real world: A study of police officers in a field shooting exercise. *International Journal of Occupational Safety and Ergonomics*, 14(2):119–131.

[104] Ostrov, L., Grimsby, G., Menon, V., Keays, M., Sheth, K., Granberg, C., DaJusta, D., Hill, M., Sanchez, E., Huang, R., et al. (2015). Mp40-12 the effect of race and skin color on near-infrared spectroscopy readings in pediatric patients with unilateral acute scrotum. *The Journal of Urology*, 193(4):e467.

[105] Peck, E. M., Yuksel, B. F., Ottley, A., Jacob, R. J., and Chang, R. (2013). Using fNIRS Brain Sensing to Evaluate Information Visualization Interfaces. In *Proc. SIGCHI*. ACM.

[106] Pickup, L., Wilson, J. R., Norris, B. J., Mitchell, L., and Morrisroe, G. (2005). The integrated workload scale (iws): a new self-report tool to assess railway signaller workload. *Applied Ergonomics*, 36(6):681–693.

[107] Pike, M., Wilson, M. L., Divoli, A., and Medelyan, A. (2012). CUES: Cognitive Usability Evaluation System. In *EuroHCIR2012*, pages 51–54.

[108] Pike, M. F., Maior, H. A., Porcheron, M., Sharples, S. C., and Wilson, M. L. (2014). Measuring the effect of think aloud protocols on workload using fnirs. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 3807–3816. ACM.

[109] Plichta, M. M., Gerdes, A. B., Alpers, G., Harnisch, W., Brill, S., Wieser, M., and Fallgatter, A. J. (2011). Auditory cortex activation is modulated by emotion: a functional near-infrared spectroscopy (fnirs) study. *Neuroimage*, 55(3):1200–1207.

[110] Rasche, P., Mertens, A., Schlick, C., and Choe, P. (2015). The effect of tactile feedback on mental workload during the interaction with a smartphone. In *Cross-Cultural Design Methods, Practice and Impact*, pages 198–208. Springer.

[111] Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE transactions on systems, man, and cybernetics*, (3):257–266.

[112] Reid, G. B. and Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in psychology*, 52:185–218.

[113] Reid, G. B., Potter, S. S., and Bressler, J. (1989). Subjective workload assessment technique (swat): A user's guide. *Wright Patterson Air Force Base, OH: Harry G. Armstrong Aerospace Medical Research Laboratory*.

[114] Roberts, A. H. (1985). Biofeedback: Research, training, and clinical roles. *American Psychologist*, 40(8):938.

[115] Rubio, S., Díaz, E., Martín, J., and Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, 53(1):61–86.

[116] Rushworth, M. F., Nixon, P. D., Eacott, M. J., and Passingham, R. E. (1997). Ventral prefrontal cortex is not essential for working memory. *The Journal of Neuroscience*, 17(12):4829–4838.

[117] Saito, S. (1993). The disappearance of phonological similarity effect by complex rhythmic tapping. *Psychologia: An International Journal of Psychology in the Orient*.

[118] Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*, volume 5126. Basic books.

[119] Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press.

[120] Sharples, S., Edwards, T., and Balfe, N. (2012). Inferring cognitive state from observed interaction. In *Proceedings of the 4th AHFE International Conference*.

[121] Sharples, S. and Megaw, T. (2015). Definition and mesurement of human workload. In Wilson, J. R. and Sharples, S., editors, *Evaluation of human work*. CRC Press.

[122] Shi, Y., Ruiz, N., Taib, R., Choi, E., and Chen, F. (2007). Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 extended abstracts on Human factors in computing systems*, pages 2651–2656. ACM.

[123] Shimomura, Y., Yoda, T., Sugiura, K., Horiguchi, A., Iwanaga, K., and Katsuura, T. (2008). Use of frequency domain analysis of skin conductance for evaluation of mental workload. *Journal of physiological anthropology*, 27(4):173–177.

[124] Shneiderman, B. and Plaisant, C. (2005). Designing the user interface: Strategies for effective human computer interaction. *Addison-Wesley, Reading, Mass*, (4th edn).

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

[126] Solovey, E., Afergan, D., Peck, E. M., Hincks, S. W., and Jacob, R. J. (2015). Designing implicit interfaces for physiological computing: guidelines and lessons learned using fnirs. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(6):35.

