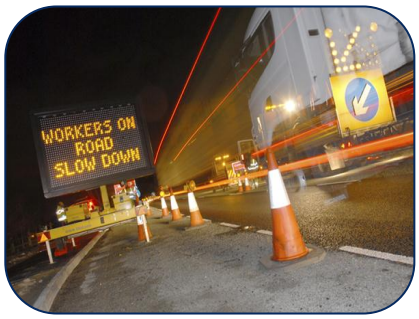
Transport Research Laboratory
Creating the future of transport



PUBLISHED PROJECT REPORT PPR749

Assessing cognitive underload during train driving: A physiological approach (CUPID)

S. Waters, A. Whitmore, D. Basacik, and N. Reed

Prepared for: RSSB

Project Ref:

Quality approved:

Sally Cotter
(Project Manager)

Prof Andrew Parkes
(Technical Referee)

2015

Disclaimer

Electronic copies of this report are available to RSSB members through the SPARK portal and on application to TRL Limited.

Hard copies can be purchased from TRL Limited. When purchased in hard copy, this publication is printed on paper that is FSC (Forest Stewardship Council) and TCF (Totally Chlorine Free) registered.

Contents amendment record

This report has been amended and issued as follows:

Version	Date	Description	Editor	Technical Referee

Contents

Executive summary

The rail industry has long recognised the effects of high workload on safety critical staff, but there has been very little research into very low workload. Cognitive underload is a state in which the demands of a task are so low that the performance of the person carrying out the task suffers. There are many scenarios in which train drivers monitor the status of the train in the train cab and environment outside the train, but either make few control inputs (eg running on green signals, running at low speeds for extended periods of time) or carry out very repetitive actions (eg running on caution signals, driving routes which have frequent station stops). In addition to characteristics of routes or traffic situations, there is an emerging concern that with increasing automation, the train driving task may not demand sufficient attention from the driver to keep them alert and engaged. This could result in safety-critical information being missed or not acted upon in the most appropriate manner.

This study built on a previous study by Robinson et al (2015) and aimed to understand whether it is possible to establish a set of reliable and objective physiological measures that could detect the presence of underload in a sample of train drivers. The study was designed following a literature review of physiological techniques used to measure low operator workload. An extended version of the Sustained Attention to Response Task (SART) was used as the vigilance task to be completed by participants. The repetitive nature of the SART has been identified in previous research as having similar characteristics to driving a train. Fifteen train drivers were recruited to take part in the study. Participants were required to complete the laptop based SART which lasted approximately 17 minutes. Physiological measures of cardiovascular, respiration and electrodermal activity were taken throughout the task. Performance measures such as reaction time, reaction time variability and accuracy were also taken based on performance on the SART. Subjective measures of boredom, fatigue and workload were collected after the task was completed.

Subjective measures of perceived boredom showed a significant increase in the second half of the task suggesting participants became more bored due to increased time on task. Even with this increase in boredom levels, the ratings during the second half of the SART were still closer to ‘not bored at all’ than ‘extremely bored’. No change was reported for sleepiness ratings, suggesting the study had successfully managed the potentially confounding effect of sleepiness.

From the results of the experiment it was evident that participants sometimes entered an automatic response mode when responding to the stimuli on the SART. This was seen through:

- Faster reaction times to the four stimuli immediately leading up to an incorrect response to a target stimulus and slower reaction times in the lead up to a correct response. Faster responses indicate more automatic information processing. These results have been reported in previous research by Robertson et al (1997).

- A lower mean heart rate in the four seconds leading up to an incorrect response to a target stimulus and a higher heart rate immediately before a correct response. Previous research has highlighted that a lower heart rate is an indicator of lower workload (Schmidt et al, 2009; Jap et al, 2009). However, the literature review has not identified any studies which have done this in the context of the SART, nor any studies which have found both faster reaction times and lowered heart rate prior to an input error.

This is a positive outcome of the study and supports the link between repetitive, automatic information processing and errors.

Given that the literature suggested that cognitive underload should lead to task performance deteriorating over time, it was surprising that this study found no such effect. Instead, participants' performance appeared to fluctuate, but was not significantly better or worse at the beginning, middle or end of the trials. There are several possible reasons for this:

- The task may have been so easy that everyone was able to complete it correctly even if their information processing deteriorated. This is very unlikely to have been the case as average response accuracy to one type of stimulus was only 67%. This suggests that the task was difficult enough.
- Alternatively, the task may have been too demanding. The SART was chosen based on claims made in the literature that was available before the study. However, very recent research by Dillard et al (2014), published following the literature review carried out in this project, has suggested that the SART may not actually promote 'mindlessness' or the withdrawal of attentional effort. Nevertheless, the fact that fluctuations in response speed and heart rate were linked to accuracy suggest that participants did not invest the same amount of effort and attention at all times. Further work could consider a vigilance task with fewer target stimuli over a longer period of time, to better ensure the task is not perceived as being demanding.
- The sample of train drivers may be more resilient to cognitive underload and could be better equipped to deal with repetitive conditions than the general population. Indeed, train drivers go through a rigorous selection process. This includes psychometric testing to identify candidates with skills such as maintaining high levels of concentration and performing well in low workload tasks. This factor, as a potential influence on the results, could be investigated by repeating the same study with a sample of non-train drivers and comparing the findings with the results of the current study.
- There may have been a practice effect which masked the expected performance decrements. Performance may have started at a lower level due to inexperience of the task but then improved as the participant adjusted to the speed and nature of the SART. At the same time, underload may have been causing the opposite effect, leading to no overall change in performance. A practice effect may also explain why drivers thought their performance improved during the

second half of the task. Participants were given a practice session of approximately 20 seconds before they started the trial, and this is standard in the research which has used the SART. However future studies could extend the length of this session.

- This was a relatively short task and participants may have been able to invest more effort as the trial progressed, in order to guard against the effects of boredom and repetition. This could have been because participants knew that they were taking part in an experiment and were being observed, and because they knew the task was relatively short: they were told that the task would only last about 15 minutes. The SART used in this study was four times longer than the standard SART, but this may not have been sufficient.

Future work in this area should, if possible, consider a vigilance task with fewer target stimuli, that monitors driver performance over a longer time period. Alternatively naturalistic studies could be conducted, in which the effects of being observed are minimised and measurement can take place over very long periods of time. Developments in technology such as camera based heart rate monitoring and the growth of accurate, non-invasive wearable technology for physiological monitoring may make it viable, in the future, to measure some of the key physiological indicators of workload on the operational railway.

The performance and heart rate effects that were observed in this study are promising. They suggest that physiological measures can indicate when someone is experiencing underload and is likely to make errors. At this stage the results are not strong enough to develop a tool to detect underload. Heart rate is influenced by a range of factors and, on its own, cannot be relied upon to make an assessment of workload. As technology develops to make it more viable to use some of the measures that this study was unable to use, such as EEG, research should be carried out to explore whether they can be used, potentially in combination with heart rate, to detect underload. An operational solution could then be used to indicate when to apply an appropriate mitigation. This could help to reduce errors associated with underload.

1 Introduction

TRL and RSSB are involved in joint research investigating the effects of very low mental workload (also known as cognitive underload) on train driver performance. This is a particularly important area of research because certain train driving scenarios involve periods of very low task demands, requiring the driver to continuously monitor and respond to repetitive stimuli in the environment, whilst at the same time maintaining a preparedness to react appropriately to infrequent critical events (Larue et al., 2010). Such conditions are associated with a reduction in arousal and attention, and reduced vigilance with increasing time on task.

Previous work by the team involved in this project looked at the effects of underload on train driving performance, and developed a method to assess the task characteristics which may indicate the potential for underload (Robinson et al, 2015), and which may be useful as part of a route risk assessment. However, it is not yet known whether drivers do indeed experience underload during tasks where the risk assessment method predicts that this is more likely. Thus, whether predicting the likelihood of underload along a route for risk assessment purposes, or as a means for determining the point at which to apply a strategy to mitigate underload, it is important to understand when drivers are, in fact, experiencing underload.

Although there are a number of studies that have attempted to detect low arousal using self-report and physiological measures, much of this work has been carried out in the context of fatigue. Many studies have also used physiological measures to look at workload, but the focus has been on high, rather than low, workload. In addition, there has been extensive work conducted in recent years to establish reliable methods and techniques for detecting and mitigating against the effects of very high operator workload, yet techniques that can reliably detect and measure the onset or presence of cognitive underload remain largely elusive. This remains an area of concern particularly in safety critical tasks such as train driving because without knowing when a driver is experiencing underload, it is not possible to deliver a targeted mitigation.

The purpose of the current research was thus to understand whether it is possible to establish a set of reliable, objective physiological measures that could detect the presence of underload in a sample of train drivers.

2 Literature Review

2.1 Aims of the literature review

The overall aim of this literature review was to inform the design of the current experiment to detect cognitive underload. The review was conducted in order to:

- Understand in detail the physiological techniques used to measure low operator workload, identified during the previous literature review. We define low workload as being the conditions under which a person experiences cognitive underload. This tends to occur when an operator is performing a task that requires little cognition, effort or attention and the operator has little motivation to invest cognition, effort and attention into performing the task (Gimeno et al, 2006).
- Identify and describe any new techniques used since the previous literature review which was carried out in 2013 (and reported by Robinson et al, 2015).
- Provide detail on other methodological considerations that affect the current study, such as the design of the participant's task.

2.2 Methodology

To find the most relevant literature on the topic, several academic databases were used such as ScienceDirect, Sage Journals, Taylor and Francis and Google Scholar. The key search terms, as listed below, were searched for in these databases with each term providing a large amount of literature on the subject. Abstracts from the literature were then read, with papers discarded or kept dependent on their relevance to the topic.

Key search terms:

Boredom

- Cognitive and affective boredom
- Under or low stimulation
- Monotony

Workload

- Underload
- Task difficulty
- Demands
- Memory load
- Effort
- Time on task

Vigilance

- Attention

- Arousal

Physiological measures

- EEG
- Electrodermal activity
- Skin conductance
- Heart rate and heart rate variability
- Blood pressure
- Respiration
- Eye activity
- Posture
- Facial expression

2.3 Summary of literature review

The focus of the review was on metrics and measures for cognitive underload. A broader review of underload literature can be found in Robinson et al (2015) and Dunn (2011), but as background, it is important to understand the following models:

- The well-established Yerkes-Dodson Law (see Dunn, 2011, for a more comprehensive discussion) states that if arousal increases, then task performance will improve up to an optimal point. After this point, performance gets worse as arousal increases. Workload theory assumes that workload is strongly related to arousal.
- Young and Stanton's (2002) malleable attentional resource theory (MART) states that the size of an individual's pool of available attentional resources will adapt to suit the demands of the task. With cognitive underload an individual's pool of available attentional resources will shrink in line with task demands. Increases in task demand then become difficult to accommodate, resulting in performance decrements.
- Hockey's (1997) compensatory control model suggests that human performance is regulated through adjusting effort and goals, and that with easy tasks, mental effort is reduced with the aim of conservation. It also implies that people can "try harder" in order to maintain or improve their performance and achieve goals, but at some subjective, behavioural and psychological cost.

The review carried out for the current study describes the different physiological indicators used in a wide range of previous studies to measure workload, how each measure was used in the context of the study, the different tasks that participants were asked to carry out and key findings from the studies. The full text of the review can be found in Appendix A.

In summary, research has shown that there are several potential indicators of low mental workload. Each physiological measurement technique has its strengths and weaknesses (see table 3). For example an advantage of using techniques such as heart rate is that they are less intrusive and restrictive, therefore allowing participants to perform the task as they normally would (Yu et al, 2011). Other measures, such as facial expression, have certain disadvantages such as requiring a large amount of training to understand how to set up the equipment, and on how to correctly analyse the data (Stone and Wei, 2011). These techniques may, therefore, not provide the most efficient use of experimenter time and effort.

For most measures, the literature shows mixed results. This can especially be seen in EEG studies. Some analyses of different wave bands (e.g. theta waves) contradict expectations and there have been inconsistent results between different studies. It may be that the relationship between low workload and the physiological effects it produces are neither straightforward nor fully understood.

The review has shown that none of the measures are perfect on their own; several physiological measures used at the same time could produce the most useful results. Based on the studies considered within this review, it seems that body movement and posture (Frank, 2006; Graf et al, 1995; Qui and Helbig, 2012 and Roge et al, 2001), as well as respiration rates (Karavidas et al, 2010; Veltman and Gaillard, 2010; Backs et al, 2000 and Yamakoshi et al, 2009) provide the most consistency between predicted and obtained results, and consistency between studies.

From the few studies which look specifically at underload, Harris et al (1988) used heart rate, heart rate variability, pupil diameter and EEG as physiological measures with some success. Braby et al (1993) used heart rate and heart rate variability but reported unpromising results. Young and Stanton (2002) used eye movement, coupled with a secondary task with limited success.

Table 1 shows the strengths and weaknesses associated with the various physiological measures covered in this literature review.

Table 1: Physiological measures' strengths and weaknesses

Physiological measure	Strengths	Weaknesses	Key references
Electrical brain activity (EEG)	A large number of past studies have used EEG as their preferred measure Provides a direct measure of brain activity	Complex analysis Difficult to setup and administer as familiarisation with the equipment will be required Equipment available is not very portable	Belyavin and Wright, (1987) Brookhuis and De Ward, (2010) Gevins et al, (1997) Gevins et al, (1998)

Physiological measure	Strengths	Weaknesses	Key references
		Invasive due to the conductive gel used to attach electrodes Conflicting past research showing mixed results in terms of the EEG correlates of low workload	Kumashiro, (2005) Lal and Craig, (2001) Murata, (2005) Wilson and Russell, (2003)
Cardiovascular	Heart rate and heart rate variability measures are less intrusive and restrictive compared with EEG Easy to administer and cheap to obtain equipment	Blood pressure measures are restrictive Reliability of data collected is particularly dependent on the type of equipment used and level of experimental control	Backs and Seljos, (1993) Bonner and Wilson, (2002) Dijksterhuis et al, (2011) Jap et al, (2009) Megaw, (2005) Schmidt et al, (2009) Stuiver et al, (2014) Yu et al, (2011)
Eye activity	Unobtrusive measures can be taken from video recordings All the main classes of information can be measured at once Consistent blink rate results found for both types of task	Results are affected by the perceptual demands of the task For manual analysis, time and effort can be high depending on measures of interest Analysis by software/hardware is possible but can be expensive to buy	Ahlstrom and Friedman-Berg (2006) Benedetto et al, (2011) Brookings et al, (1996) Chen and Epps, (2013) Lecret and Pottier, (1971) Piquado et al, (2010) Stern el al, (1994)

Physiological measure	Strengths	Weaknesses	Key references
			Van Orden et al, (2000) Veltman and Gaillard, (1998)
Respiration	Respiration rate measures allow for simple analysis Easy to set up and relatively non-invasive Promising results from previous research on certain measures such as respiration rate	Potential confound between changes in respiration rate caused by physical activity and cognitive underload	Backs et al, (2000) Karavidas et al, (2010) Leino et al., (2001) Veltman and Gailard, (1998) Veltman and Gaillard, (2010) Yamakoshi et al, (2009)
Electrodermal	Recently developed wrist sensors are less restrictive than finger worn electrodes Relatively cheap Promising results from previous research	Finger sensors are restrictive and very sensitive to movement related artifacts (though other measurement techniques are available)	Collet et al, (2014) Mehler et al, (2012) Poh et al, (2010) Richter et al, (2010)
Facial expression	Non-invasive Measures can identify barely visible facial movements	Large amount of training is required to understand the setup of the equipment Requires specialist software The electrodes attached to the participants' face limit facial activities Complex and time consuming analysis needed Few studies have used facial expression as a measure of low workload	Capa et al, (2008) Ekman and Friesen, (1978) Dinges et al, (2005) Stone and Wei, (2011) Veldhuizen et al, (2003)

Physiological measure	Strengths	Weaknesses	Key references
		The evidence base appears to be weak, and this is in part due to the wide range of measures and techniques that have been used.	
Posture and movement	Consistent and expected results for both types of task Cameras and pressure sensors are not intrusive	Based on behaviour, not mental processes Time needed to analyse video footage (if using camera-based system) Few studies have used posture and movement as a measure of low workload	Balaban et al, (2004) Frank, (2006) Graf et al, (1995) Qui and Helbig, (2012) Jagannath and Balasubramanian, (2014) Roge et al, (2001)

2.4 Selection of physiological measures

In selecting the physiological measures for use within this study, a balance was struck between selecting measures which gave the most consistently promising results in previous studies, and pragmatic issues such as the feasibility of this small scale project to obtain equipment, the level of intrusiveness in collecting the data, and the practicalities of producing a suitable analysis with the available resources. In addition, consideration was given to selecting a range of techniques, as seemingly no single measure emerged from the review as being completely consistent in detecting cognitive underload.

Based on the analysis in Table 3, facial expression and EEG measures were ruled out on the basis that both of these techniques require substantial training, involve complex and time consuming data analyses, and because the attachment of electrodes is a fairly invasive procedure compared with other techniques.

Body posture movements were thought to have the potential to yield some useful results, however, the limited supporting evidence for using this technique as a measure of low workload, coupled with the time needed for manual examination and coding of data (unless a suitable computer based solution could be identified), made this technique less appealing for use in the current study than alternative techniques.

The potential for including eye activity measures (in particular blink rates and eye movements) was pursued, as these have shown promising results from previous research, are easy to set up and are relatively non-invasive. However, difficulties were encountered in obtaining equipment which would automate data analysis.

Based on the evidence presented in this literature review, and taking into account practical considerations, cardiovascular, respiration and electrodermal measures were selected for inclusion in the study.

2.5 Participant task

It was evident in all the studies reviewed that there are two main task designs that investigators use to investigate low cognitive workload conditions. These are:

- combining task difficulties ranging from very easy tasks (low workload) to hard tasks (high workload) and
- vigilance decrement tasks (involving monotonous tasks).

From the studies reviewed, some of the measures such as EEG show results which differ depending on which type of task is chosen. Each type of physiological measure has been adopted in studies which use both types of task. Overall it seems that the vigilance decrement tasks provide results which tend to be better matched with the effects expected in underload conditions. It is unclear if the easy, low workload tasks are in fact easy and repetitive enough to result in underload effects. This is most obvious in the case of eye activity (namely blink rates and durations). Monotonous tasks seem to produce statistically significant results consistent with physiological effects associated with underload, whereas experimenters who use the different workload tasks report conflicting results. Task design is therefore very important to consider when attempting to measure cognitive underload. The concern about cognitive underload in train drivers stems from the long, and at times, boring and repetitive nature of their task. On this basis a decision was taken to use a vigilance task during the experiment.

2.6 Summary of implications for experimental design

This review investigated research that has been undertaken to investigate the effects of low operator workload, and to identify measurement techniques to inform the design of the planned of study. Although a range of different physiological measures have been used in research studies for the assessment of operator cognitive underload, only a few are viable for use outside of a laboratory setting. In particular, it is suggested that cardiac, respiration and electrodermal measures are most suitable and these measures were selected for inclusion in the trial completed by professional train drivers.

The review also highlighted various options for the participant task. Considering how promising the different task designs have been in previous studies (see Section 2.5), and the nature of the train driving task, a vigilance task was chosen for the current study.

3 Method

3.1 Participants

Fifteen participants were recruited to take part in the study, with 14 males and 1 female. The participants were train drivers from Southeastern and South West Trains.

Table 2: Sample age, age when certified and experience (N = 15)

	Age (years)	Age at certification (years)	Train driving experience (years)
Mean	50.6	35.0	14.7
St Dev.	8.0	8.5	11.3
Max	64	54	44
Min	37	22	1

3.2 Experimental design

Participants were required to complete a modified version of the Sustained Attention to Response Task (SART - described in section 3.3) which is a computer-based repetitive vigilance task that was expected to induce conditions representative of cognitive underload. In selecting this task, a number of key issues were considered including the outcomes from the two literature reviews that were carried out in earlier stages of this research. Firstly, the reviews suggested that vigilance decrement tasks (involving lengthy monotonous tasks) were more likely to produce results in the literature that were consistent with the physiological effects associated with underload, compared with studies in which workload was manipulated (high vs low workload conditions). Secondly, performance on the SART generally reflects the premise of the effect of cognitive underload on task performance, namely:

- Decreased vigilance
- Increased reaction times
- Increase in number of performance errors
- Increase in lapses in attention to target stimuli

Thirdly, the SART is considered to be a robust and valid measure of sustained attention, and furthermore, performance on the SART has been found to correlate with attentional failures in everyday life (Manly et al., 1999) indicative of real-world applicability and ecological validity.

Finally, the issue of cognitive underload in train drivers stems largely from the lengthy, repetitive nature of the task in which drivers must continuously monitor and respond to repetitive stimuli in the environment, whilst at the same time maintaining a preparedness to react to infrequent critical events. On this basis, the SART has certain

similarities with train driving, and has been referenced in the literature as such (e.g. Robertson & Garavan, 2004).

In order to observe the predicted performance decrement with time on task, participants were required to complete four consecutive SART trial blocks (as opposed to one block as per the original SART). Physiological parameters thought to be sensitive to low workload conditions (EDA, heart rate and HRV, respiration rate and breath depth) were simultaneously recorded. This approach allowed for the relationship between task performance and physiological correlates to be investigated.

So as not to influence task performance, subjective ratings of workload were also obtained via a post-trial questionnaire in which participants were required to rate the level of effort required in the first half of the trial compared with the second half.

Cognitive underload is thought to be a separate construct to fatigue, although it is acknowledged that the physiological indicators for both of these constructs are likely to be highly correlated with one another, making it difficult to disentangle the potential contributory effects of fatigue and/or underload in the resulting physiological data. For this reason, the study also asked participants to rate their level of fatigue (using the Karolinska sleepiness scale - KSS) before and after the trial. While it was predicted that task performance and physiological indicators of underload would be highly correlated, no such correlation was expected to exist between task performance and fatigue ratings, providing evidence in favour of a genuine presence of cognitive underload rather than fatigue with increased time on task.

The following outcome variables across the four SART stimulus presentation blocks were examined:

- Task performance measures:
    - Reaction times to non-target stimuli
    - Variability of reaction times to non-target stimuli
    - Frequency of commission errors (where a participant fails to withhold a response to the target stimulus) and omission errors (where a participant fails to respond to non-target stimuli)
- Subjective measures
    - Sleepiness/alertness
    - Subjective workload level
- Physiological measures
    - Heart rate and HRV, electrodermal activity, respiration rate and depth of breath

3.3 Materials

3.3.1 Sustained attention to response task (SART)

A computer based version of the sustained attention to response (SART) task was used to monitor participant's sustained attention to repetitive stimuli for the duration of the experiment. The original SART procedure, developed by Robertson et al. (1997), involves presenting 225 single digits (the numbers 1-9) randomly on a computer screen. Each individual digit is presented for 250ms, followed by a 900ms "mask". Participants are instructed to respond to each digit by pressing the spacebar as quickly as possible with their preferred hand, with the exception of the digit '3' which serves as the target 'No-Go' stimulus to which participants must withhold a response. The target appears 25 times in a random order (i.e. a 'No-Go' target probability of 0.11). The digit stimuli are presented centrally on the screen in one of five random font sizes. The "mask" consists of a round circle with an 'X' inside and is also presented centrally on the screen. Digit and mask stimuli are presented in white on a black background. The stimulus onset asynchrony is 1150ms.

The original SART lasts for just over 4 minutes which is considered long enough to detect performance decrements using a sample from the general population (Robertson et al., 1997). However, the recruitment process for train drivers excludes people with poor ability to maintain vigilance and attention. Thus, it was expected that the train driver population may generally perform better on vigilance tasks, such as the SART, compared with the general population, and may be more resilient to underload-inducing conditions. As a consequence, it was predicted that the performance decrement would take longer to appear for train drivers. For this reason, the current study adopted a longer version of the SART, with four blocks of 225 stimulus presentations, as opposed to the single block used in the original. This resulted in an experimental duration of approximately 17 minutes, increasing the repetitive nature of the task, and the likelihood of inducing conditions representative of cognitive underload. It also allowed for the trial to be broken into four distinct trial 'blocks', allowing task performance and physiological measures to be evaluated with respect to time-on-task across the four blocks.

3.3.2 Physiological recording equipment

Participants' physiological indicators of workload and arousal were measured using the following equipment:

- **Movisens EDA 'Move' Sensor -** for measuring electrodermal activity. This device involved placing two electrodes onto the participant's palm which were connected to a portable wrist-worn recording device.
- **Movisens EKG 'Move' Sensor –** for measuring cardiac activity (heart rate and heart rate variability). This device involved the participant wearing a chest strap with a removable recording device.

- **Nexus-10 MKII –** for measuring respiration and breath depth. This involved an additional chest strap, worn below the cardiac chest strap, which was connected to the Nexus recording device.

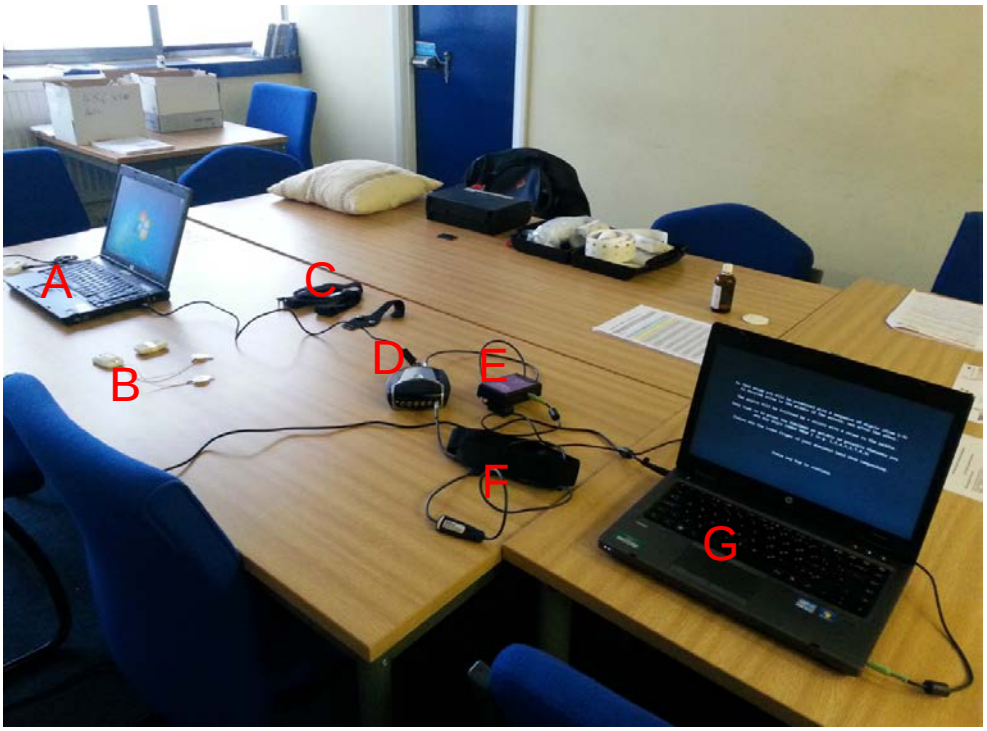


Figure 1- Experiment set-up

The components labelled in the image above were:

A- Laptop to record physiological measures

B- Movisens EDA 'Move' Sensor

C- Movisens EKG 'Move' Sensor

D- Nexus-10 MKII

E- Nexus trigger interface

F- Nexus respiration strap

G- Laptop to run SART

3.3.3 Questionnaires

Participants were asked to provide feedback before and after the trial in order for the experimenters to obtain subjective fatigue ratings, as well as detailed demographic

information and subjective workload ratings via a post-trial questionnaire. More specifically, the questionnaires used in the study were:

1. Demographic questionnaire (see Appendix B) - To collect basic demographic information about the participant (e.g. age, gender, years of train driving experience etc.), administered at the end of the experiment.
2. Boredom scale (see appendix B) – To understand participants' boredom levels during the first and second halves of the SART trial. The boredom scale was created for this study and required participants to rate their perceived boredom on a scale of 1 (not bored at all) to 10 (extremely bored). The decision was made that a simple boredom rating would be more feasible to administer and provide more accurate results than the more complex and lengthy stress arousal checklist used previously by Robinson et al (2015).
3. Karolinska sleepiness scale (KSS) (see Appendix C) - To gain an understanding of participant's level of sleepiness/alertness before and after the SART. Participants were required to rate their level of alertness on a scale of 1 (extremely alert) to 9 (extremely sleepy).
4. Subjective workload scale (see Appendix D) – Administered at the end of the trial, participants were required to compare their perceived level of effort for the first half of the trial compared with the second half. The NASA-TLX was used as the measure of subjective workload.

3.4 Procedure

Upon their arrival, participants were provided with some basic information about the trial and were asked to provide informed consent prior to taking part. They were then asked to sit comfortably in front of a laptop computer that was used to present the SART, and the physiological recording equipment was fitted by the experimenter. Participants provided a KSS rating, and then completed a brief practice of the SART so that they could become familiar with the task (18 stimulus presentations, with two targets). There was a brief pause at the end of the practice session which provided an opportunity for the experimenter to check that the participant understood the task and was happy to start the main experiment.

Four consecutive blocks of stimulus presentations were then presented whilst physiological parameters were simultaneously recorded. At the end of the trial, the participants were required to provide a second KSS rating, and were asked to complete the workload questionnaire. At the end of the trial, the physiological recording equipment was removed, and participants were required to complete a post-trial demographics questionnaire before leaving.

3.4.1 Participant information

Participants were given an information sheet to read before the trial began. This provided a description of what the study was about, and details of how the study would proceed and what they would be expected to do during the trial. This sheet was identical

to the information which participants were given in advance of the trial. It included a request that participants restrict their caffeine intake on the day of the trial so as not to effect the results. The information sheet also contained a description of the physiological sensors that they were required to wear during the trial.

3.4.2 Participant instructions

A standard set of instructions were provided to the participants prior to commencing the experiment. They were informed that they would be completing a computer based attention task in which a sequence of random digits from 1-9 would be presented, one after the other, on the laptop screen. Participants were instructed to respond to any digit other than the number '3' by pressing the spacebar as quickly as possible with their dominant hand. They were told that they should not respond if the digit '3' was presented. Participants were asked to respond as quickly and as accurately as possible.

Participants were also informed that they would be wearing physiological recording equipment during the trial.

3.4.3 Physiological data recording

EDA, heart rate, HRV and respiration data were recorded while participants completed the SART. The EDA sensor was worn on the non-dominant arm, whilst the ECG (heart rate) sensor was worn on the dominant arm. In addition, the Nexus 10 MKII device was be used to measure respiration and breath depth. This involved participants wearing a chest strap, which could be worn over their clothing

3.5 Research Hypotheses

3.5.1 Hypotheses relating to task performance

H1 Reaction time to 'Go' stimuli will decrease significantly with time on task

H2 Reaction time variability to 'Go' stimuli will increase with time on task

H3 The frequency of errors of commission and omission will increase with time on task

H4 The reaction times to the four stimuli before a 'No-Go' event will be higher before a correct 'withheld' response compared to an incorrect 'non-withheld' response

3.5.2 Hypotheses relating to driver's subjective ratings

H5 Perceived levels of fatigue will not differ significantly between pre and post experiment fatigue ratings

H6 Subjective workload levels will be significantly lower for the second half of the trial compared with the first half

H7 Perceived levels of boredom will be significantly higher for the second half of the trial compared with the first half

3.5.3 Hypotheses relating to physiological indicators of arousal and workload

H8 Heart rate (beats per minute) will reduce significantly with time on task

H9 Heart rate variability (HRV) will increase significantly with time on task

H10 Skin conductance level will reduce significantly with time on task

H11 Respiration rate will reduce significantly with time on task

H12 Breath depth will reduce significantly with time on task

H13 Collectively, the physiological and subjective indicators of workload and arousal will correlate with the task performance decrement measures with time on task

H14 Heart rate during the four seconds before a 'No-Go' event will be higher before a correctly 'withheld' response compared to an incorrectly 'non-withheld' response

H15 Skin conductance during the four seconds before a 'No-Go' event will be higher before a correct 'withheld' response compared to an incorrect 'non-withheld' response

H16 The frequency of Skin Conductance Responses (SCR) will decrease in the four seconds after a commission error

H17 SCR amplitude will decrease in the four seconds after a commission error

H18 Respiration rate during the four seconds before a 'No-Go' event will be higher before a correct 'withheld' response compared to an incorrect 'non-withheld' response

4 Results

Table 5 shows a breakdown of the number of stimulus presentations per participant, along with error rates. An error for 'Go' trials was failing to respond by pressing the spacebar (error of omission); an error for 'No-Go' trials was failing to withhold a response to the target stimulus (error of commission).

Table 5- Stimulus presentations and error rates

Stimulus type	Presentations per participant	Mean Errors	Std. Deviation	Accuracy (%)
'Go'	800	8.3	9.4	99
'No-Go'	100	32.6	16.6	67

4.1 Performance measures

4.1.1 Reaction times with time on task

Hypothesis H1 was that average reaction times to the 'Go' stimuli would decrease significantly with time on task.

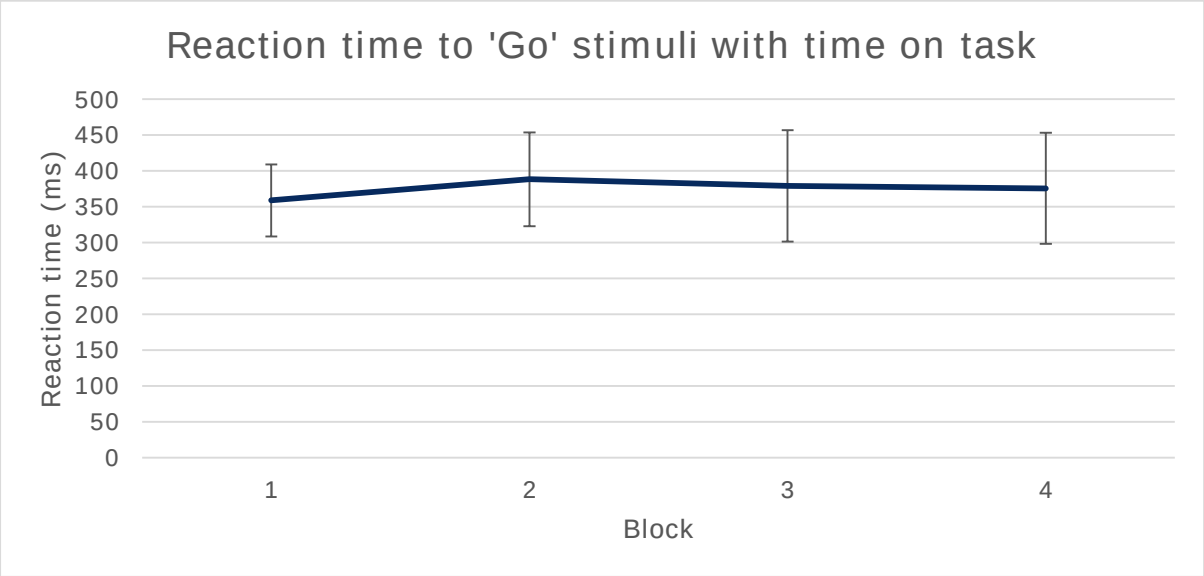


Figure 2- Reaction times with time on task

Figure 2 shows that there was very little change in the reaction times to the 'Go' stimuli with time on task. A one-factor repeated measures ANOVA found no significant difference between the four blocks ($F(3, 42)=1.332$, $p=.277$).

4.1.2 Reaction time variability with time on task

Hypothesis H2 was that the average reaction time variability to 'Go' stimuli would increase significantly with time on task.

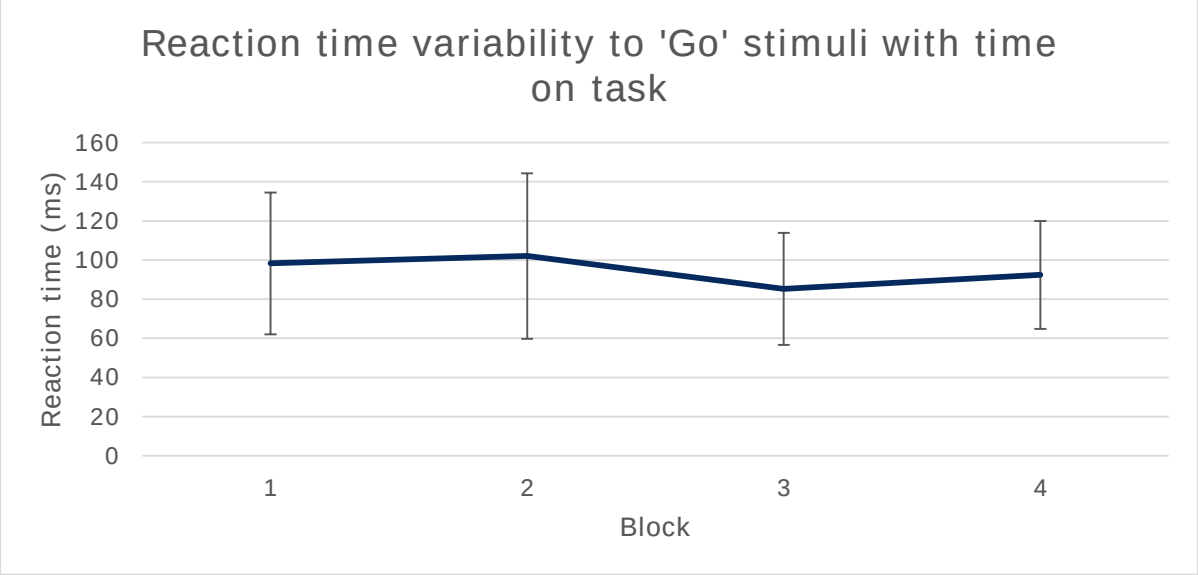


Figure 3- Reaction time variability to 'Go' stimuli with time on task

Figure 3 shows that the differences in the reaction time variability to 'Go' stimuli were minimal over the four blocks. A one-factor repeated measures ANOVA confirmed that there was not a significant difference between the four blocks ($F(3, 42)=2.035$, $p=.135$).

4.1.3 Errors of commission and omission

Hypothesis H3 was that the frequency of errors of commission and omission for both 'Go and 'No-Go' stimuli will increase with time on task.

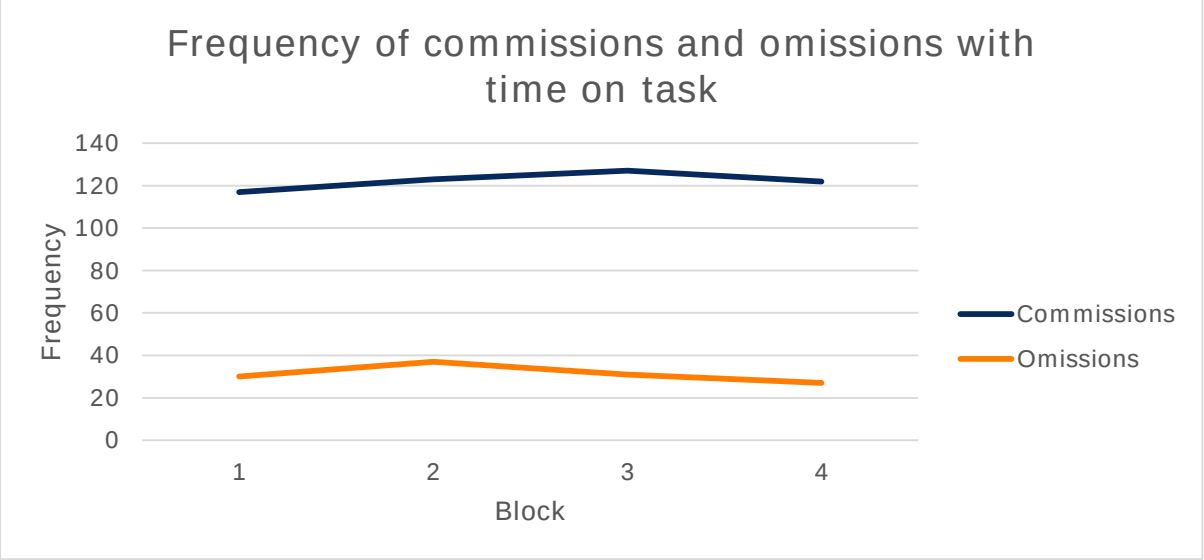


Figure 4- Frequency of commissions and omissions with time on task

Figure 4 shows the differences in the number of commission errors with time on task. A one-factor repeated measures ANOVA showed that there was not a significant difference over the four blocks ($F(3,36)=1.125$, $p=.352$).

Figure 4 also shows the difference in the number of omission errors with time on task. A one-factor repeated measures ANOVA test revealed that there was no significant difference over the four blocks ($F(3,6)=.014$, $p=.925$).

4.1.4 Reaction time four stimuli before 'No-Go'

Participants' reaction times to stimuli immediately before their correct and incorrect responses to 'No-Go' stimuli were calculated and compared (following Robertson, 1997). The aim was to investigate whether reaction time to four 'Go' stimuli preceding a 'No-Go' stimulus would be different if the participant responded correctly to the 'No-Go' stimulus, compared to situations where they responded incorrectly to the 'No-Go' stimulus. It was important to control for the influence of other 'No-Go' stimuli on reaction time, so the data was cleansed to exclude situations in which a 'No-Go' stimulus was present among the four stimuli preceding another 'No-Go' stimulus. Out of the 1500 'No-Go' stimuli presented to participants, 570 (38%) of these were filtered out due to this overlap. Fewer data points would have been lost if shorter periods of time were analysed; however, this would have been inconsistent with the previous analysis by (Robertson, 1997) and would have given even shorter periods in which to detect physiological change.

Hypothesis H4 was that the reaction times to the four stimuli before a 'No-Go' event would be greater before a correct 'withheld' response compared to an incorrect 'non-withheld' response.

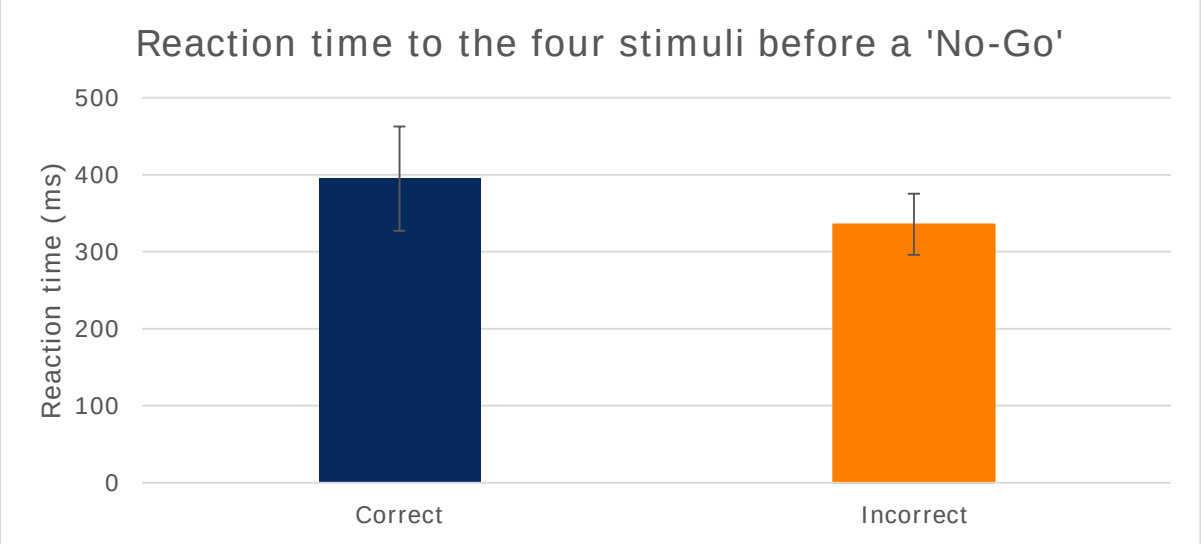


Figure 5- Reaction time to the four stimuli before a 'No-Go'

Figure 5 shows mean reaction times to the four stimuli before the 'No-Go' stimuli. There is a clear difference between the reaction times of stimuli four before a correct 'withheld' response than four before an incorrect 'non-withheld' response. A paired samples t-test confirmed that this difference was statistically significant (t(14)=-3.649, p=.003).