[127] Solovey, E. T., Girouard, A., Chauncey, K., Hirshfield, L. M., Sassaroli, A., Zheng, F., Fantini, S., and Jacob, R. J. (2009). Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines. In *Proc. UIST*, pages 157–166. ACM.

[128] Stemberger, J., Allison, R. S., and Schnell, T. (2010). Thermal imaging as a way to classify cognitive workload. In *Computer and Robot Vision (CRV), 2010 Canadian Conference On*, pages 231–238. IEEE.

[129] Stuiver, A., De Waard, D., Brookhuis, K., Dijksterhuis, C., Lewis-Evans, B., and Mulder, L. (2012). Short-term cardiovascular responses to changing task demands. *International Journal of Psychophysiology*, 85(2):153–160.

[130] Svensson, E., Angelborg-Thanderez, M., Sjöberg, L., and Olsson, S. (1997). Information complexity-mental workload and performance in combat aircraft. *Ergonomics*, 40(3):362–380.

[131] Tan, D. and Nijholt, A. (2010). Brain-computer interfaces and human-computer interaction. In *Brain-Computer Interfaces*, pages 3–19. Springer.

[132] Taoda, K., Kawamura, M., Wakara, K., Fukuchi, Y., and Nishiyama, K. (2001). Heart rate variability during long truck driving work. *Journal of human ergology*, 30(1-2):235–240.

[133] Tattersall, A. J. and Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39(5):740–748.

[134] Thurlings, M. E., van Erp, J. B., Brouwer, A.-M., and Werkhoven, P. J. (2010). Eeg-based navigation from a human factors perspective. In *Brain-Computer Interfaces*, pages 71–86. Springer.

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

[136] Villringer, A. and Chance, B. (1997). Non-invasive optical spectroscopy and imaging of human brain function. *Trends in neurosciences*, 20(10):435–442.

[137] Villringer, A., Planck, J., Hock, C., Schleinkofer, L., and Dirnagl, U. (1993). Near infrared spectroscopy (NIRS): a new tool to study hemodynamic changes during activation of brain function in human adults. *Neuroscience letters*, 154(1):101–104.

[138] Warm, J. S., Dember, W. N., and Hancock, P. A. (1996). Vigilance and workload in automated systems.

[139] Wassenaar, E. and Van den Brand, J. (2005). Reliability of near-infrared spectroscopy in people with dark skin pigmentation. *Journal of clinical monitoring and computing*, 19(3):195–199.

[140] Welford, A. T. (1968). Fundamentals of skill.

[141] Whiting, H. T. A. (1969). *Acquiring ball skill: A psychological interpretation*. Lea & Febiger.

[142] Wickens, C. D. (1991). Processing resources and attention. *Multiple-task performance*, 1991:3–34.

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

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

[145] Wickens, C. D., Gordon, S. E., and Liu, Y. (2004). *An introduction to human factors engineering*. Pearson Prentice Hall Upper Saddle River.

[146] Wickens, C. D., Hollands, J. G., Banbury, S., and Parasuraman, R. (2015). *Engineering psychology & human performance*. Psychology Press.

[147] Wierwille, W. W. and Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 35(2):263–281.

[148] Wilson, G. F. and Russell, C. A. (2003). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(3):381–389.

[149] Wolpaw, J. R., McFarland, D. J., Vaughan, T. M., and Schalk, G. (2003). The wadsworth center brain-computer interface (bci) research and development program. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):1–4.

[150] Ye, J. C., Tak, S., Jang, K. E., Jung, J., and Jang, J. (2009). NIRS-SPM: statistical parametric mapping for near-infrared spectroscopy. *Neuroimage*, 44(2):428–447.

[151] Yeh, Y.-Y. and Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 30(1):111–120.

[152] Yuksel, B. F., Afergan, D., Peck, E. M., Griffin, G., Harrison, L., Chen, N. W., Chang, R., and Jacob, R. J. (2015). Braahms: A novel adaptive musical interface based on users' cognitive state. In *Proceedings of the International Conference on New Interfaces for Musical Expression NIME*.

[153] Yuksel, B. F., Oleson, K. B., Harrison, L., Peck, E. M., Afergan, D., Chang, R., and Jacob, R. J. (2016). Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5372–5384. ACM.