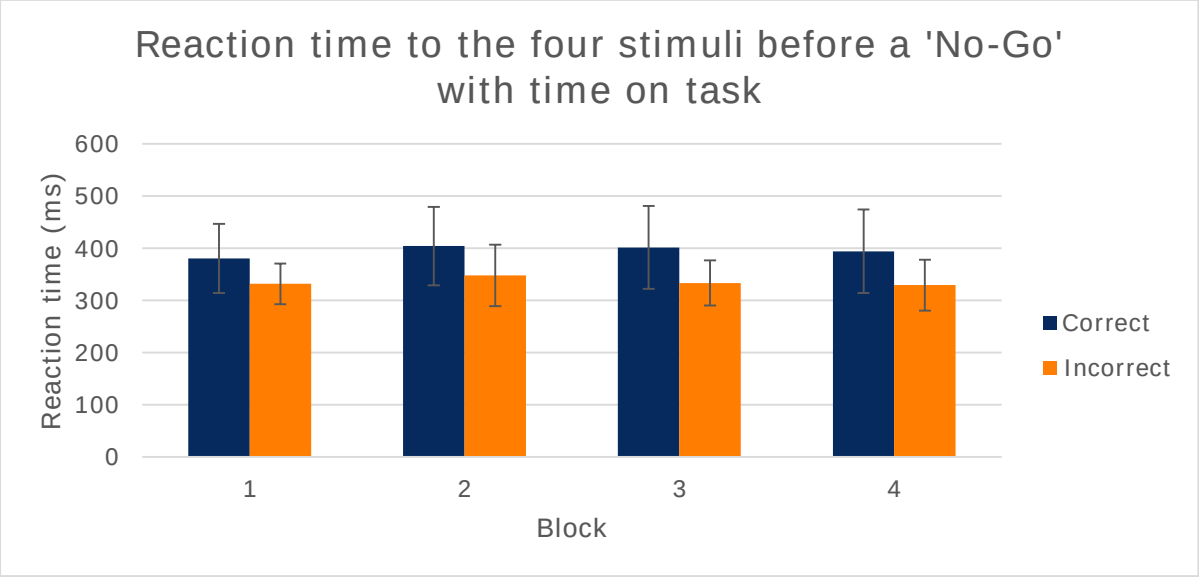


Figure 6- Reaction time to the four stimuli before a 'No-Go' with time on task

Figure 6 shows the reaction time to the four stimuli before a 'No-Go' over time on task, the pattern was very similar over the 4 blocks. One-factor repeated measures ANOVA revealed that there was no significant difference over the four blocks: correct response (F(3,39)=.927, p=.437) and incorrect response (F(3,33)=.381, p=.767)

4.2 Subjective measures

4.2.1 Karolinska Sleepiness Scale (KSS)

Hypothesis H5 was that the perceived levels of fatigue would not differ significantly between pre and post experiment fatigue ratings.

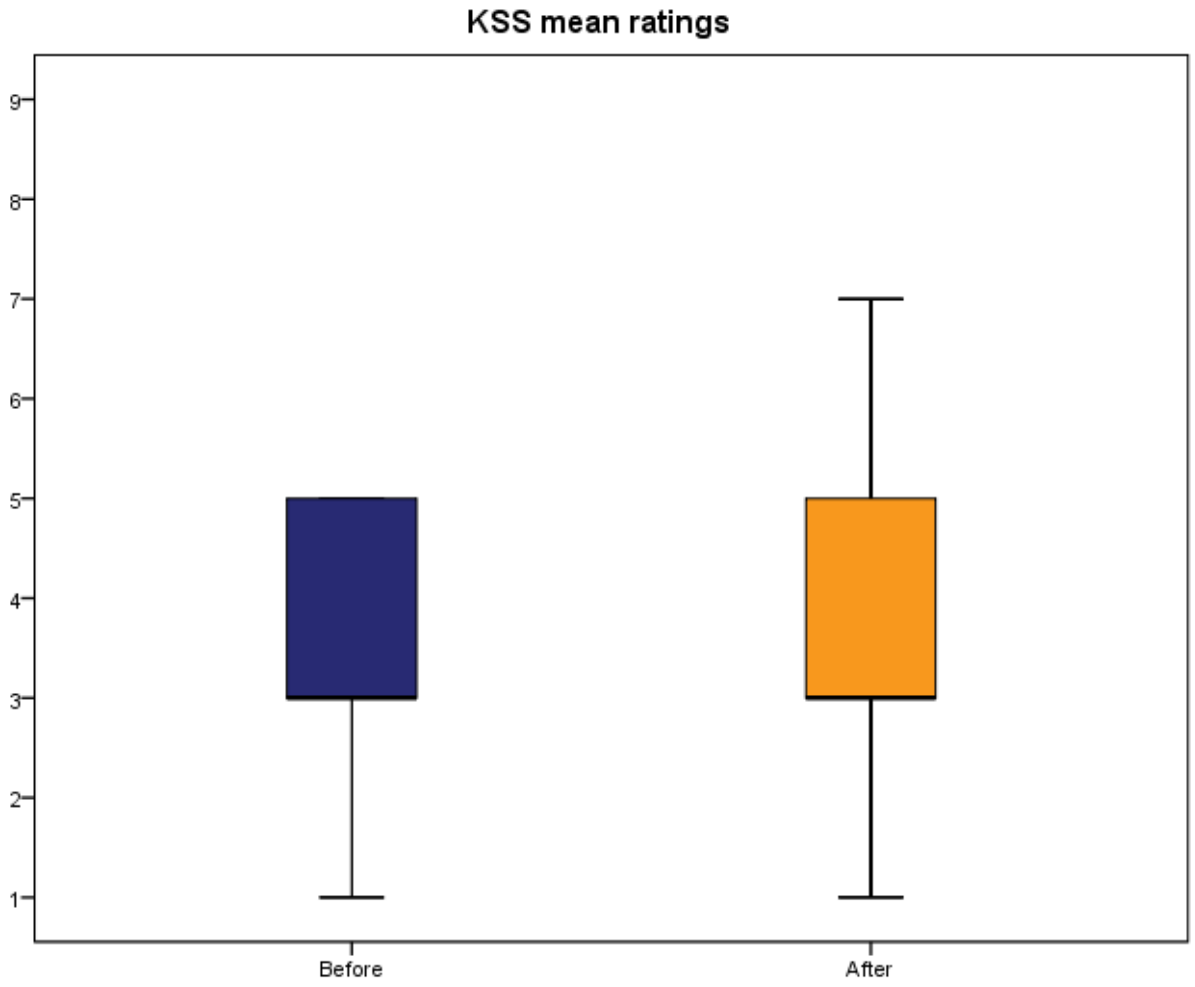


Figure 7- KSS mean ratings

Figure 7 shows that there was very little difference in the sleepiness ratings before and after the trial. The results were not normally distributed, therefore a Wilcoxon Signed Ranks Test was used. Analysis showed that the results were not significantly different (Z=-.322, p=.748).

4.2.2 NASA TLX

Hypothesis H6 was that the perceived levels of workload would be significantly lower for the second half of the trail compared with the first half.

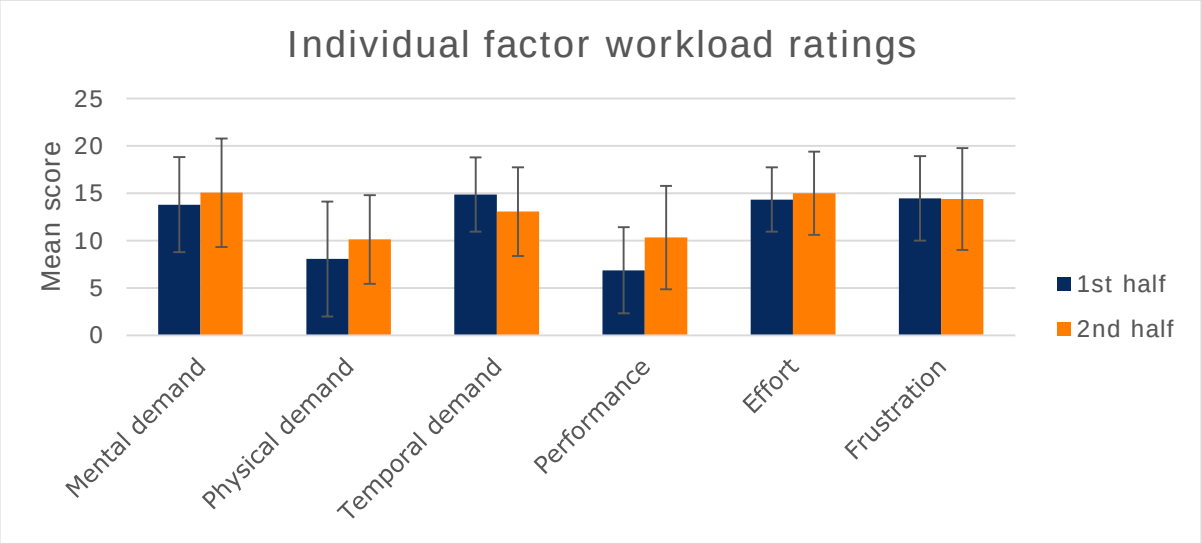


Figure 8- Individual factor workload ratings

Figure 8 shows the average ratings for each individual factor in the NASA TLX. Paired samples t-tests showed that the perceived levels of performance significantly decreased in the second half compared with the first ($t(14)=2.525$, $p=.024$). A paired samples t-test showed that the perceived levels of physical demand approached significance ($t(14)=-1.661$, $p=.119$) with higher levels of perceived workload reported in the second half compared with the first. No significant differences were observed for the other subscales.

Overall workload has not been calculated for participants in this study. The reasoning for this is that for the NASA TLX, a higher perceived performance rating is usually thought to be linked to a lower overall workload, so this subscale is usually reversed and scores added to the other scores. This procedure would have been misleading in our study, because our hypothesis is that performance improves with increasing workload.

4.2.3 Boredom ratings

Hypothesis H7 was that the boredom ratings would significantly increase for the second half of the trial compared with the first half.

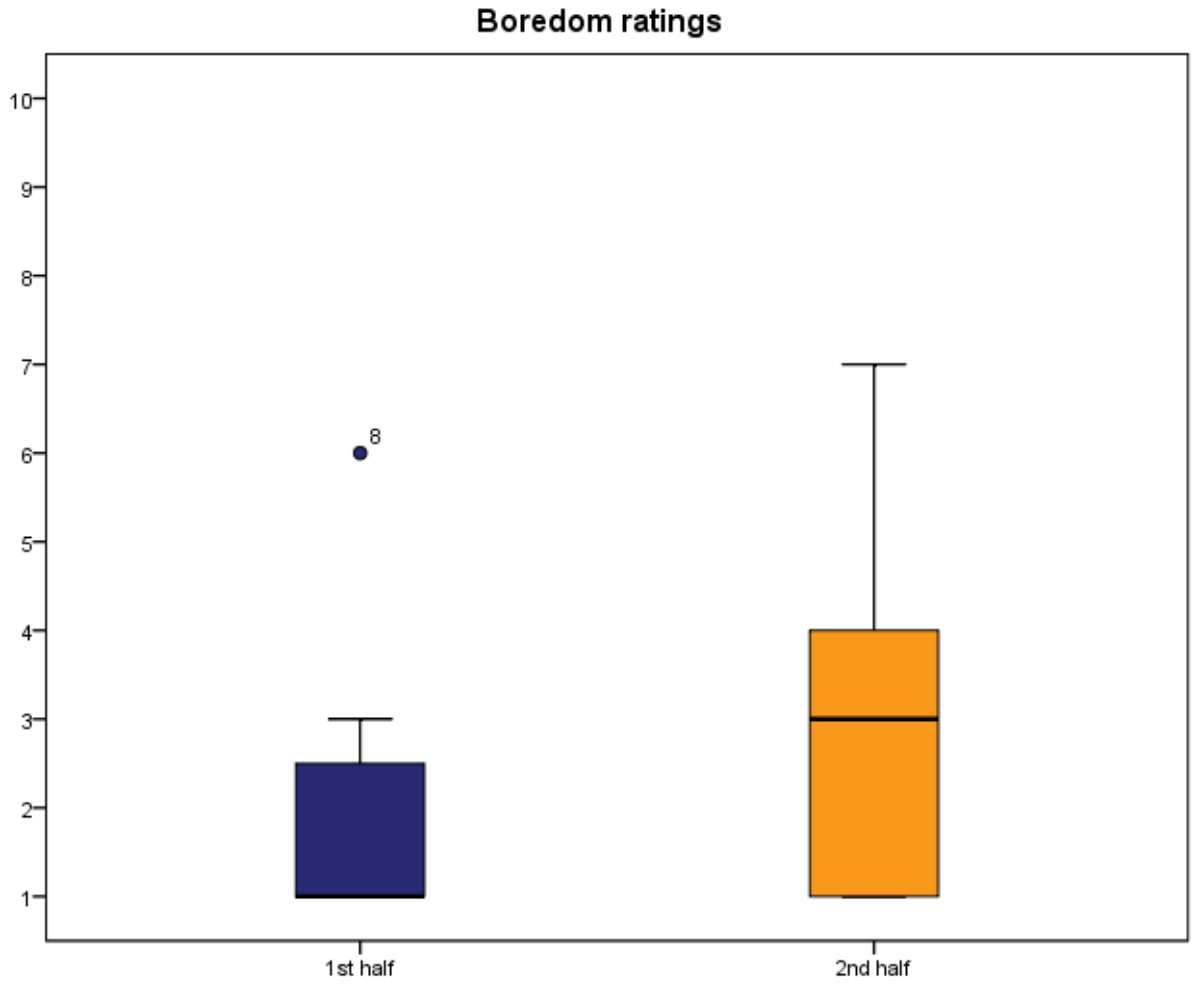


Figure 9- Boredom ratings

Figure 9 shows that boredom ratings were considerably higher for the second half of the trial than for the first half. Data were not normally distributed so a Wilcoxon Signed Ranks test was conducted. The analysis confirmed that this difference was statistically significant (Z=-2.549, p=.011).

4.3 Physiological measures

4.3.1 Cardiac Measures over time

Prior to conducting the analyses on the heart rate and HRV data, a visual inspection of each participant's data was carried out in order to identify any artefact that could have been caused by poor signal connection or excessive movement. Further inspections were carried out by observing the distribution of the data using histograms and box-plots. From the heart rate data, one participant's dataset was identified as containing a large

number of extreme outliers and was subsequently excluded from further analysis. From the heart rate variability data two participants' data was identified as containing large numbers of extreme outliers and were excluded from further analysis.

Hypothesis H8 was that the heart rate would reduce significantly with time on task.

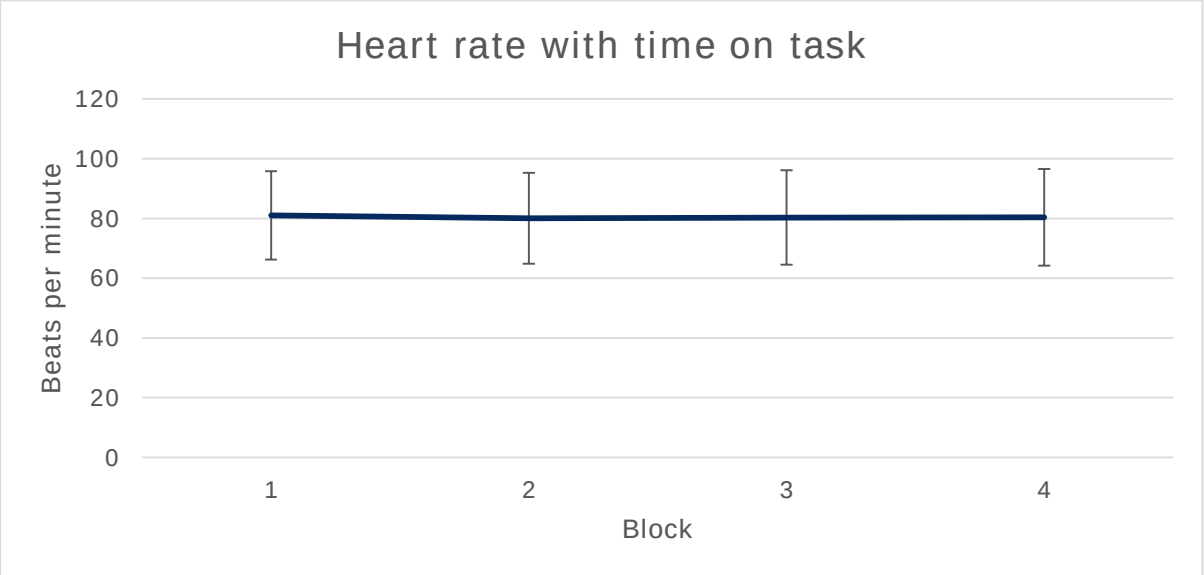


Figure 10- Heart rate with time on task

Figure 10 shows very little difference in the heart rate with time on task. A one-factor repeated measures ANOVA confirmed that there was not a significant difference between the four blocks ($F(3, 39)=.798$, $p=.502$).

Hypothesis H9 was that heart rate variability (HRV) would increase significantly with time on task. **Table 6** shows commonly used HRV metrics that were calculated for each of the four SART blocks in both the timing and frequency domains:

HRV metrics in the timing domain:

- **Mean RR:** the mean of R-R intervals, where R-R is calculated from the peak of one R wave to the peak of the next R wave in the QRS complex (normal heartbeat).
- **SDNN:** The standard deviation of R-R intervals
- **RMSSD:** The root-mean square of successive differences of R-R intervals
- **PNN50:** The percentage of adjacent R-R intervals that are greater than 50ms

HRV metrics in the frequency domain:

- **LF Power:** Fast Fourier Transform (FFT) power density spectrum in the low frequency band (0.04-0.15Hz). LF is associated with the heart's parasympathetic and sympathetic activity and typically increases with mental workload.
- **HF Power:** FFT power density spectrum in the high frequency band (0.15-0.4Hz). HF is associated with respiratory activity and typically reduces with mental workload

- **LF/HF Ratio:** The ratio between LF and HF band power. Mental workload is typically associated with an increase in the LF/HF ratio.

Table 6. Mean HRV metrics across the four blocks

Variable	Block 1	Block 2	Block 3	Block 4	*p* value
Mean RR (ms)	770.74	781.21	781.83	782.95	.327
SDNN (ms)	41.31	34.81	35.31	34.90	.077
RMSSD (ms)	29.00	26.95	24.57	25.39	.162
pNN50 (%)	11.77	10.13	8.00	8.25	.049*
LF Power (ms^2)	472.79	545.14	397.96	342.64	.326
HF Power (ms^2)	360.64	317.41	240.49	231.32	.160
LF/HF Ratio (ms^2)	2.51	2.33	2.85	2.94	.640

**significant at the 0.05 level*

Repeated measures ANOVA revealed that only the pNN50 metric, which is the percentage of adjacent R-R intervals greater than 50ms, showed a significant difference across the four blocks. Post hoc comparisons showed that block three had significantly lower mean pNN50 compared with block 1 ($p=.031$), but there were no other significant differences between any other blocks (all ps >.05). These results are contradictory to our hypothesis and a recent study by Cinaz et al (2011) who reported a decrease in the pNN50 time domain with increased mental workload, and may suggest that the SART increased arousal.

In order to examine beat-to-beat variability with greater temporal granularity, further analyses were performed on the HRV data. This involved creating small overlapping epochs with window lengths of 2 minutes and a step size of 30 seconds. For each epoch, mean values were calculated, and plotted on individual graphs to show overall patterns of the HRV metrics across the duration of the SART (see Figures 11-16).

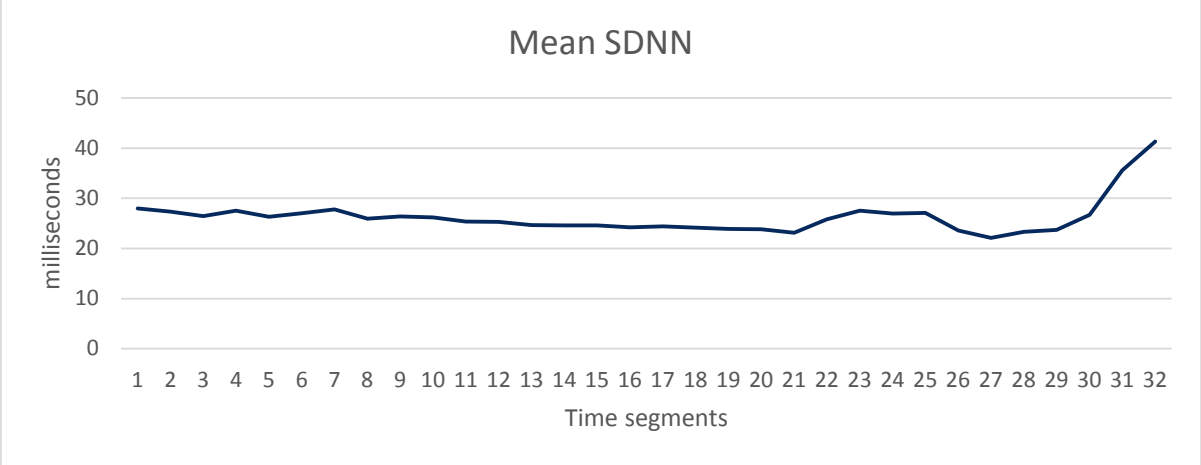


Figure 11- Mean standard deviation of R-R intervals over time

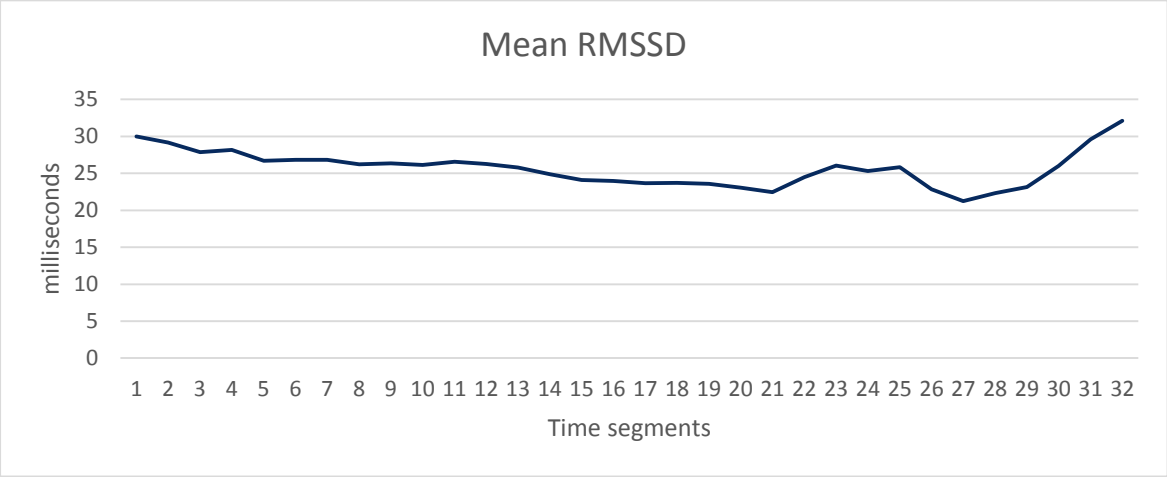


Figure 12- Mean root-mean square of successive differences of R-R intervals over time

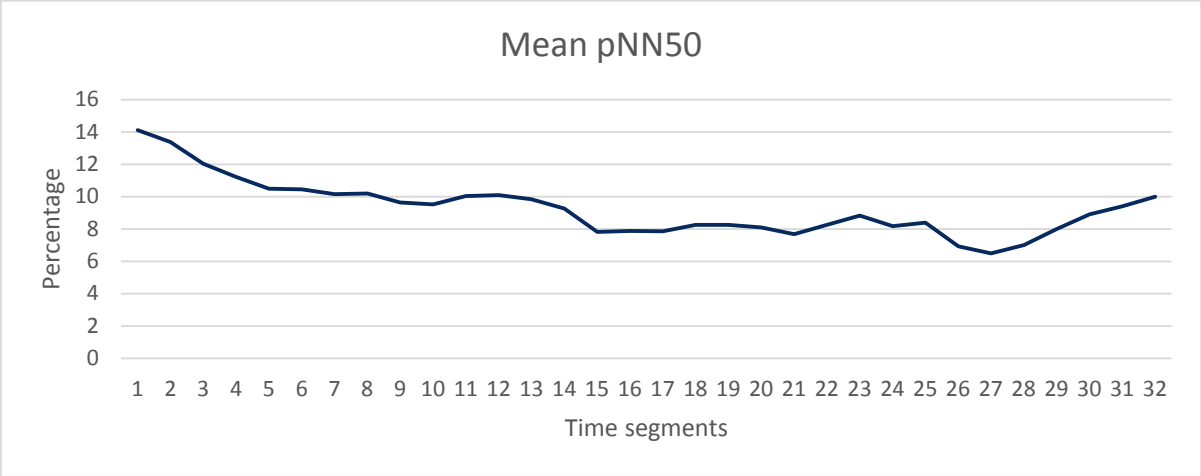


Figure 13- Mean percentage of adjacent R-R intervals greater than 50ms over time

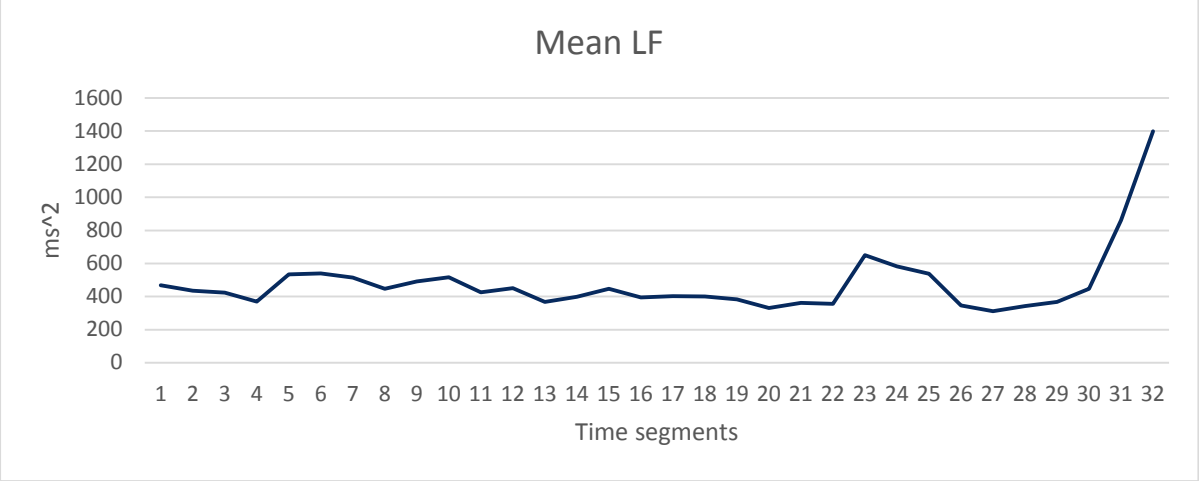


Figure 14- Mean low frequency band over time

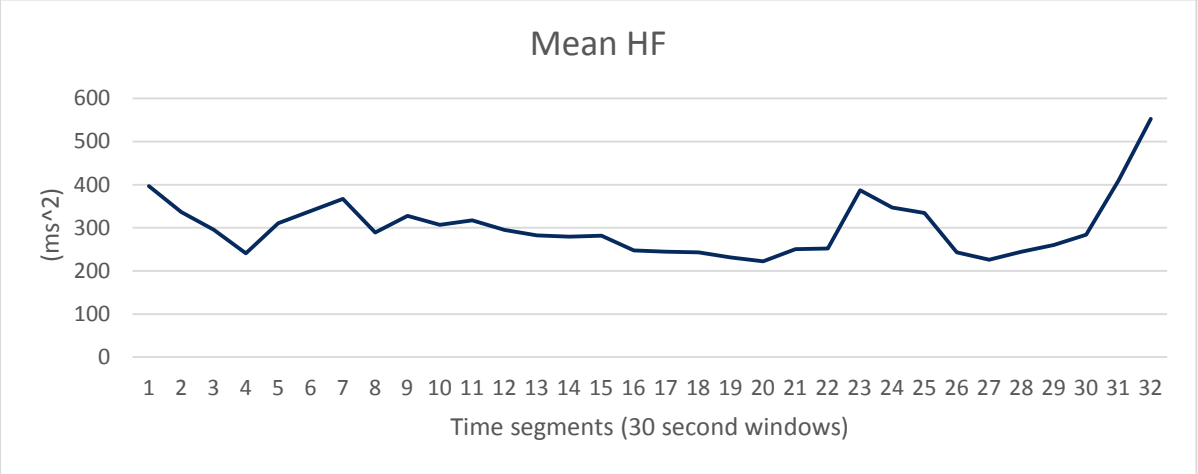


Figure 15- Mean high frequency band over time

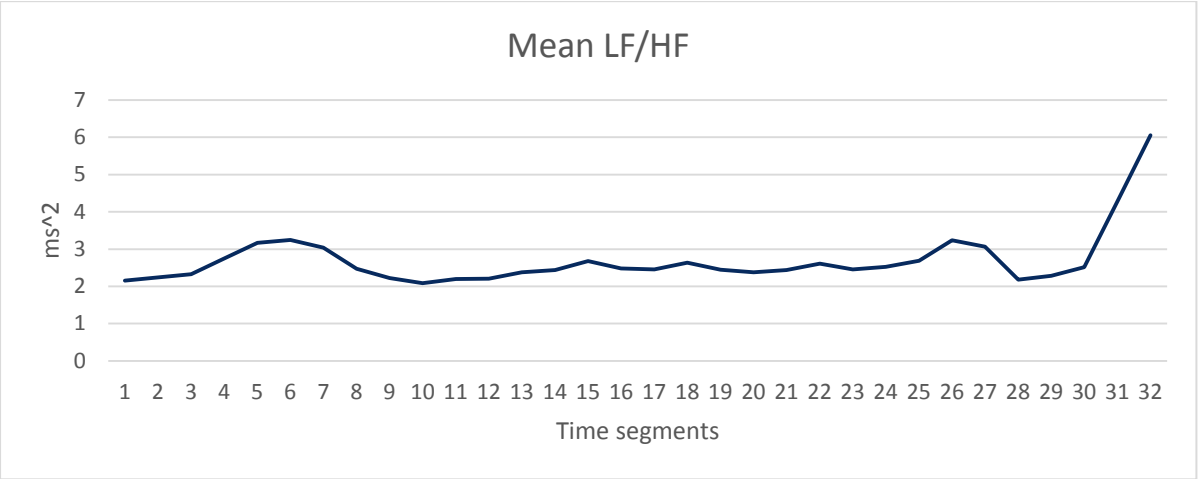


Figure 16- Mean ratio between low frequency and high frequency band power over time

Figures 11-16 show that, in general, all of the HRV metrics tended to show a small decrease over time, however there was a noticeable increase towards the end of the task. The reason for this increase is unclear. Appropriate steps were taken to make sure that participants were not in view of a clock or timer, to ensure that anticipation of the end of the task was not a factor. Data checks were also carried out to ensure correct synchronisation of the different data feeds.

4.3.2 Electrodermal activity over time

Prior to conducting the analyses on the electrodermal activity, a visual inspection of each participant's data was carried out in order to identify any artefact. From the skin conductance data, one participant's data set was identified as containing excessive artefact due to a poor connection and was subsequently excluded from further analyses.

Hypothesis H11 was that there would be significant reduction in skin conductance level with time on task.

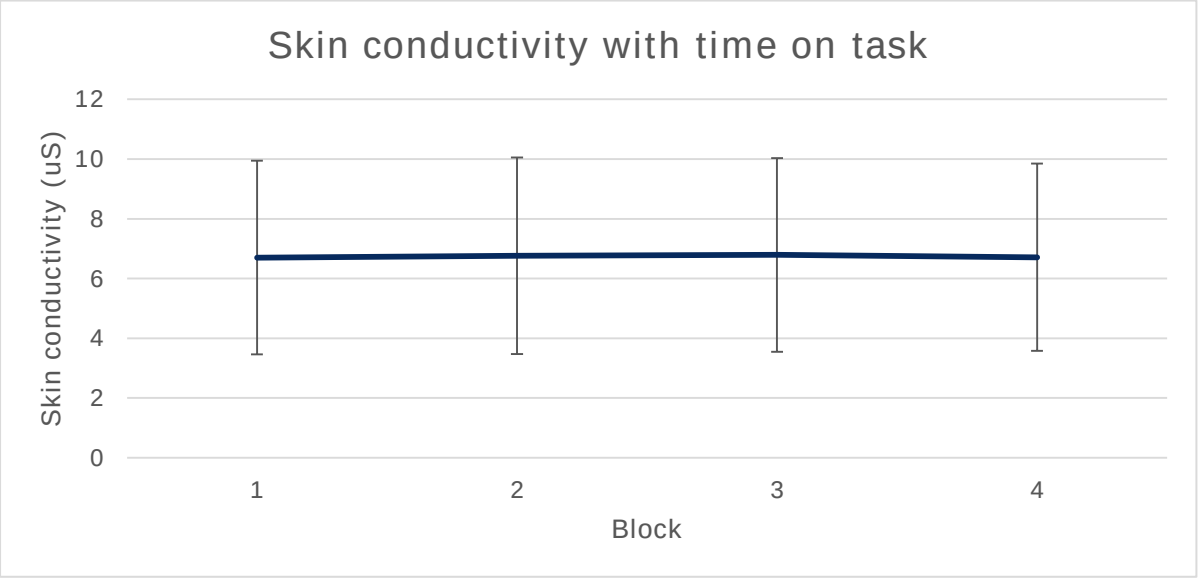


Figure 17- Skin conductivity level with time on task

So as to prevent the SCL data from being artificially elevated, the amplitudes of occurring skin conductance responses (SCRs) were first subtracted from the background SCL signal. Figure 17 shows the mean corrected SCL over the four blocks. A one-factor repeated measures ANOVA confirmed that there was not a significant difference between the four blocks ($F(3, 39)=.257$, $p=.856$).

4.3.3 Respiration measures over time

Prior to conducting the analyses on the respiration measures, a visual inspection of each participant's data was carried out in order to identify any artefact. All participants' data sets were included in further analysis.

Hypothesis H12 was that there would be a significant reduction in the respiration rate with time on task.

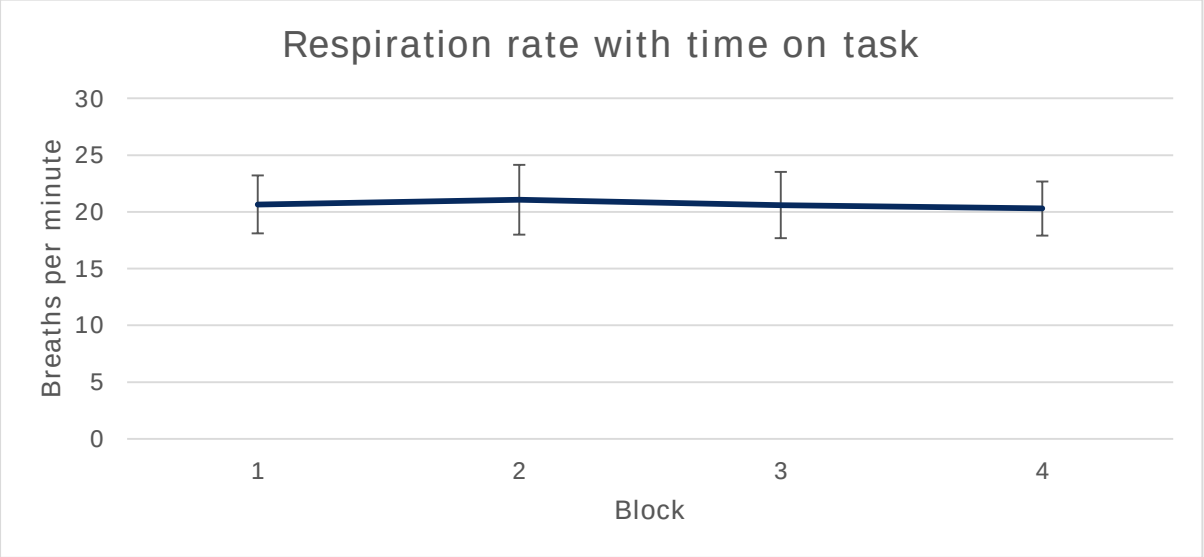


Figure 18- Respiration rate with time on task

Figure 18 shows the average respiration rate over the four blocks. A one-factor repeated measures ANOVA confirmed that there was not a significant difference between the four blocks ($F(3, 42)=1.915$, $p=.142$).

Hypothesis H13 was that there would be a significant reduction in breath depth with time on task.

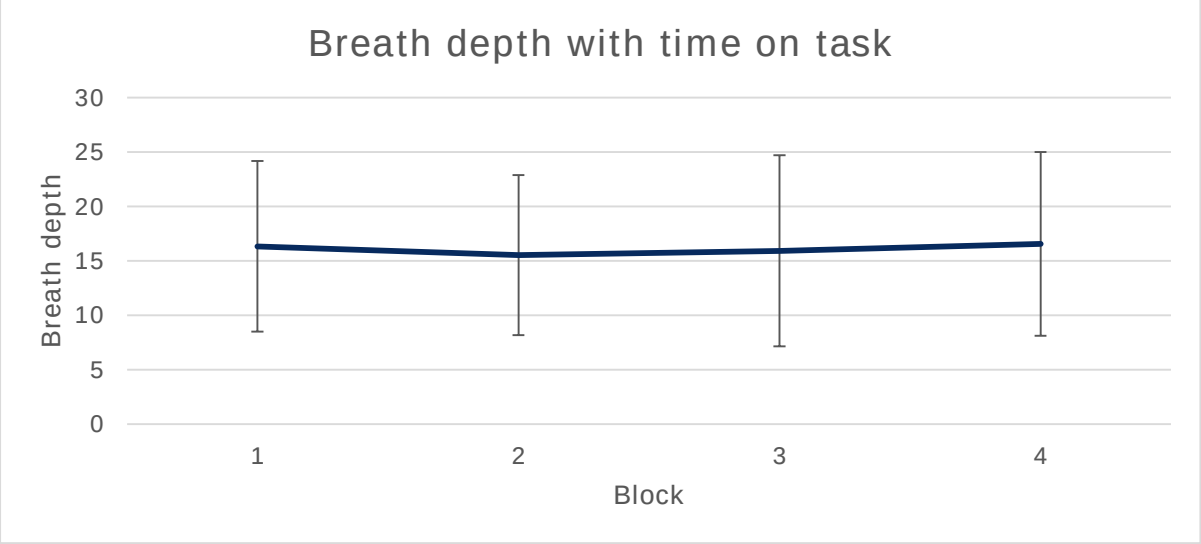


Figure 19- Breath depth with time on task

Figure 19 shows the average breath depth over the four blocks. A one-factor repeated measures ANOVA confirmed that there was not a significant difference between the four blocks ($F(3, 42)=.341$, $p=.796$).

4.3.4 *Physiological measures during four-second epochs of interest*

Participants' physiology in the four second epochs immediately preceding a 'No-Go' stimuli were analysed, to determine whether any physiological changes were present alongside the differences in performance described in Section 4.1.4.

The heart rate variability measures were unsuitable for the four second epoch analyses, as valid HRV metrics cannot be calculated over such short periods of time.

Similarly, the breath depth measure was calculated over the previous 16 seconds of data meaning that it was also not suitable for inclusion in the four second epoch analysis.

4.3.4.1 Heart rate

H14 was that participants' average heart rate during the four seconds before a correct response to a 'No-Go' stimulus would be significantly higher than the mean heart rate four seconds before an incorrect response.

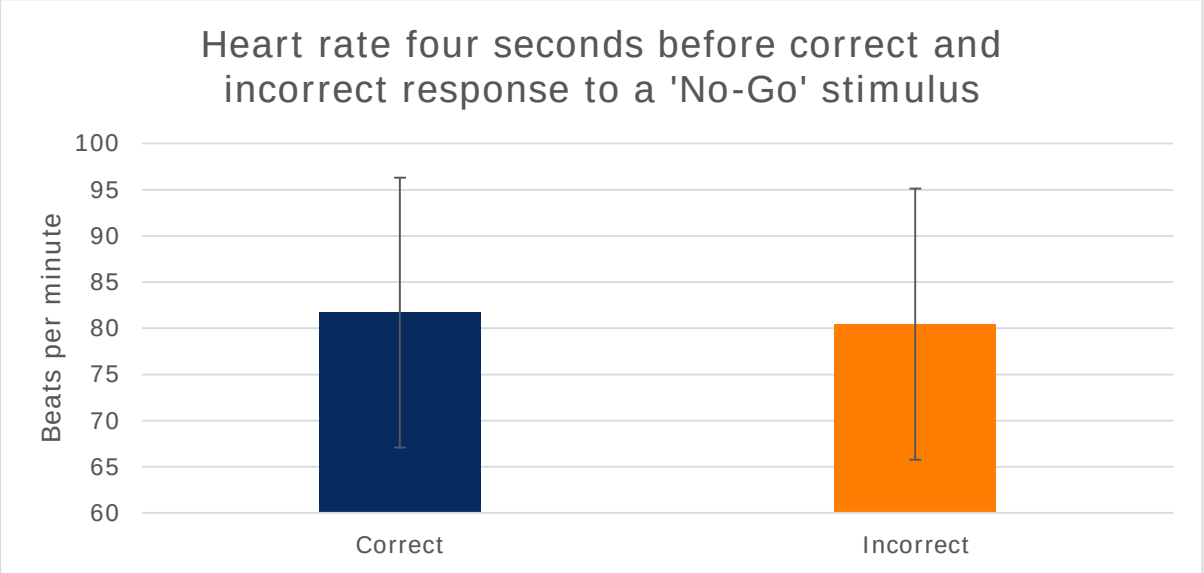


Figure 20- Average heart rate four seconds before correct and incorrect responses to a 'No-Go' stimulus

Figure 20 illustrates that there was a difference in the heart rate four seconds before correct and incorrect responses to a 'No-Go' stimulus, with the average heart rate slightly higher in the correct responses. A paired samples t-test confirmed that the difference between correct and incorrect response was statistically significant ($t(13)=2.282$, $p=.040$).

4.3.4.2 Skin conductance level

H15 was that participants' average SCL during the four seconds before a correct response to a 'No-Go' stimulus would be significantly higher than the average SCL four seconds before an incorrect response.

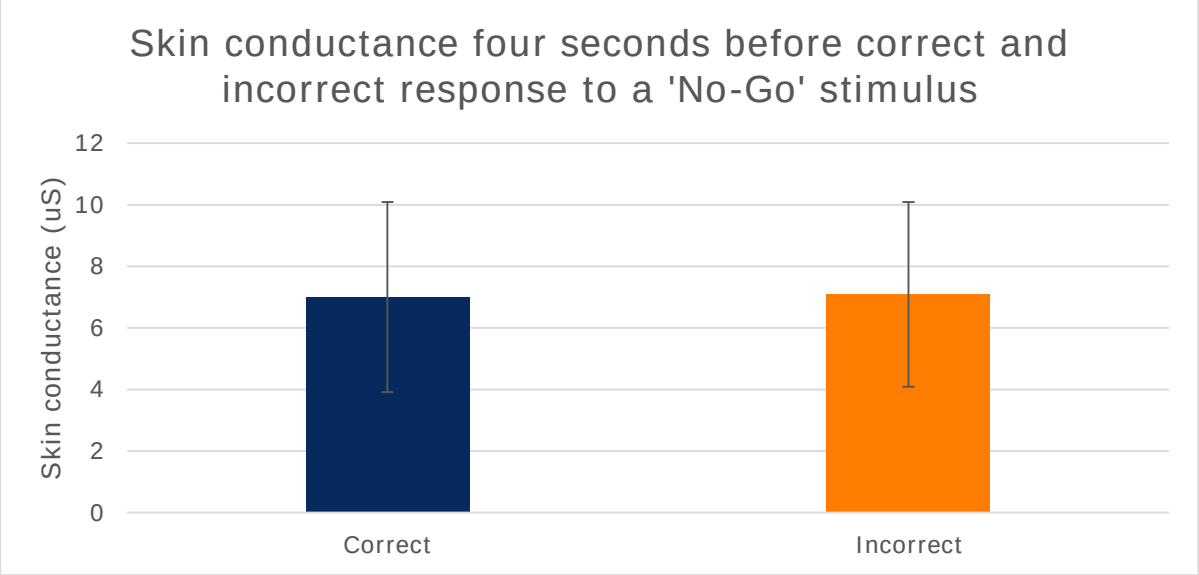


Figure 21- Average skin conductance during the four seconds before correct and incorrect responses to a 'No-Go' stimulus

Figure 21 shows that there was very little difference in the average SCL four seconds before correct and incorrect responses to a 'No-Go' stimulus. A paired samples t-test showed that the difference was not statistically significant ($t(13)=-1.519$, $p=.153$).

H16 was that there would be a reduction in the frequency of SCRs immediately following a commission error with time on task.

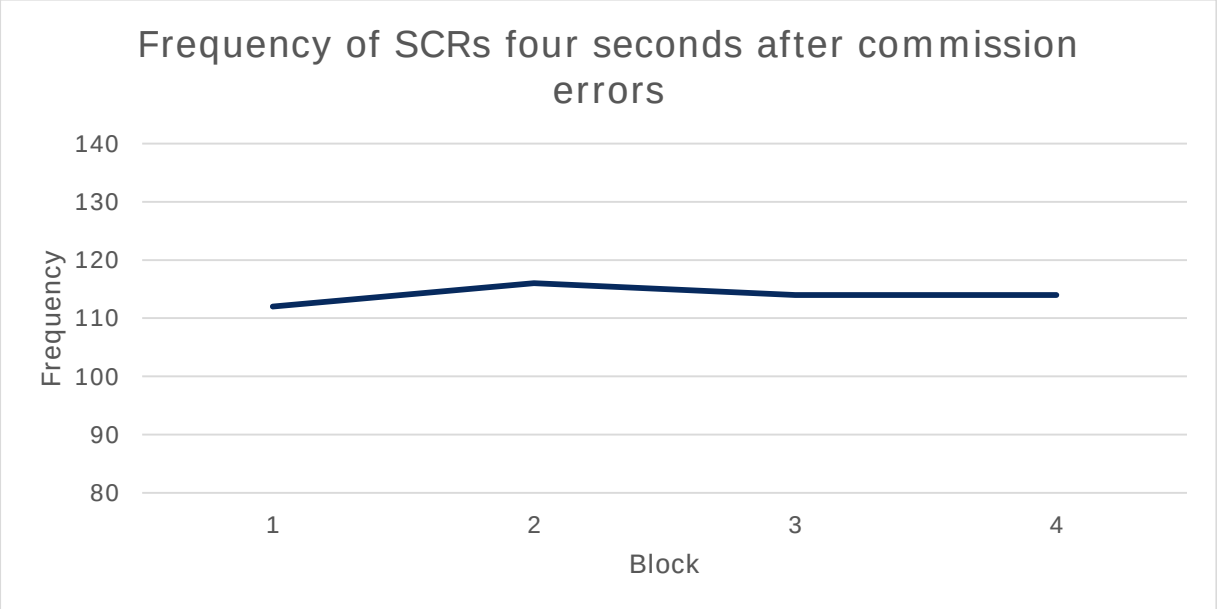


Figure 22- Frequency of SCRs four seconds after commission errors

A skin conductance response (SCR) was calculated as a rise in trace amplitude that began between one and four seconds post-stimulus onset (i.e. after the 'No-Go' event) that exceeded 0.05μS (Dawson et al., 2000). Figure 22 shows that there were few

changes in SCR frequencies immediately after a commission error with time on task. A one-factor repeated measures ANOVA confirmed that there was not a significant difference between the four blocks (F(3, 33)=.687, p=.566).

H17 was that would be a reduction in the SCR amplitude four seconds after a commission error with time on task.

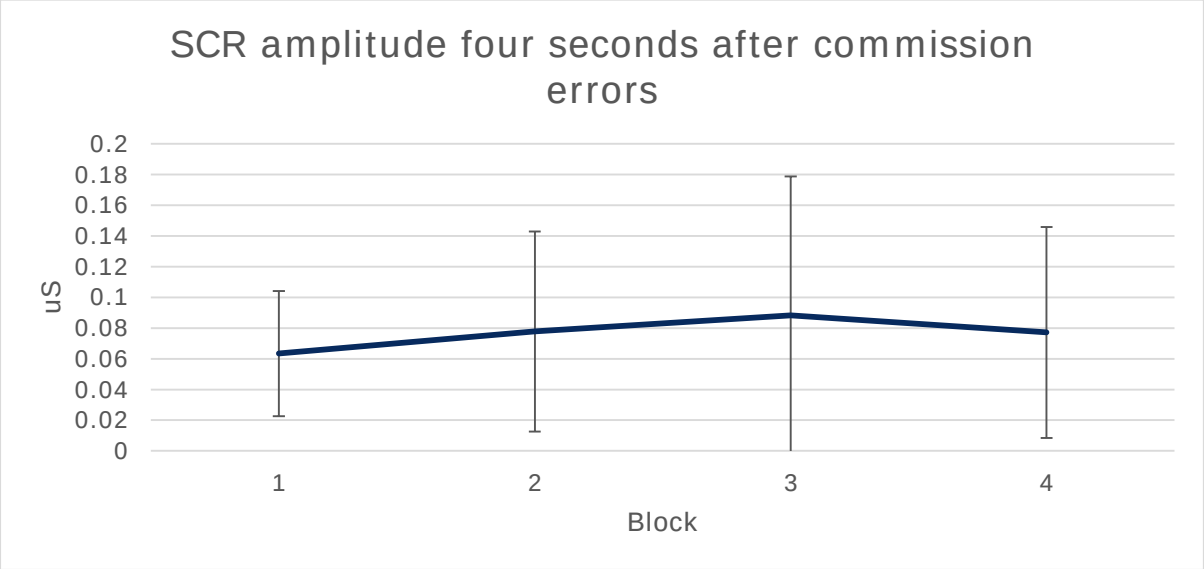


Figure 23- SCR amplitude four seconds after commission errors

Figure 23 shows average SCR amplitudes immediately following a commission error with time on task. A one-factor repeated measures ANOVA confirmed that there was not a significant difference between the four blocks (F(3, 33)=.2.714, p=.061).

4.3.4.3 Respiration rate

H18 was that participants' average respiration rate during the four seconds before a correct response to a 'No-Go' stimulus would be significantly higher than the average respiration rate four seconds before an incorrect response.

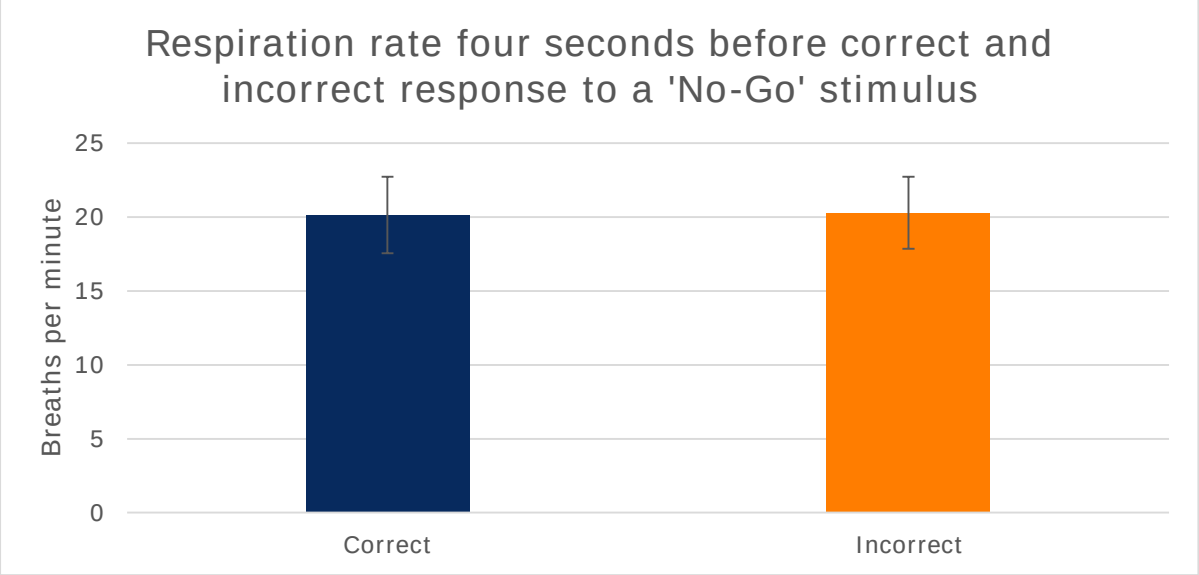


Figure 24- Average respiration rate during the four seconds before correct and incorrect responses to a 'No-Go' stimulus

Figure 24 presents the average respiration rate four seconds before correct and incorrect responses to a 'No-Go' stimulus. A paired samples t-test found that there was not a significant difference between the correct and incorrect responses ($t(14)=-.461$, $p=.652$).

5 Discussion

This study was conducted to understand whether it is possible to establish a set of reliable, objective, non-intrusive and easy-to-administer physiological measures that could detect the presence of cognitive underload in a sample of certified train drivers.

A literature review was conducted to inform the design of subsequent experimental trials. The review found that there were few studies that had used physiological measures in order to investigate cognitive underload, and so literature on high workload was also examined.

The literature review highlighted several potential physiological indicators of cognitive underload, including cardiovascular measures, EEG, eye activity, respiration, electrodermal activity, facial expression, posture and movement. Each had advantages and disadvantages. Strengths and weaknesses of each were compared, taking into account validity, consistency of results from past studies and practical issues to do with administering them.

Cardiovascular, respiration and electrodermal measures were selected for the current study on the basis that some previous studies had found them to be promising, because they were not intrusive and were relatively easy to administer. EEG was considered intrusive and difficult to administer for the current study, and the workload literature also showed that it produced inconsistent results. Eye activity and facial expression were given serious consideration, but equipment that would automate the analysis could not be obtained within the scope of this small study.

The literature review also highlighted the importance of the choice of task that participants would carry out. Previous studies had either used vigilance decrement tasks or tasks with differing levels of difficulty (where difficult was thought to relate to high workload and easy was thought to relate to low workload). Vigilance decrement tasks were seen to provide results which better matched the effects expected in underload conditions. Due to the long, and at times, repetitive nature of train driving tasks, a vigilance task with these characteristics was identified as being the most appropriate type of task for the study. The sustained attention to response task (SART) was selected. This involves the rapid and random presentation of digits (1 to 9) on a computer screen. Participants have to press the spacebar as quickly as possible, each time a digit appears, but they must refrain from pressing the spacebar when the digit ‘3’ appears. Thus the digit ‘3’ is the ‘no-go’ stimulus, whereas all other digits are ‘go’ stimuli. During the current study, the SART was extended to last 17 minutes, as it was thought that the train driver population might be able to maintain their attention for longer periods of time than the general population.

Fifteen participants completed the laptop based SART. Participants were all train drivers from South West Trains and Southeastern. Along with the physiological measures described above, participants’ performance on the SART was monitored, and subjective measures of workload, fatigue and boredom were taken after the task was completed.

The SART was successful in making participants feel more bored as time went on. Participants reported having felt more bored during the second half of the task compared to the first. However participants' average boredom rating in the second half the task was still nearer 'not bored at all' than 'extremely bored'. Dunn (2011) states that boredom is associated with low workload situations and is likely to negatively affect performance. It may be that participants did not get bored enough to truly experience cognitive underload while completing the SART.

Participants' sleepiness levels were measured using the Karolinska Sleepiness Scale (KSS). Mean sleepiness levels did not change from the first half of the task to the second. This is a positive outcome for the study as it suggests that the SART increased boredom, but not sleepiness, which would have been a confounding factor.

It was predicted that participants' task performance would deteriorate over time. The results did not show this to be the case. Neither the number of errors made by participants, their reaction times nor the variability of their reaction times seemed to change over time. Surprisingly, the performance sub-scale of the NASA-TLX showed a significant increase in perceived task performance from the first half of the task to the second half. Contrary to our hypothesis, this suggests participants thought their performance improved in the second half. This subjective assessment of performance is not supported by objective measures of task performance, which do not show any improvement. This suggests that participants' impressions of their performance did not match actual performance. Further investigation is needed to understand why this might be the case. One possible explanation is that participants thought their performance got better, or the task got easier, with practice. A learning effect could also have masked performance decrements due to underload; a practice

e effect could have caused improvements in performance as time went on, while underload could have resulted in performance decrements as time passed. These two effects may have combined to result in the overall picture which shows no change in task performance with time. To avoid a possible learning effect, the 20 second practice session prior to starting the task could be extended to allow participants to get used to the SART in future studies. None of the other NASA-TLX factors showed significant differences.

Physiological measures were predicted to change over time (a reduction in heart rate, breath depth respiration rate and skin conductance was expected, alongside an increase in heart rate variability), in line with task performance measures. The results show that they did not change. This is perhaps unsurprising in light of the lack of change in task performance over time.

It is not entirely clear why time-on task effects were not found for performance and physiological measures in this study, but possible interpretations are provided towards the end of this section.

The data were explored to understand whether participants' reaction times changed immediately before they made an error. Reaction times to the four 'go' stimuli before a 'no-go' stimulus were found to be faster before an incorrect response to a 'no-go'

stimulus than before a correct response. These differences were consistent throughout the trials and statistically significant. Previous research by Robertson et al (1997) found similar results and they suggest that this difference reflects a reduction in active attention to the task. It is likely that the faster reaction times before an incorrect response are caused by participants 'drifting' into an automatic response mode. Robertson et al (1997) argue that 'local fluctuations in attention or "lapses" may provide a better account of poor performance on [the SART] task than a simple decrement over time'. This interpretation is consistent with a comment provided by one of the train drivers who took part in the study: 'being in the same position and carry out [*sic.*] the same task caused lapses in concentration levels, putting one into autopilot mode'.

Analysis was carried out to investigate any patterns in the physiological measures in line with the task performance measure investigated above. The analysis of task performance looked at 4 presentations before these no-go stimuli. This approximates to a 4-second time period, and so the analysis of physiological measures also focused on these four-second periods of time immediately before a no-go stimulus.

Interestingly, mean heart rate was significantly lower in the four seconds before participants made an incorrect response to a no-go stimulus, compared with a correct response. This lowered heart rate is consistent with findings from previous studies on vigilance by Schmidt et al (2009) and Jap et al (2009) who both reported a decrease in heart rate to be associated with lowered attention. This further supports the idea that the drivers were entering an automated response mode. The same effect was not found for the SCL or average breath rates. These measures may not have been sensitive enough to be detected in the four-second periods that we analysed; for example SCL changes fairly slowly over time (Braithwaite et al, 2015).

It was also predicted that the number of skin conductance responses might reduce over time, if participants became less-aware of or less-sensitive to the errors that they were making. However, no change over time was found in the number of skin conductance responses (SCR) during the four seconds following an error. This suggests that the participants may have been just as aroused when making an error towards the end of SART as they were at the start.

The study did not find task performance to get worse as time went on, so it is important to critique the method. Previous research by Larue et al (2009) reported that performance impairments can be observed in less than five minutes of the SART. Robertson and Garavan (2004) also suggest that errors on the SART are a reasonable model for what happens when a train driver passes through a signal at danger (failure to react to a no-go stimulus).

A key question is whether the task was too easy for the group of train drives who took part in the study. A previous study by Robinson et al (2015) used simulation and found that some drivers were detecting and responding to all of the stimuli they needed to respond to, and this meant that there were no differences in performance to analyse. In the current study, however, accuracy of responses to the no-go stimulus was only 67%. None of the participants had near-perfect performance. In addition, there was clear

evidence of a fluctuation in active monitoring of the task. This suggests that the task was difficult enough, and an improvement to that used by Robinson, et al.

Robinson et al. also questioned the accuracy of the measures that were taken in the training simulator which they used. The PC-based SART was run using E-Prime software which is highly accurate research software (Stahl, 2006), and so measurement issues are not likely to explain these results.

Another consideration is whether train drivers tend to be more resilient to underload inducing situations than the general population. Train drivers go through a selection process involving psychometric tests which aim to identify candidates with skills such as maintaining high levels of concentration and performing well in low workload tasks. Effectively, train drivers might perform better than the average person on the SART or may be better equipped to deal with repetitive conditions due to the nature of their job and the fact they are regularly exposed to repetitive environments. This may explain why performance decrements were seen in other studies but not in this one. This potential issue was identified at the beginning of the study and was the reason why the SART was extended to 17 minutes. This is four times the original length of the task. Nevertheless, the expected performance decrements were still not observed.

It may be that performance did not deteriorate over time because participants were able to overcome their feelings of boredom and to motivate themselves sufficiently to maintain their concentration levels for the duration of the task. Dunn (2011) and Hockey (1997) highlight motivation and effort as being a potential buffer against cognitive underload.

These methodological questions have significant implications for further studies on cognitive underload with train drivers. Where short measurement periods are used, at least, it may be that low workload produces fluctuations in performance, as suggested by Robertson et al (1997), rather than a decrease in performance over time. In studies where physiological measures are being taken, the results described here suggest the need for a different approach to experimental design, with sufficient time between target stimuli to allow physiological changes to occur between two consecutive measurements, and to prevent the measurements from being affected by multiple rapidly occurring stimuli. In order to overcome the potential effects of short term motivation or effort resulting from participation in an experiment, it may be that significantly longer measurement periods and naturalistic study designs could give a better indication of whether time on task effects do occur with underload. This may become more viable as measurement techniques and technologies develop and become feasible to implement outside of laboratory conditions, and in real-world operational environments.

This study did show that participants' performance fluctuated over the course of the trial, and the results suggest that they drifted into an automatic response mode, with their heart rate reducing and responses speeding up, just before they made errors in responding to target stimuli in the SART. Thus, the results provide an insight into the relationship between task performance and physiology during low workload conditions, as well as highlighting experimental design considerations for other work in this area.

6 Conclusions

This study has found that train drivers' performance fluctuated during a PC-based, 17-minute repetitive task, which was selected to introduce performance challenges representing those encountered in train driving. Immediately before participants made an error in responding to a target stimulus, their response times to the repetitive stimuli shortened and their heart rate decreased. When they responded correctly, they did so more slowly and their heart rate was higher. This suggests that the repetitive task resulted in participants drifting into a more automated mode, leading to lapses in attention. Although previous studies have identified these effects to be associated with underload, this is the first study to report this relationship between reaction times and heart rate at the same time.

This study did not find a deterioration in task performance or a change in physiology as participants progressed through the task. There are a number of potential reasons for this. It is possible that the train drivers who participated in the study were more resilient to boredom and repetition than the general population. A further study involving a control group of non-train drivers may be useful to test this hypothesis. There may have also been a practice effect which could have masked any potential performance decrements. A further study with an extended practice session may help to reduce any practice effect and allow decrements to be identified. On the other hand, the SART may have been too 'stimulating' for the train drivers, meaning they never experienced underload. They may have been motivated, or put in additional effort, to maintain performance for the duration of the task. Future work should, if possible, consider a vigilance task with fewer target stimuli, and should monitor driver performance over a longer time period. Alternatively, naturalistic studies could be conducted, in which the effects of being observed are minimised and measurement can take place over very long periods of time.

The performance and heart rate effects immediately preceding a target stimulus are promising. These results are a building block for identifying when a train driver is experiencing underload. Further research is needed before an operational solution can be developed, as heart rate on its own cannot be relied upon to make an assessment of workload. Heart rate is affected by a range of factors including caffeine intake, air temperature and body movement. Nevertheless future research and development activities could build on this work to identify and test different physiological measures which could be used, in combination with the heart rate measure, to detect potential instances of underload. Such studies would be a further step towards the ultimate aim of developing an operational solution to indicate when to apply an appropriate mitigation and reduce errors associated with underload. Developments in technology are likely to have a positive impact on this research area. As measures such as sophisticated camera-based physiological acquisition techniques become viable, research should be carried out

to explore whether they can be used, potentially in combination with heart rate, to detect underload.

Acknowledgements

The authors of this study would like to thank Southeastern and South West Trains for supporting the study, and all of the train drivers who took part. We would like to thank colleagues at Wuppertal University and Anglia Ruskin University who provided the physiological monitoring equipment that was needed, as well as technical support. Finally, we would like to acknowledge the efforts of Huw Gibson and Prof Andrew Parkes who carried out the technical review of this report.

7 References

Advisory Group for Aerospace Research and Development. (1989). Human performance assessment methods (AGARDograph-308). Neuilly-sur-Seine, France: North Atlantic Treaty Organization, Author.

Ahlstrom, U., Ferne J., Friedman-Berg, F.J. (2006). Using eye movement activity as a correlate of cognitive workload. International Journal of Industrial Ergonomics. 1 (36), 623-636

Backs, R.W., Seljos, K.A., (1994). Metabolic and cardiorespiratory measures of mental effort: the effects of level of difficulty in a working memory task. Int. J. Psychophysiology. 16, 57–68.

Backs. R.W., Navidzadeh, T., Xu, X. (2000). Cardiorespiratory Indices of Mental Workload during Simulated Air Traffic Control, Human Factors and Ergonomics Society Annual Meeting Proceedings. 13: 89-92.

Baddeley A. (1968). A 3-min reasoning task based on grammatical transformation. Psychometric Science. 10:341–2.

Balaban, C. D., Cohn, J., Redfern, M., Prinkey, J., Stripling, R., & Hoffer, M. (2004). Postural control as a probe for cognitive state: Exploiting human information processing. International Journal of Human-Computer Interaction. 17, 275–287.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and structure of processing resources, Psychological Bulletin, 91: 276–292.

Bechtereva, N. P. (1981). Neurophysiological correlates of mental process in man, Psychophysiology today and tomorrow. 11–21.

Belyavin, A., Wright, N. A. (1987). Changes in electrical activity of the brain with vigilance. Electroencephalography and Clinical Neurophysiology. 66 (2), 137-144.

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., Montanari, R. (2011). Driver workload and eye blink duration, Transportation Research Part F-traffic Psychology And Behaviour. 14(3), 199-208.

Boucsein, W. (1993). "Psychophysiology in the workplace- Goals and Methods." Psychophysiology of Mental Workload. 35-41.

Bonner, M. A., Wilson, G.F. (2002). Heart rate measures of flight test and evaluation. International Journal of Aviation Psychology. 12:63-77.

Burgess PW, Shallice T. (1996). Response suppression, initiation and strategy use following frontal lobe lesions. Neuropsychologia. 34:263–73.

Braby, C.D., Harris, D., Muir, H.C. (1993). A psychophysiological approach to the assessment of work underload. Ergonomics . 36 (9), 1035-1042.

Braithwaite, J.J., Watson, D.G., Jones, R., Rowe M. (2015). A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments. Revised version 2.0.

Brookhuis, K., de Waard, D. (2010). Monitoring drivers' mental workload in driving simulators using physiological measures. Accident Analysis and Prevention. 42 (3), 896-903.

Brookings, J.B., Wilson, G.F., Swain, C.R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. Biological Psychology. 42 (3), 361-377.

Brown, M., Marmor, M. Vaegan. (2006). ISCEV Standard for Clinical Electro-oculography (EOG), Documenta Ophthalmologica. 113:3. 205-212.

Campagne, A., Pebayle, T., Muzet, A. (2005). Oculomotor changes due to road events during prolonged monotonous simulated driving. Biological Psychology. 68, 353-368.

Capa, R.L., Audiffren, M., and Ragot, S. (2008). The interactive effect of achievement motivation and task difficulty on mental effort. International Journal of Psychophysiology, 70, 144-150.

Casali, J.G., Wierwille, W.W. (1984). On the measurement of pilot perceptual workload: A comparison of assessment techniques addressing sensitivity and intrusion issues. Ergonomics. 27:1030-1050.

Chen, S.Y., Epps, J. (2013). Automatic classification of eye activity for cognitive load measurement with emotion interference, Computer Methods And Programs In Biomedicine. 110(2), 111-124.

Cinaz, B., Arnrich, B., Marca, R., Troster G. (2011). Monitoring of mental workload levels during an everyday life office-work scenario. Personal and Ubiquitous Computing. 17(2), 229-239.

Collet, C., Salvia, E., Petit-Boulanger, C. (2014). Measuring workload with electrodermal activity during common braking actions, Ergonomics.57(6), 886-896.

Comstock, J. R., & Arnegard, R. J. (1992). The Multi-Attribute Task Battery for human operator workload and strategic behavior research. (NASA Tech. Memorandum No. 104174). Hampton, VA: NASA.

Dawson, M., Schell, A., Filion, D. (2000). The electrodermal system. In Cacioppo, J., Tassinary, L. & Berntson, G. (Eds.) Handbook of psychophysiology, 2nd ed. (pp. 200-223). Cambridge University Press: New York.

Dijksterhuis, C., Brookhuis, K.A., De Waard, D. (2011). Effects of steering demand on lane keeping behaviour, self-reports, and physiology. A simulator study. Accident Analysis and Prevention. 43, 1074-1081.

Dillard et al., (2014). The sustained attention to response task (SART) does not promote mindlessness during vigilance performance. Human Factors, 56(8), pp. 1364-1379.
Dinges D.F., Kribbs N.B., Bates B.L., Carlin M.M. (1993). A very brief probed recall memory task: Sensitivity to sleep loss. Sleep Research. 22:330.

Dinges D.F., Orne E.C., Evans F.J., Orne M.T. (1981). Performance after naps in sleep conducive and alerting environments, Biological rhythms, sleep and shift work. Advances in sleep research. New York: Spectrum Publications. 7:539 –52.

Dinges D.F., Powell J.W. (1985). Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. Behavior research methods instruments & computers impact factor. 17:652–5.

Dinges, D.F., Rider, R.L., Dorrian, J., McGlinchey, E.L., Rogers, N.L., Cizman, Z., Goldenstein, SK., Vogler, C., Venkataraman, S., Metaxas, D.N. (2005). Optical computer recognition of facial expressions associated with stress induced by performance demands, Aviation Space And Environmental Medicine. 76(6), B172-B182.

Dunn, N. J. (2011). Monotony: The effect of task demand on subjective experience and performance, School of Risk and Safety Sciences, University of Wales.

Ekman, P. and Friesen, W.V. (1978). Facial action coding system. Palo Alto, CA: Consulting Psychologists Press.

Elsmore T.F. (1994). Synwork1: a PC-based tool for assessment of performance in a simulated work environment. Behavioural Respiration Methods. 26:421– 6.

Feyer, R. G. (2007). Bridging the gap: Exploring interactions between digital human models and cognitive models. In V. G. Duffy (Ed.), Digital human modelling. Berlin, Germany: Springer-Verlag.

Fisch, B.J., (1991). Spehlmann's EEG Primer, Second revised and enlarged edition. Elsevier Science BV, Amsterdam, The Netherlands.

Frank, G. (2006). Monitoring seated postural responses to assess cognitive state (Doctoral dissertation). University of Pittsburgh, Pittsburgh, PA.

Gardner, E. (1975). Fundamentals of neurology. Philadelphia: Saunders.

Gevins, A., Smith, ME., Leong, H., McEvoy, L., Whitfield, S., Du, R., Rush, G. (1998). Monitoring Working Memory Load during Computer-Based Tasks with EEG Pattern Recognition Methods. Human Factors: The Journal of the Human Factors and Ergonomics Society. 40 (1), 79-91.

Gevins, A., Smith, M.E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. Theoretical Issues in Ergonomics Science. 4:113–131.

Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. Cerebral Cortex, 7, 374–385.

Gimeno, P.T., Cerezuela, G.P., Montanes, M.C. (2006). On the concept and measurement of driver drowsiness, fatigue and inattention: Implications for countermeasures. International Journal of Vehicle Design, 42, 67-86.

Gundel, A., & Wilson, G. F. (1992). Topographical changes in the ongoing EEG related to the difficulty of mental tasks. Brain Topography, 5, 17–25.

Graf, M., Guggenbuhl, U., Krueger, H. (1995). An assessment of seated activity and postures at five workplaces. International Journal of Industrial Ergonomics, 15, 81–90.

Grandjean, E. (1988). Fitting the Task to the Man. London: Taylor and Francis

Grier, R.A., Warm, J.S., Dember, W.N., Matthews, G., Galinsky, T.L., Szalma, J.L., Parasuraman, R. (2003). The Vigilance Decrement Reflects Limitations in Effortful Attention, Not Mindlessness. Human Factors, 45: 349-359.

Haider, E., Rohmert,W., (1976). Blink frequency during four hours of simulated truck driving. European Journal of Applied Psychology. 35, 137–147.

Harris, Randall L., Alan T., Comstock, J., Raymond, Jr. (1988). Physiological assessment of task underload, 2nd Annual Workshop on Space Operations Automation and Robotics. 221-226.

Hitchcock, E., Warm, J. S., Matthews, G., Dember, W. N., Shear, P. K., Tripp, L., Parasuraman, R. (2003). Automation cueing modulates cerebral blood flow and vigilance in a simulated air traffic control task. Theoretical Issues in Ergonomics Science, 4, 89–112.

Hockey, G.R.J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. Biological Psychology. 45, 73-93.

Hockey, G. R. J., Nickel, P., Roberts, A. C., & Roberts, M. H. (2009). Sensitivity of candidate markers of psychophysiological strain to cyclical changes in manual control load during simulated process control. Applied Ergonomics, 40: 1011–1018.

Iqbal, Z, M. Lateef, M. Ashraf, and A. Jabbar (2004). Anthelmintic activity of Artemisia brevifolia in sheep. J Ethnopharmacol 93: 265–268.

Jagannath, M., Balasubramanian, Venkatesh. (2014). Assessment of early onset of driver fatigue using multimodal fatigue measures in a static simulator, Applied Ergonomics. 45, 1140-1147.

Jap, B.T., Lal, S., Fischer, P., Bekiaris, E. (2009). Using EEG spectral components to assess algorithms for detecting fatigue. Expert Systems with Applications. 1 (36), 2352-2359.

Kamzanova, A., Kustubayeva, A., Matthews, G. (2014). Use of EEG Workload Indices for Diagnostic Monitoring of Vigilance Decrement. The Journal of the Human Factors and Ergonomics Society, 56:1136-1149.

Karavidas, M.K., Lehrer, P.M., Lu, S.E., Vaschillo, E., Vaschillo, B., Cheng, A. (2010). The effects of workload on respiratory variables in simulated flight: A preliminary study, Biological Psychology. 84(1): 157-160.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. Journal of Experimental Psychology. 55:352-358.

Kobayashi, S., Hara, T., Goi, Y. (1996). Predicting changes in driver's drowsiness level during prolonged driving: extracting long blinks from EOG data. JSAE Review. 17, 435–458.

Kumashiro, M. (2005). Evaluation of Human Work. Boca Raton: Taylor and Francis group. 21:618.

Lacey, J.I., Lacey, B.C. (1978). Two-way communication between the heart and the brain: Significance of time within the cardiac cycle. Research in the Psychobiology of Human Behaviour. Baltimore: John Hopkins University Press. 99-113.

Lal, S.K.L., Crag, A. (2001). A critical review of the psychophysiology of driver fatigue. Journal of Sleep Research, 8, 255-262.

Lang P.J., Bradley, M.M., Cuthbert, B.N. (2008). International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual, Technical Report A-8, University of Florida.

Larue, G.S., Rakotonirainy, A., Pettitt, A.N. (2009). A model to predict hypovigilance during a monotonous task. Proceedings of the 2009 Australasian Road Safety Research, Policing and Education Conference: Smarter, Safer Directions, 10-12 November 2009, Sydney Convention and Exhibition Centre, Sydney, New South Wales.

Larue, G.S., Rakotonirainy, A., Pettitt, A.N. (2010). Real-time performance modelling of a sustained attention to response task. Ergonomics, 53: 1205-1216.

Larue, G.S., Rakotonirainy, A., Pettitt, A.N. (2011). Driving performance impairments due to hypovigilance on monotonous roads, Accident Analysis and Prevention. 43(6), 2037-2046.

Lecret, F., & Pottier, M. (1971). La vigilance, facteur de securite dans la conduite automobile. Le Travail Humain, 34, 51–68.

Lei, S.G., Roetting, M. (2011). Influence of Task Combination on EEG Spectrum Modulation for Driver Workload Estimation. Human Factors. 53 (2), 168-179

Leino, K., Nunes, S., Valta, P., Takala. J. (2001). Validation of a new respiratory inductive plethysmograph, Acta Anaesthesiologica Scandinavica, 45: 104–111.

Louhevaara, V., Kilborn, A. (2005). Evaluation of Human Work. Boca Raton: Taylor and Francis group. 15:445-446.

Manly, T., Robertson, I., Galloway, M., Hawkins, K. (1999). The absent mind: further investigations of sustained attention to response. Neuropsychologia, 37: 661-670.

Mattes, S. (2003). The lane change task as a tool for driver distraction evaluation. In H. Strasser, H. Rausch, & H. Bubb (Eds.), Quality of work and products in enterprises of the future (pp. 57–60). Stuttgart: Ergonomia Verlag.

Megaw, T. (2005). Evaluation of Human Work. Boca Raton: Taylor and Francis group. 18:538.

Mehler, B., B. Reimer, and J. F. Coughlin. (2012). Sensitivity of Physiological Measures for Detecting Systematic Variations in Cognitive Demand from a Working Memory Task: An On-Road Study Across Three Age Groups. Human Factors 54 (3): 396–412.

Mehler, B., Reimer, B., Coughlin, J.F., Dusket, J.A. (2009). The impact of incremental increases in cognitive load on physiological arousal and performance in young adult drivers, Transportation Research Record: The Journal of the Transportation Research Board. 2138: 6-12.

Mulder, L. J. M. (1992). Measurement and analysis methods of heart rate and respiration for use in applied environments. Biological Psychology, 34, 205–236.

Murata, A. (2005). An Attempt to Evaluate Mental Workload Using Wavelet Transform of EEG. Human Factors: The Journal of the Human Factors and Ergonomics Society. 47 (3), 498-508.

Nickel, P. & Nachreiner, F. (2003). Sensitivity and diagnosticity of the 0.1Hz component of heart rate variability as an indicator of mental workload. Human Factors, 45(4), pp. 575-590.

Pfaff, U., Fruhstorfer, H., Peter, J.H., (1976). Changes in eyeblink duration and frequency during car driving. Pflueger Archive 362, R21.

Piquado, T., Isaacowitz, D., Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults, Psychophysiology. 47: 560-569

Poh, M., Swenson, N.C., Picard, R.W. (2010). A Wearable Sensor for Unobtrusive, Long-Term Assessment of Electrodermal Activity. IEEE Transactions On Biomedical Engineering. 57(5): 1243-1252

Recarte, M.A., Pérez, E. Conchillo, A. Nunes L.M. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating, The Spanish Journal of Psychology, 11, 374–385.

Reyes del Paso, G.A., Gonzalez, I., Hernandez, J.A., (2004). Baroreceptor sensitivity and effectiveness varies differentially as a function of cognitive-attentional demands. Biol. Psychol. 67, 385–395.

Richter, P., Wagner, T., Heger, R., Weise, G. (1998). Psychophysiological analysis of mental load during driving on rural roads - A quasi-experimental field study, Ergonomics. 41(5), 593-609.

Robinson, D., Waters, S., Basacik, D., Whitmore, A., Reed, N. (2015). A pilot study of low workload in train drivers.

Robertson, I., Garavan, H. (2004). Vigilant attention, in M Gazzaniga (eds). The cognitive neurosciences. 632-633.

Robertson, I., Manly, T., Andrade, J., Baddeley, B., Yiend, J. (1997). 'Oops': Performance correlates of everyday attentional failures in traumatic brain injured and normal subjects. Neuropsychology. 24(6), 636-647.

Roge, J., Pebayle, T., & Muzet, A. (2001). Variations of the level of vigilance and of behavioral activities during simulated automobile driving. Accident Analysis & Prevention, 33, 181–186.

Rognin, L., Hébraud, C., Hoffman, E., Pène, N., Zeghal, K. (2004). Assessing the impact of a new air traffic control instruction on flight crew activity. AIAA Guidance, Navigation, and Control Conference and Exhibit. 16 - 19 August 2004, Providence, Rhode Island

Qiu, J., Helbig, R. (2012). Body Posture as an Indicator of Workload in Mental Work, Human Factors. 54, 626-635.

Salvucci, D. D., Goldberg, J.H. (2000). Identifying fixations and saccades in eye-tracking protocols, in: Proceedings of the 2000 Symposium on Eye Tracking Research and Applications, New York, pp. 71–78.

Sammer, G. (1998). Heart period variability and respiratory changes associated with physical and mental load: non-linear analysis, Ergonomics, 41: 746–755.

Schell, A.M., Dawson, M.E., Nuechterlein, K.H., Subotnik, K.L., Ventura, J. (2002). The temporal stability of electrodermal variables over a one-year period in patients with recent-onset schizophrenia and in normal subjects. Psychophysiology. 39 (2), 124–132.

Schmidt, E.A., Schrauf, M., Simon, M., Fritzsche, M., Buchner, A., Kincses, W.E. (2009). Drivers' misjudgement of vigilance state during prolonged monotonous daytime driving. Accident Analysis and Prevention. 1 (41), 1097-1091.

Smilek, D., Carriere, J., Cheyne, A. (2010). Failures of sustained attention in life, lab, and brain: Ecological validity of the SART. Neuropsychologia, 48: 2564-2570.

Stahl, C. (2006). Software for generating psychological experiments. Experimental Psychology, 53(3), 218.

Stanton, N., Neville A. (2002). Malleable Attentional Resources Theory: A New Explanation for the Effects of Mental Underload on Performance. The journal of Human Factors and Ergonomics Society. (44), 365-375.

Stanton, N., Hedge, A., Brookhuis, K.A., Salas, E., Hendrick, H.W. (2004). Handbook of Human Factors and Ergonomics Methods. CRC Press, London.

Stern, D., Boyer, D., Schroeder. (1994). Blink rate: A possible measure of fatigue Human Factors, 36, 285–297.

Sternberg, S. (1966). High-speed scanning in human memory. Science, 153, 652-654.

Sternberg, S. (1969). Memory-scanning: mental processes revealed by reaction-time experiments. Psychologist. 57, 421-457.

Stone, R.T., Wei, C.S. (2011). Exploring the linkage between facial expression and mental workload for arithmetic tasks, Proceedings of the Human Factors and Ergonomics Society. 616-619

Stroop JR. (1935). Studies of interference in serial verbal reactions. Journal of Experimental Psychology.18:643– 62.

Stuiver, A., Brookhuis, K., de Waard, D., Mulder, B. (2014). Short-term cardiovascular measures for driver support: Increasing sensitivity for detecting changes in mental workload, International Journal of Psychophysiology. 92:35-41.

Taylor G.J, Ryan D, Bagby RM. (1985). Toward the development of a new self-report alexithymia scale. Psychother Psychosom; 44: 191–9.

Van Orden K.F., Jung T.P., Makeig S. (2000). Combined eye activity measures accurately estimate changes in sustained visual task performance. Biological Psychology. 52 (3): 221-240.

Van Roon, A.M., Mulder, L.J.M., Althaus, M., Mulder, G. (2004). Introducing a baroreflex model for studying cardiovascular effects of mental workload. Psychophysiology. 41, 961–981.

Veldhuizen, I. J. T., Gaillard, T., De Vries, A.W.K.J. (2003). The influence of mental fatigue on facial EMG activity during a simulated workday. Biological Psychology, 63, 59-78.

Veltman, J. A., & Gaillard, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. Ergonomics, 41(5), 656–669.

Wallin, B.G., Fagius, J. (1986). The Sympathetic Nervous System in man: Aspects Derived from Microelectrode Recordings. Trends in Neurosciences. 9: 63-67.

Wientjes, C.J., (1992). Respiration in psychophysiology: methods and applications. Biol.
Psychol. 34, 179–203.

Wilson, G., Russell, C. (2003). Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. Human Factors: The Journal of the Human Factors and Ergonomics Society. 45 (4), 635-643.

Yamakoshi, T., Rolfe, P., Yamakoshi, Y., Hirose, H. (2009). A novel physiological index for Driver's Activation State derived from simulated monotonous driving studies, Transportation Research Part C: Emerging Technologies. 17(1), 69-80.

Young, M.S., Stanton, N.A. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance, Human Factors. 44(3), 365-375.

Appendix A Literature review

A.1 Electrical brain activity

A.1.1 Introduction

Electroencephalogram (EEG) is the recording of the brain's electrical activity. An electroencephalograph is a device which measures this activity and consists of two parts; an amplifier and a recording device. Electrical activity generated by the brain is amplified and then recorded by the recording mechanism (Kumashiro, 2005). Activity is usually recorded by multiple scalp electrodes and classified according to rhythms defined by the frequency of delta, theta, alpha and beta bands (Fisch, 1991). Past research has shown that delta waves are associated with the transition from drowsiness to sleep with theta waves indicating developing drowsiness (Lal and Craig, 2001; Brookhuis and De Ward, 2010). These slower waves are thought to be associated with underload and should be present in underload conditions. However, Gevins et al (1997) have reported that an increase in theta activity is associated with sustained mental effort or concentration. There seems to be some uncertainty regarding the relationship between low workload and theta activity. Alpha waves indicate wakefulness; however they also occur when in a relaxed state and reduced readiness to respond (Grandjean, 1988). A high amount of beta wave activity indicates that a person is generally awake and alert (Lal and Craig, 2001, Brookhuis and De Ward, 2010).

There seems to be two main methods for analysing the frequency of the different brain waves. The first of these methods is the waveform recognition method where brainwave data are drawn on a chart and visually measured and recorded according to a scale by experienced human experimenters. Waveforms are then converted to frequencies and organised according to rate of occurrence. The second method requires a power spectrum for analysing frequencies from measured data using a analysis by a computer. Measured data are broken down into sine waves and expressed in terms of their "power

spectrum" at each frequency band (Kumashiro, 2005) Brainwave data in graphical form can be analysed using the waveform recognition method where waveforms are identified and recorded by experienced human experimenters or by a method which requires.

A.1.2 Tasks performed by participants in EEG studies

Gevins et al (1998) and Murata (2005) both used continuous matching tasks in their study requiring participants to indicate whether a current stimulus matched a stimulus from a previous trial. There was both Each study employed tasks with a spatial and verbal aspect to the task components. Gevins et al (1998) asked participants to complete these two variations separately, whereas Murata's (2005) participants completed both at once meaning they had to remember both the identity and location of the stimulus. Each study had varying levels of task difficulty, including easy, lower workload conditions through to hard, higher workload conditions meaning comparisons between these could be made.

Lei and Roetting (2011) used a driving simulator task, measuring brain waves at base, low and high mental workload levels. To stimulate brain activity when driving, a similar matching task to Gevins et al (1998) and Murata (2005) was used. The N-back task, first introduced by Kirchner (1958) consists of indicating whether a current stimulus matches one from n steps back. The value of N can therefore be adjusted to make the task more or less difficult. Lei and Roetting (2011) used this technique for 1-back and 2-back tasks to simulate low and high mental workload respectively.

Wilson and Russell (2003) also looked at difficulty levels; however, they used the NASA Multi Attribute Task Battery (Comstock & Arnegard, 1992) at base, low and high difficulty levels. Task difficulties were manipulated by the number of events which occurred in 5 minute trials.

For vigilance tasks Kamzanova et al (2014) asked participants to perform a slightly modified version of the Hitchcock et al (2003) task. Participants were asked to monitor an 'air traffic control display' consisting of a 'city' bound by three circular boundaries and two 'aircraft' represented by grey lines. Aircraft approached the 'city' from opposite directions and participants were required to click when the two grey lines of the aircraft aligned; this happened very infrequently. Brainwaves from both cue and non-cue conditions were measured in the study. The cue was the word 'LOOK' and participants were told that the target would follow the cue in the next 5 stimuli.

Harris et al (1988) used a fault acknowledgment task to induce boredom and underload. The participant's task was to respond to a highlighted fault area on a desk top computer generated jet engine pictorial.

A.1.3 Findings

There have been many studies that have used EEG as the main physiological measure of workload, with several past studies finding it to be highly sensitive to changes in task difficulty (Gevins, et al, 1998; Gundel & Wilson, 1992). These studies differ both in

terms of the number of electrodes used on their participants and the task used to measure brain waves.

Table 3: Summary of Electroencephalographic Indices Used in Study (Kamzanova et al, 2014)

Index	Source	Measurement Sites	Cognitive Function
Alpha-1 (suppression)	Klimesch (1999)	All cortical sites	Alertness and expectancy
Alpha-2 (suppression)	Klimesch (1999)	All cortical sites	Semantic processing
Engagement Index (EI)	Freeman et al. (1999)	Parietal montage, at Cz, P3, Pz, P4	Alertness and engagement
Frontal theta	Gevins, Smith, McEvoy, & Yu (1997)	Beta: (alpha+theta) at F1, F2, F3, F4	Mental effort
Task Load Index (TLI)	Gevins & Smith (2003)	Theta Fz: alpha Pz	Mental effort and cognitive load

Table 1 lists some of the studies that have used EEG and shows the cognitive function being tested. Tasks used to induce low workload vary depending on the cognitive function that the study intends to measure; it is important to identify which cognitive function the study is measuring as brain activity differs between each.

EEG results are often categorised by the accuracy with which the analysis of the brain activity can predict which pattern matches the task difficulty; this is termed 'correct classification'. Gevins et al (1998), Wilson and Russell (2003) and Murata (2005) all used three different difficulty levels in their tasks and all report correct classifications above 80% showing a successful outcome when analysing the brain activity data when classifying task difficulty.

When comparing brainwaves from different task difficulties, Murata (2005) reported increased total power of both alpha and beta waves as task difficulty was increased. To describe these results Murata (2005) explained that alpha and beta waves are manifested in increased alertness due to higher mental activity (Bechtereva, 1981; Gardner 1975), therefore providing a reasonable explanation for the increase in fast waves as a function of task difficulty. Some studies did not find a statistical significance between low and moderate task difficulties (Gevins et al, 1998, Murata, 2005). This seems to be the case due to the low power and minimal differences between signals caused by the lack of difference between the low and moderate tasks themselves. Murata's (2005) results conflict with those found by Gevins et al (1998) and Lei and Roetting (2011) who found that theta activity increased with increased task difficulty, whereas alpha activity significantly decreased from the lowest to the highest working memory load conditions. This is the opposite of what was originally expected. Gevins et al (1998) also reported findings that beta activity did not significantly differ between task or load in the majority of sites used.

When looking at changes in EEG activity in relation to task vigilance, Belyavin and Wright (1987) reported theta activity results that conflict with those of Gevins (1998), with increases in theta activity associated with increased time on task. It has been suggested that the difference in results indicates that increased theta activity may reflects two

different mechanisms;, one that is related to task performance (low and high task difficulty tasks) and another which is determined by the arousal state of the subject (monotonous tasks) (Brookings et al, 1996). Gevins et al (1998) state that increases in frontal theta are associated with sustained mental effort or concentration, contradictory to Lal and Craig (2001). The differing tasks in these two studies may produce the different results, providing an explanation as to why the Gevins et al's (1998) results were not as expected. Kamzanova et al (2014) reported that the alpha-1 index (lower frequency alpha) showed the closest correspondence to performance and may be the most accurate wave form to use to measure loss of vigilance over time. Kamzanova et al 2014) state 'alpha-1 is associated with a topographically diffuse synchronisation of activity across the entire scalp, indicating a generally inactive state'. This suggests that high alpha-1 activity should be present in underload conditions. It provides an explanation as to why decreased alpha activity is seen in the studies where task difficulty is increased from low workload to high workload. Kamzanova et al (2014) also found that theta activity was higher when the task was less demanding, again contradicting earlier findings (e.g. Gevins and Smith, 2003; Hockey et al, 2009). However it should be highlighted that these previous studies typically involved complex tasks requiring working memory that differ in their processing demands from monotonous tasks used by Kamzanova et al (2014). In one of the few studies which used EEG to look specifically at underload (Harris et al, 1988), it was hypothesised that continued performance of the underload task would lead to a shift in brain activity from fast waves to slow waves (due to the monotony of the task). However, the results did not support this hypothesis. Harris et al (1988) do explain that placing an operator in an underload situation does not guarantee performance deficits or that the operator will be bored. They explain that their unexpected results may be due to the fact that some participants reported engaging in their own mental variety tasks in order to maintain their levels of alertness, while carrying out the fault acknowledgement task described above. One result reported was that the variability of the lower frequency EEG bands (delta and theta) did show an increase in variability as time performing the vigilance task increased.

A.2 Cardiovascular

A.2.1 Introduction

From past studies it is evident that cCardiac activity is one of the most popular physiological measures of workload. A reason for this is that cardiac measures are easy and cheap to obtain, particularly heart rate and heart rate variability (Megaw, 2005). Yu et al (2011) state 'Comparing with EEG, ECG is easy to be recorded and is non-invasive to subjects'. Electrocardiography (ECG) is used to measure the electrical activity of the heart over a period of time. Electrodes are attached to the surface of the skin and cardiac activity is recorded by an external device. This external receiver identifies the R-waves of the ECG signal and measures the number of beats in a given time period (Louhevaara and Kilborn, 2005). Stuiver (2014) explains that in general, results have found that periods of rest or low workload are associated with decreased heart rate and blood pressure in combination with increased heart rate variability and blood pressure

variability, while the opposite has been found for periods of high mental workload (Reyes del Paso et al, 2004; Wientjes, 1992). This is usually explained as the effect of the Autonomic Nervous System (ANS), with an increase in activation of the parasympathetic nervous system in low workload conditions and an increase in activation of the sympathetic nervous system in high workload conditions (Van Roon et al, 2004). Blood pressure has also been measured on a select few studies in relation to workload, with pressure being taken at the finger. However, few studies measure blood pressure due to it restricting participant's' movement (Stuiver et al, 2014).

A.2.2 Tasks performed by participants in ECG studies

Stuiver et al (2014) used a driving simulator task, measuring physiological response in both low (6 cars on the road) and high density (35 cars on the road) traffic levels as well as response in no-fog and fog conditions for both of the traffic levels. These conditions were used to simulate different levels of mental workload; low density traffic with no fog being the lowest workload and high density traffic with fog being the highest. The main element of the driving task was switching lanes.

Jap et al (2009) also measured participants' cardiovascular activity used during a driving simulator task to measure cardiovascular activity. Participants were required to complete two drives, one 15 minute session of alert driving to act as the baseline and another 1 hour session of monotonous driving with few road stimuli. Heart rate and blood pressure measures were then compared between the underload and high workload conditions.

A vigilance state monotonous real world driving task was used by Schmidt et al (2009). Each participant was required to drive for 4 hours and to ensure monotony they drove in a low traffic density level whilst complying with traffic rules. The route was divided into 4 sections and an average heart rate was taken at each of these. The researcher sat with the participant at all times during the drive to monitor continuously and intervene if necessary.

Braby et al (1993) used a monotonous flight journey on a flight simulator to induce underload conditions. Participants were required to monitor flight instruments such as airspeed, altitude and vertical speed indicators, whilst wind turbulence caused rare fluctuations outside the predefined tolerances. These signals occurred at random and were monitored and recorded by participants.

To induce different types and levels of mental workload, Nickel and Nachreiner (2003) used the AGARD-STRES-Battery (AGARD, 1989). The battery provided differing task requirements, with the tasks differing in type (e.g. spatial vs mathematical) and level (e.g. number of items in a memory set) of mental processing. Participants were required to perform the task as fast as possible but whilst attempting to avoid any errors.

A similar mental processing task was used by Del Pasco et al (2004), who used three tasks to induce different levels of mental workload. Mental arithmetic, memory and visual attention tasks were used, however it is important to point out that the tasks only lasted 3 minutes and it is highly unlikely that underload conditions were induced. This

study indicates the change in cardiovascular activity from a baseline (rest) level to periods of mental stimulation.

A.2.3 Heart rate

Results from cardiac analysis of mental workload are varied, with some studies reporting high sensitivity of heart rate to changes in workload (e.g. Bonner and Wilson, 2002), whereas other studies report unconvincing results that fail to demonstrate reliable influences on heart rate (e.g. Casali and Wierwille, 1984). Megaw (2005) states that there are two main factors behind these inconsistencies, firstly the data can be highly affected by physical workload components, secondly that there is a lack of understanding as to why heart rate should differ depending on the level of mental workload. Lacey and Lacey (1973) proposed a hypothesis named the 'intake-rejection hypothesis' to provide an explanation of the effect workload has on heart rate. According to this, heart rate increases with the intake of information and decreases with the rejection of information.

Backs and Seljos (1993) reported statistically significant results after an analysis on the heart period (time between heart beats) during task performance, with the heart period significantly decreasing for males and females as mental load increased. However, Stuiver et al's (2014) results from their heart rate analysis were not statistically significant, with little difference between the low and high workload conditions. Stuiver et al (2014) stated that some of their non-significant results may be due to the difference in task load between conditions being very subtle. This therefore shows the importance of the task when comparing low and high workload measures and shows that the measure may not be useful if attempting to measure subtle differences. In one of the few studies measuring underload specifically, Braby at al (1993) reported similar unpromising results from the heart rate measurement. Using a task aimed to induce underload, only 10 out of the 16 participants experienced the expected decrease in heart rate across the task. Brookings at al (1996) and Dijksterhuis et al (2011) found very similar results when comparing heart rate measurements when completing the required task to baseline measurements. Both studies reported non-significant results but did find heart rates to increase when participants were completing the task. The papers do not state which analysis method was used, but the 'lack of significance' may have been due to the large between subject variability in heart rates.

In studies on vigilance, Schmidt et al (2009) and Jap et al (2009) reported results consistent with what should be expected from conditions of underload. Schmidt et al (2009) reported a clear linear decrease in heart rate over the four hours of monotonous driving. Heart rate decreased from an average of 76 beats per minute in the first section of the drive to an average of 72.5 in the last stage of the drive. Jap et al (2009) ran a similar vigilance study and reported similar results, with heart rate decreasing from 72 beats per minute pre-study to 65 beats per minute post-study. Again, it seems that results are greatly affected by the task which the investigator uses to induce 'low workload' conditions.

A.2.4 Heart rate variability (HRV)

The most widely used methods of measuring HRV can be grouped into time-domain (i.e. variability in beat-to-beat intervals) and frequency-domain (with the 0.1Hz component thought to be most sensitive to changes in workload) (Kumashiro, 2005).

Promising heart rate variability results were reported by Dijksterhuis et al (2011) who showed a clear decrease in heart rate variability in the harder tasks. Stuiver et al (2014) also reported that heart rate variability showed a clear decrease when workload of participants was increased. However, the heart rate variability of the participants who were subject to higher workloads from the start slightly increased when workload was increased further. This contradicted findings by Reyes del Paso et al (2004) and Wientjes (1992). Stuiver et al (2014) state that these results may have been obtained due to the high density traffic inducing an already high mental workload and causing a ceiling effect for the variability measures. Issues with the heart rate variability were also identified by Nickel and Nachreiner (2003); they state that the different levels of mental workload were not reflected in the 0.1Hz component of heart rate variability. Their results show that the sensitivity of the heart rate variability was extremely low when comparing different workload levels. However it is important to identify that their study is slightly different than one looking at the effect of heart rate variability on low workload or underload. Their study attempts to find out if HRV is sensitive to differences in mental strain. It is possible to discriminate between periods of low mental activity to periods of high activity from their results, showing that their 'inconsistent' findings may be due to a difference in aims.

A.2.5 Blood pressure

In a study using differing task difficulties, Stuiver et al (2014) took measures of blood pressure as well as heart rate and reported an increase in the systolic blood pressure when mental workload was increased.

The Rresults consistent with of Struiver et al (2014) were also acquired consistent with those observed by Jap et al (2009) who found that both systolic and diastolic blood pressure decreased with time on a monotonous journey. However, Tthese results contradict those found by Yakakoshi et al (2009) who used blood pressure as their main and only cardiovascular measure in their study on the physiological response to monotony whilst driving. Whereas past other results reported a decrease in blood pressure with time on monotonous task, Yakakoshi et al (2009) reported an overall statistically significant ($p<0.01$) gradual increase in blood pressure. They state that 'despite the drivers being in monotonous situations, they must still face demands, such as 'to keep an eye on surroundings' or 'to shake off their drowsiness''. They describe suggest that long monotonous driving situations can make participants considerably more stressed.

A.3 Eye activity

A.3.1 Introduction

Parameters such as blink rate and duration, pupil diameter, saccadic lengths and fixation frequency have all been used in past studies to estimate mental workload of different tasks (e.g. Brookings et al, 1996; Van Orden et al, 2000; Young et al, 2002). These parameters can be measured in several ways. For example, electro-oculography (EOG) is a technique which involves measuring the corneo-retinal standing potential between the front and back of the eye (Brown et al, 2006). Pairs of electrodes can be placed above each eye and eye movements can be measured from the potential difference occurring between the electrodes (Brown et al, 2006). Blink frequency and duration can also be recorded by electro-oculography (Megaw, 2005). Eye trackers can also be used to measure eye activity, with past studies using it to measure pupil diameter (e.g. Piquado et al, 2010). As Chen and Epps (2013) state, one of the big advantages of using eye activity as a measure of mental workload is that the three main classes of information (pupillary response, eye blink and eye movement) can all be measured by one sensor. In general, past studies have found that during periods of low workload, the average blink rate has been found to decrease (Lecret and Pottier, 1971; Van Orden et al, 2000). Research has also found that pupil diameter increases as a function of cognitive processing demands (Iqbal et al, 2004). Decreases in workload have been linked to decreases in fixation durations frequency of saccades (Rognin et al, 2004). Chen and Epps (2013) state that fixations are seen as a stationary state over regions of interest, generally interest is maintained during high workload conditions, increasing fixation duration. Therefore it would be expected that during underload conditions, blink rates would increase and pupil diameter, saccade frequency and fixation durations would decrease.

A.3.2 Tasks performed by participants in eye activity studies

In a study looking directly at underload, Young and Stanton (2002) used a driving simulator task where participants were required to catch up to and then maintain a constant distance behind a leading vehicle. A visual-spatial secondary task was also introduced, requiring participants to make a judgement as to whether rotated stick figures holding flags were the same or different. This secondary task was first used by Baber (1991). Eye movement and fixation were coded by video analysis and the total time spent looking at the secondary task (at the bottom left of the screen) was recorded.

Benedetto et al (2011) also used a driving simulator task to induce different workload levels on their participants. Their simulator required participants to perform the Lane Change Test, which required participants to change lanes at least 18 times on a 3km straight three-lane road. The Surrogate Reference Task was used as the secondary task (Mattes, 2003). Different difficulty levels were used to induce different levels of mental workload on participants with various easy and difficult tasks being used for comparisons.

Ahlstrom and Friedman-Berg (2006) used a high fidelity simulator emulating modern workstations used in select flight terminals. Pre-recorded weather conditions were presented to participants who were required to operate traffic within the airspace and issue commands to simulation pilots. Variation in traffic density determined the different

levels of mental workload. Eye activity was recorded by an oculometer consisting of an eye and head tracking system.

Different levels of cognitive load were induced by using arithmetic tasks with varying difficulty (Chen and Epps, 2013). Different arousal levels were also recorded by showing participants the International Affective Picture System (IAPS) (Lang et al, 2008).

Campagne et al (2005) instructed participants to drive 6 laps around a monotonous track on a driving simulator. At random intervals, conditions where high attention levels were required were presented to the participants. This allowed monotonous measures to be compared with measures from high attentional states.

There is a wealth of research into the use of eye activity as a measure of cognitive demand (Ahlstrom and Friedman-Berg, 2006). Reports view eye activity as being a less intrusive yet effective technique at measuring different levels of workload. However as Brookings et al (1996) and Veltman & Gaillard (1998) state, it is important to consider should be noted that eye activity is affected by other factors such as visual workload, fatigue and light levels as well as mental workload.

A.3.3 Eye blink

It is important to consider Stern et al (1994) and Recarte et al's (2008) findings on blink rate to help understand how this measure could be affected by other factors. Stern el al (1994) state that blink rate is affected by other factors such as perceptual demand of tasks and cognitive variables. Recarte et al (2008) took these findings further and provided experimental evidence that blink rate is affected by both mental workload and visual demand. They found that both these factors act in opposition to each other when participants perform any task, with the former leading to an increase in blink rate and the later to a decrease. It is therefore important to understand that the visual element of the task used to elicit a blink rate response will affect findings. However, Benedetto et al (2011) report when understanding blink duration, some past studies have focused their measurements on the 'endogenous' blink (due to perception and information processing), which can be distinguishable from other blinks (voluntary and reflex). They suggest that the endogenous blink could be an effective measure of mental processing activity.

The Ahlstrom and Friedman-Berg (2006) study reported that blink duration showed a significant effect between different levels of workload. They report that the mean blink duration of their participants decreased from a mean 0.262s at the lowest workload to 0.224s in the highest. Ahlstrom and Friedman-Berg (2006) and Benedetto et al (2011) did however report unconvincing findings in their analyses of blink frequency. Their studies on participants' responses to differing workload levels failed to find any difference between the easy low level workload conditions and the hard high level workload conditions. Ahlstrom and Friedman-Berg (2006) state 'the slope of the regression line was not significantly different from zero'. Benedetto et al (2011) did report findings of a higher blink rate in the low workload (baseline conditions), however concluded that blink rate was in itself a complex variable that was highly affected by

inter-subject variability and a range of other factors. These 'unconvincing' results may have been due to the study comparing results of several of the high workload conditions. Simple visual comparison of the low and high workloads does seem to suggest a difference in results; however the differences did not reach statistical significance. Chen and Epps, (2013) found poor correlations between blink rate and both the cognitive load and arousal factors. Results fluctuated between different difficulty levels, and no clear relationship between mental load and blink rate could be distinguished. When comparing the single (low workload) against the dual (high workload) task demands Benedetto et al (2011) found no significant differences in blink duration. Again they did not find a notable change in mean blink duration between conditions.

However, the monotonous and repetitive tasks produced consistent and expected results, indicating boredom has an effect on eye blink activity (Campagne et al, 2005). Both Haider and Rohmert (1976) and Pfaff et al (1976) report large increases in blink rate with time on task, explained as an increase in visual fatigue (Stern et al, 1994). Other studies show that blink durations significantly increased with time on task (Kobayashi et al, 1996). These results suggest an occurrence of drowsiness associated with those effects expected in underload conditions. Campagne et al (2005) confirmed these results from the past studies mentioned above, showing clear evidence of a drop in eye blink and duration when participants were exposed to a sudden variable environment in an otherwise monotonous journey. Increased blink frequencies were also recorded throughout the two2 hour drive.

A.3.4 Pupil diameter

The pupil primarily dilates due to luminance, however TEPRs (Task-Evoked Pupillary Responses) can be identified and used as an indicator of mental processing load (Beatty, 1982). Current eye-tracking systems can provide pupil diameter on both the x-axis and y-axis and size computed (Benedetto et al, 2011).

Many past studies have found pupil diameter to be an effective technique of measuring mental workload (e.g. Benedetto et al, 2011; Ahlstrom and Friedman-Berg, 2006; Chen and Epps, 2013). Chen and Epps (2013) state that among the three main types of eye activity measured in their study (blink frequency, pupil size and saccade amplitude), 'pupil features demonstrate the best attributes for discriminating different levels of cognitive load'. Results from ANOVA analysis of the pupil diameter findings show significant effects of cognitive load. Similar results were reported by Ahlstrom and Friedman-Berg (2006) who, although found there to be large individual differences in baseline diameter, found pupil diameter was smaller in the easier tasks. Benedetto et al (2011) reported mixed results; they successfully showed a significant difference in pupil diameter between the baseline and hard tasks. However Benedatto et al state 'we then expected to find differences between the easy and difficult conditions: planned contrast did not support this thought'.

A.3.5 Eye movement

Fixations and saccades can be identified and then separated by automatic identification algorithms using vertical and horizontal positions of the pupil derived from eye tracking devices (Salvucci and Goldberg, 2000).

Ahlstrom and Friedman-Berg (2006) reported statistically significant results for saccade distance, with results showing a clear linear decrease from low workload levels to high. However, unsuccessful findings were reported on the relationship between saccade frequency and mental workload with the regression line found to be not statistically significant from zero. Chen and Epps (2013) stated that, out of the eye activity measures used in their study, eye movement provided the most unpromising results. Several participants did show potential for discriminating between different load levels, e.g. one participant had saccade amplitude 1.16cm in the easiest task which increased to 1.99cm in the hardest.

A.4 Respiration

A.4.1 Introduction

There is considerable evidence that changes in task difficulty are associated with differing respiratory patterns (Sammer, 1998). There are many respiratory parameters that can be taken to indicate differing levels of workload, these include respiratory rate, breath depth, breath volume and inspiratory time (Veltman and Gailard, 1998; Yamakoshi et al, 2009). Respiration rate seems to be used most commonly, along with minute ventilation and tidal volume. Minute ventilation is the volume of gas inhaled or exhaled from a person's lungs per minute. Tidal volume is defined as the normal volume of air inspired and expired during regular breathing. Respiration is often measured by the expansion of the chest and abdomen by respiratory inductance plethysmography (Leino et al., 2001). Elastic belts are attached around the chest and peaks and valleys from the signals can then be used for the calculation of the different parameters (Veltman and Gailard, 1998). There are also other methods that can measure respiratory activity. If no information is needed on the respiratory depth, then a simple nose-thermistor (attached to glasses) can be used. The signal produced is very sharp and easily distinguishable and is almost never affected by body movements (Mulder, 1992). However this technique does not account for breaths through the mouth.

A.4.2 Tasks performed by participants in respiration studies

Yamakoshi et al (2009) used a monotonous driving simulator task to expose participants to low demand. Two experiments were run. The first exposed participants to a monotonous screen of autonomous driving, requiring no intervention. This experiment was terminated when participants displayed the appearance of a 'serious accident level' (where participants were beginning to fall asleep). The other involved driving in a simulator at a constant speed and staying in a lane. Respiration measurements were taken from a 5 minute rest period (baseline), 120 minutes of driving and then a further 5 minutes of rest after the task. This condition would be terminated if participants

reached the 120 minutes of driving or if the vehicle moved out of a specific lane (indicating poor performance).

Karavidas et al, (2010) also used simulator settings for their study on the effect of differing task difficulty levels on respiration. However, their study used a flight simulator. Participants completed 11 flight tasks of varying difficulty including an easy baseline task with minimal stimulation. Tasks were categorised as low, moderate or high based on their difficulty.

In a similar flight simulator experiment, Veltman and Gaillard (2010) first ran a flight pursuit task requiring participants to maintain a large distance behind a target jet. This task was designed to be easy; due to the large following distance allowing enabled easy recognition of anticipation of manoeuvres made by the target jet. The other task required participants to fly through a tunnel whilst simultaneously performing memory tasks of varying difficulty.

Backs et al (2000) trained participants on the TRACON video game, used by Brookings et al (1996). Participants were required to complete three mental workload conditions in a single session. Low, medium and high mental workload scenarios were obtained by varying air traffic density from 5, 10 and 15 aircraft respectively.

Results from workload studies seem unanimous in their findings that a lower task difficulty and longer periods of monotony result in decreased respiration rates, minute ventilations and breath depth (e.g. Karavidas et al, 2010; Veltman and Gaillard, 2010; Backs et al, 2000; Yamakoshi et al, 2009). This probably reflects both the decrease metabolic demand from decreased neural and muscular activity (Veltman and Gaillard, 1998).

A.4.3 Respiration rate

Respiration rate is the most popular measure of respirational activity within the workload studies reviewed, mainly due to the ease of the measurement and analysis of the data.

When mental load was low and participants were subject to simple task requirements, respiration rate was reported to be significantly much lower than during the harder tasks in a study performed by Karavidas et al (2010). There was found to be no apparent difference in respiration rate between medium and low task difficulties, so these were combined in the analysis stage on the study. A mixed model analysis comparing respiration rates between high and low/medium tasks found rate to be on average 1.78 breaths/minute lower for the combined lower and medium difficulty tasks than for the harder tasks. Karavidas et al (2010) also state that a statistically significant correlation was found when comparing the results of the respiration rates with those of the subjective NASA-TLX scores (which is a subjective measure of workload). Veltman and Gaillard (2010) measured respiration rate by the cycle time of each breath taken and reported the same conclusions as both Karavidas et al (2010) and Backs et al (2000). It is clear from Veltman and Gaillard's (2010) results that there is again a notable difference between cycle times, with times being greater in the lower difficulty tasks than for the higher difficulties.

Yamakoshi et al (2009) studied the effect of monotony on respiratory rates and reported expected results of a decreased respiration rate with increased time on task. The report found higher respiration rates were maintained for a large proportion of the start of the experiment until they finally started dropping. Yamakoshi et al (2009) explains this as 'the subject would, at least initially, attempt to fight off drowsiness, in other words to 'activate' themselves'. The extended period of driving did, however, lead to what Yamakoshi et al (2009) describe as 'the driver eventually giving up', resulting in the lower respiration rates. It is encouraging to find that both task difficulty and vigilance tasks find results that are consistent with expectations; however it is important to consider that there have been very few vigilance studies that have measured respiration rates.

A.4.4 Minute ventilation

Karavidas et al (2010) report statistically significant results when comparing the lower and higher difficulty tasks. Mixed model analysis found mean ventilation was higher for the high difficulty task and lower for the lower difficulty tasks ($p=0.014$).

A.4.5 Tidal volume

Karavidas et al (2010) reported no significant difference when comparing tidal volumes between the low and high workload tasks. It is hard to know why non-significant results were found as the author does not provide an explanation.

A.5 Electrodermal activity

A.5.1 Introduction

Electrodermal activity (EDA) is less commonly used than the other physiological measures covered within this review. However, there are studies which do use it as an indicator of workload (e.g. Collet et al, 2014; Richter et al, 2010; Mehler et al, 2012). The physiological recordings are taken from the autonomic nervous system (ANS) as it is known to give a very close estimation of subjects' arousal (Boucsein, 1993). Sympathetic endings innerve sweat glands, giving responses that are sensitive to mental stimulation (Wallin and Fagius, 1986). There are two ways to express electrodermal variations; skin resistance and skin conductance. EDA is commonly measured by the use of electrodes placed on the fingertips of participants. It is important that these sensors do not come into contact with external stimuli as results can be highly affected. This therefore raises the issue that actions using the hand from which the measures are taken cannot be as functional may be restricted when the sensors are attached. However, Mehler et al (2012) used a different design of electrode, about which they state 'the thin surface design of the electrodermal sensors minimized interference with a natural grip of the steering wheel associated with the use of more traditional cup style electrode'. This resulted in less task interference. Poh et al (2010) state that more recent developments in EDA technology has led to wrist-worn devices which allow for novel, unobtrusive, unstigmatising measures of EDA to be taken. These techniques largely

negate the issues associated with finger-worn electrodes and are generally much more comfortable to wear during long-term EDA assessments. They also allow for quicker set up and an increased ease of use than previous techniques used in years gone by. Past studies have found that skin conductance level has increased with increased cognitive demand (Mehler et al, 2012).

A.5.2 Tasks performed by participants in electrodermal studies

Collet at el (2014) requested participants to drive an experimental vehicle on a private circuit. The task required participants to control their longitudinal speed with respect to speed instructions and to perform a series of normal braking tasks in a two hour driving session. Emergency braking was also scheduled at the end of the session. Braking was issued at different speeds and therefore varied levels of mental strain.

Mehler et al (2012) used a real driving scenario, keeping to normal road rules. The N-back test (Kirchner, 1958) was used as a secondary cognitive task, with participants required to recall digits in three levels of difficulty (low, medium and high). For the low difficulty (0-back), participants simply repeated back the number presented. The medium difficulty (1-back) required participants to store one number in their memory then respond with one number back in the sequence. For the hardest level, participants held two previous items in their memory and responded with the number two items back in the sequence (Mehler et al, 2012).

A similar driving task was used by Richter et al (2010); however, instead of using a secondary task, different road types of varying difficulty were used. This change in difficulty was based on their curvature change rate. Six roads were used in total, with each participant driving for a total of eight hours. Subject baseline conditions and pre-loads were assessed prior to driving.

Larue et al (2011) used a lane keeping driving simulator task to induce monotony on participants. Four different scenarios were run, with participants being asked to drive and respect road rules for 40 minutes. The four experiment scenarios differed in both road and roadside variability (see table 2).

Table 4: Driving scenarios (Larue, 2011)

The four experiment scenarios.

		Roadside variability	
		Low	High
Road design variability	Low	Scenario 1	Scenario 2
	High	Scenario 3	Scenario 4

Road geometry varied in curvature and altitude and the roadside design varied in terms of road sign frequency and in terms of scenery.

A.5.3 Findings

Past studies have found that an increase in Electrodermal ActivityEDA indicates a readiness for action and indicates that one's attention is directed toward a stimulus (Schell et al, 2002; Stanton et al, 2004). Skin conductance is therefore expected to decrease during underload conditions, unless feelings of stress are induced (which some studies seem to indicate e.g. Yakakoshi et al, 2009), in which case skin conductance may increase.

Colet et al (2014) and Mehler et al (2012) reported promising results when comparing the response of electrodermal activity to low and high workloads. Colet et al (2014) considered electrodermal response duration (EDR) through the mean ohmic perturbation duration (OPD). Colet et al (2014) state 'the OPD is directly under the control of orthosympatheitic endings innervating sweat glands and modulating their activity, in response to stimuli'. The lower the deceleration (low difficulty/decreased mental strain), the shorter the OPD. Analysis of EDR frequency showed a significant difference between the four workloads, with higher decelerations resulting in higher EDR frequencies. Similar results were also reported by Mehler et al (2012) who found that the mean skin conductance level (SCL) of their participants was significantly affected by changes in cognitive load. The SCL was lower in the low difficulty task when compared to the higher difficulty. As was observed in the original simulator study (Mehler et al, 2009), SCL changed depending on the cognitive demands. However, Richter et al (2010) did not find such promising results in their study, reporting a non-linear relationship between varying mental workloads and electrodermal activity. Richter et al (2010) state that the reason for these inconsistent results may have been due to further cognitive processes affecting the results. They also state that marginal variables (e.g. oncoming traffic) may have had a large impact on SCL.

With a look to the vigilance decrement on task, Larue et al (2011) reported that the longer the time participants spent on task, the lower the level of skin conductance. This was consistent with their predictions. Skin conductance was shown to be correlated to the vigilance state in almost all metrics. Larue et al (2011) state that the SCL diminished in the case of the alertness decrement from 0.3 to 0.1. This shows a reduced readiness for action and suggests the driver's attention is no longer focused on the task.

A.6 Facial expression

A.6.1 Introduction

Facial expression as a measure of mental workload has not been commonly used in previous studies. Past work which has investigated facial expression as an indicator of mental workload have used either facial electromyography (EMG) (Veldhuizen et al, 2003), or temporally static 2-dimensial pictures observed and standardised using the facial affect coding system (FACS) developed by Ekman and Friesen (1978). Facial electromyography measures muscle activity by identifying and then amplifying electrical impulses that are generated by muscle fibres. Although EMG is useful in that it can measure barely visible facial movements, it does have its limitations. It requires specialised software and a trained operator as well as attaching electrodes onto the

participants face, limiting facial activities (Stone and Wei, 2011). Increased activity in the corrugator supercilii, frontalis and orbicularis oris inferior muscles have been found to correlate with mental effort when counteracting reductions in performance due to boredom and fatigue (Capa et al, 2008). This increase in muscle activity would be expected during underload conditions.

FACS defines facial expression to 44 'action units' to describe each independent facial movement. Stone and Wei (2011) state that according to the FACS, facial expressions can be categorised and their variations can be compared to changes in a person's mental workload. However, Stone and Wei (2011) also add that the FACS method requires a large amount of training and evaluation can be very time consuming.

Recent studies have looked into the use of optical computer recognition as a less intrusive method to measure low workload from facial expressions (Dinges et al, 2005). However, similar to the other measures, it is very difficult to analyse the data and high amounts of training are required.

A.6.2 Tasks performed by participants in facial expression studies

Veldhuizen et al (2003) measured EMG activity during the performance of a standardised Sternberg memory-scanning task (Sternberg, 1966, 1969) on six occasions during a simulated workday. Veldhuizen et al (2003) specifically attempted to measure fatigue throughout the day. The task used was a self-paced short term memory searching task meaning lapses in attention result directly in performance changes. EMG activity was recorded on the left-hand side measuring two particular muscles, the facial corrugator supercilii and the frontalis muscles.

Stone and Wei (2011) ran an arithmetic task of three varying difficulty levels (low, medium and high). Participants were required to calculate six sets of random numbers for each of the difficulty levels (i.e. 1 digit multiplied by 1 digit, 2 digits multiplied by 2 digits and 3 digits multiplied by 3 digits). The facial video was then analysed and coded using FACS.

Dinges et al (2005) measured subject's physiological response during both high and low workload performance demands. High workload was induced by involving more difficult performance tasks and greater time pressures. Low mental workload tasks required participants to complete easier tasks with no time pressure. Some of the workload tasks used included: the Stroop word colour interference task (Stroop, 1935), the psychomotor vigilance task (Dinges, 1985), the probed recall memory task (Dinges, 1993), the descending subtraction task (Dinges, 1981), the digit symbol substitution task, the serial addition subtraction task (Taylor et al, 1985), the synthetic workload task (Elsmore, 1994), the meter reading task, the logical reasoning task (Baddeley, 1968), and the Haylings sentence completion task (Burgess and Shallice, 1996). Changes in facial expression were measured by optical computer recognition (OCP).

A.6.3 Findings

Due to the scarcity of facial expression studies it is difficult to come to conclusions about the robustness of the different measuring techniques. There are several facial detection techniques as well as different tasks used to invoke a response; therefore evidence is weak as to whether the facial expression findings are expected from low workload conditions. For example, Veldhuizen et al's (2003) study attempts to focus on the relationship between muscular activity and the effects of fatigue. Whereas Stone and Wei (2011) would be expecting a decrease in facial movements as they study the effects of low and high workload. A within study comparison would not be reliable due to the lack of studies on each technique and aim.

The linkage between facial expressions and different levels of workload is analysed using FACS by Stone and Wei (2011). The total numbers of action units (used to describe each independent facial movement) with and without intensity were investigated. A one way ANOVA analysis showed a significant difference in action units between low and high workloads, with much fewer action units observed in the lower workload conditions. The other study which measures facial expression changes as a result of task difficulty is by Dinges et al (2005). Their complex analysis of eyebrow and mouth regions by optical computer recognition discriminated between high and low-stressor with success of '75-88% in all subjects'. It is important to consider that Dinges et al (2005) were attempting to induce stress on their participants so the results may not be as valid in a workload study.

In one of the few studies which looked directly at the effects of fatigue on facial features, Veldhuizen et al (2003) reported promising findings when analysing EMG activity. They state 'that a significant positive linear trend component was found for the corrugator and frontalis muscle'. These findings of increased muscle activity whilst experiencing fatigue are consistent with findings by (Capa et al, 2008).

A.7 Posture and movement

A.7.1 Introduction

The limitation of using body behaviour as a measure of mental workload is that it is based on the behaviour of the human, rather than mental processes (Qui and Helbig, 2012). However, there have been several studies which report promising correlations between workload and body behaviour (e.g. Graf, Guggenbuhl, and Krueger, 1995; Qui and Helbig, 2012; Roge, Pebayle, and Muzet 2001). Because many work environments in which mental workload is being investigated require operators to sit at a chair for periods of time (e.g. operating a train), observing body behaviours such as posture and movements may provide a useful non-intrusive measurement technique to detect low workload. Body posture is measured most commonly and successfully by a dynamic postural assessment chair as used by Balaban et al (2004). Data from pressure pads can be analysed and the differing postures can be observed from the data. It is possible that surface electromyographic analysis can also be used to measure posture as well as movements, with electrodes placed in numerous positions across the participant's body (Jagannath and Balasubramanian 2014). Another simple technique used is human

observation (possibly via video analysis) of body posture which can be categorised using a posture classification system (Graf et al, 1995).

A.7.2 Tasks performed by participants in body posture studies

Qui and Helbig (2012) used four psychological tests to simulate various mental activities. These were:

- A simple reaction time task requiring participants to respond to a visual stimuluss
- A compensatory visual tracking task, where participants adjusted a mouse to keep it within a circlec
- Two-column addition task testing participants' ability to sum simple addition problems in terms of speed and accuracy.

Jagannath and Balasubramanian (2014) ran a monotonous driving simulator task designed to induce boredom and fatigue (effects associated with underload). A driving environment taken from a popular driving game ('Need for Speed') was used in which the participant was required to drive for 60 minutes on a highway with low traffic density.

Roge et al (2001) instructed participants to drive for two hours on the 'Vigilance Analysis Driving Simulator'. This involved a car cabin placed on a mobile base allowing movements of the cabin. Participants were required to drive on a motorway whilst respecting driving rules. There were no other cars present on the motorway so to make the drive as monotonous as possible.

A.7.3 Findings

Body behaviour, especially body posture, has not been studied widely in mental workload studies (Qui and Helbig, 2012). However, behaviour does have an observable aspect that can reflect mental processes (Feyer, 2007). Statistically significant results have been found in the studies that have observed the correlation between behaviour and mental workload both in task difficulty comparison and vigilance tasks.

A.7.4 Body movements

Past studies have shown that participants move less during tasks with higher difficulty (high workload) than lower difficulty tasks (low workload) (Frank, 2006). When subjects completed the harder task, the seats' central pressure point travelled 40.3% less than in the easier task.

Other studies which look at the effect of a reduction in vigilance on movement have concluded that an increase in time on task produces an increase in body movements (Roge et al, 2001). Roge et al (2001) measured body movements by 2 cameras, one at the front and one at the side. They state 'activities were counted and categorised into four categories; self-centred gestures (e.g. hand touches thorax, hand touches opposite arm), non-verbal activities (e.g. sighs, yawnings), ludic activities (e.g. whistling, playing

with hair) and postural adjustments (e.g. moving of the trunk forward, tilting sideways)'. A significant difference was seen in the number and duration of movements and activities after the first driver section was complete compared with after the 4th driving section. Therefore an increased number of body movements would be expected in conditions of underload.

A.7.5 Body posture

Mulder (1979) states that, in response to new information, the human orients towards a stimulus, resulting in a forward leaning change in posture which would be expected in high workload tasks. Graf et al (1995) and Qui and Helbig (2012) both reported that for less demanding tasks, participants would lean back in their seats and maintain more comfortable positions, whilst for harder tasks the distance between the head and display significantly decreased. Therefore during underload conditions we would expect an increase in pressure on the back of the chair.

When looking at the correlation of fatigue and posture change through the implementation of a monotonous task, Jagannath and Balasubramanian (2014) reported a statistically significant correlation between time on task and percentage seat pressure distribution. Distribution increased with time on task as participants sat further back on their chair as well as leaning back. This provides further evidence that in low workload conditions participants take up a much more relaxed posture.

Appendix B End of trial questionnaire

To be completed by Researcher Participant Number: _____ Time slot: _____ Date: _____/_____/_____

Background information

1.	**What was your age at your last birthday?**		
2.	**Are you Male or Female? (please tick)**	Male	Female
4.	**How old were you when you became certified to drive a train?**		
5.	**How many years train driving experience do you have?**		

6.	**Please tick the category which best describes your role:**			
Current mainline driver	Trainer of other drivers	Trainee train driver	Other (please specify)	
7.	**If you selected "trainer of other drivers", when did you cease regular train driving? (please write in the space provided)**			

8. **Please rate your level of boredom during the first half of the task (tick as appropriate)**

Not bored at all — *Extremely bored*

1	2	3	4	5	6	7	8	9	10

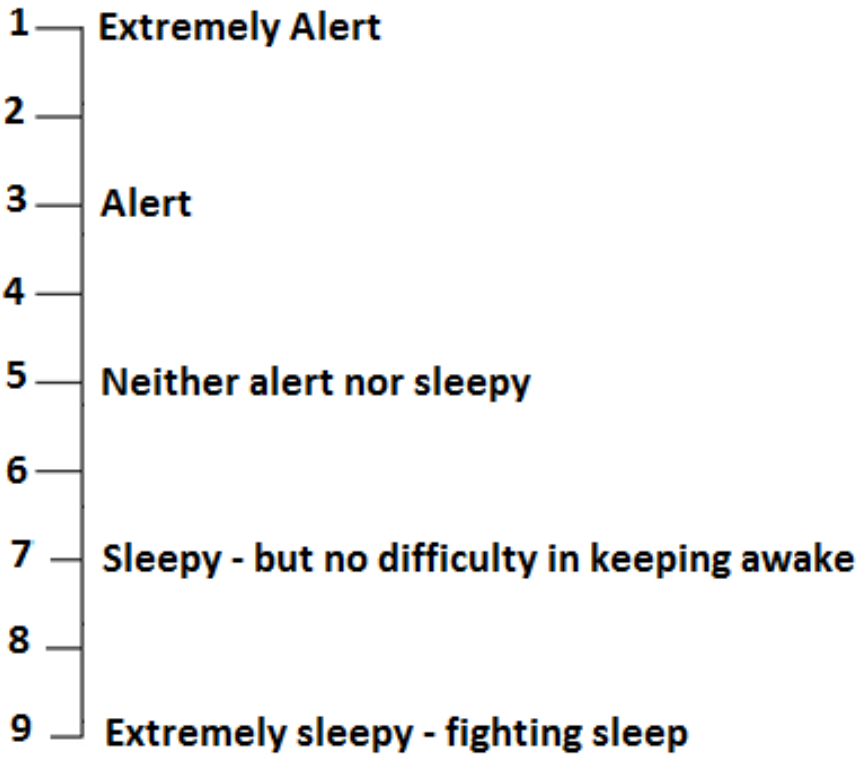
9. **Please rate your level of boredom during the second half of the task (tick as appropriate)**

Not bored at all — *Extremely bored*

1	2	3	4	5	6	7	8	9	10

10. **If you felt different at the beginning compared to the end, why do you think that might have been?**

Appendix C Karolinska sleepiness scale (KSS)



Appendix D NASA TLX

To be completed by Researcher

Participant Number: _____ Drive number: _____ Date: _____/_____/_____

NASA TLX

Your experience of the last drive

NASA TLX

For the following questions please think about the drive you just completed an "X" along each scale at the point that best indicates your experience.

Some of the scales may seem strange at first glance. If you're not confident that you have understood the descriptions of the scales, please do not hesitate to ask an experimenter for further clarification

1 **Mental Demand:** How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the drive easy or demanding, simple or complex, exacting or forgiving?

Low _____ High

2 **Physical demand:** How much physical activity was required (e.g. pushing, pulling, turning, controlling, activating, etc.)? Was the drive easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

Low _____ High

3 **Temporal demand**: How much time pressure did you feel due to the rate or pace at which the drive occurred? Was the pace leisurely or rapid and frantic?

Low _____ High

4 **Performance**: How successful do you think you were in accomplishing the goals of the drive? How satisfied were you with your performance in accomplishing these goals?

Low _____ High

5 **Effort**: How hard did you have to work (mentally and physically) to accomplish your level of performance?

Low _____ High

6 **Frustration**: How discouraged, stressed, irritated, and annoyed verses gratified, relaxed, contented, and complacent did you feel during your drive?

Low _____ High