International Journal of Human-Computer Studies 145 (2021) 102522

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Physiological indicators of task demand, fatigue, and cognition in future digital manufacturing environments

Elizabeth M. Argyle*[,a], Adrian Marinescu[a], Max L. Wilson[b], Glyn Lawson[a], Sarah Sharples[a]

[a] Human Factors Research Group, University of Nottingham, United Kingdom
[b] Mixed Reality Laboratory, University of Nottingham, United Kingdom

ARTICLE INFO

Keywords:
mental workload
fatigue
task-unrelated thought
functional near-infrared spectroscopy
heart rate
breathing rate
facial thermography
physiological sensing

ABSTRACT

As Digital Manufacturing transforms traditionally physical work into more system-monitoring tasks, new methods are required for understanding people's mental workload and prolonged capacity for focused attention. Many physiological measures have shown promise for detecting changes in cognitive state, and recent advances in sensor technology offer minimally-invasive ways to monitor our cognitive activity. Previous research in functional near-infrared spectroscopy, for example, has observed changes in cerebral hemodynamic response during periods of high demand within tasks. This work investigated the relationships among task demand, fatigue, and attention degradation in a sustained attention task, and their effect on heart rate, breathing rate, nose temperature and hemodynamic response in the prefrontal cortex and middle temporal gyrus. Analysis revealed a small but significant effect of fatigue on heart rate relative to baseline, breathing rate and hemodynamic response. Task demand had a small but significant effect on breathing rate and nose temperature, both relative to baseline, but no difference between levels of demand was observed in heart rate or hemodynamic response. Our results provide insight into what physiological data can tell us about cognitive state, ability to focus, and the impact of fatigue over time.

1. Introduction

In the era of Industry 4.0, manufacturing firms are seeking to deploy new technology-enabled systems to improve productivity, reduce costs, and enhance safe and efficient operations. The integration of novel technologies is associated with a step-change in how many tasks are performed, with a trend towards increasingly passive, cognitive work rather than active, manual work, particularly where automation is concerned. In digital manufacturing environments, it is particularly important to consider how such changes influence human factors, such as mental workload (MWL), fatigue (throughout this manuscript we will use fatigue as referring to mental fatigue), and attentional capacity, and the role of the human worker. Focused attention is a requirement for safe and effective performance of work across a range of domains (Edwards et al., 2012; Naweed, 2013), but detection and prediction of attention degradation in real-time is challenging due to the subjective and potentially disruptive nature of probe-based approaches (Smallwood and Schooler, 2006). Mind wandering, or task-unrelated thought (TUT), is a form of attentional degradation, and it has been defined as a drift in attention away from a task (Durantin et al., 2015), where intended or otherwise, executive control shifts away from the primary task to the processing of personal goals (Smallwood and Schooler, 2006). Research suggests that frequency of this type of attentional degradation this may be inversely related to MWL (Zhang and Kumada, 2017). An improved understanding of factors influencing both constructs, as well as improved methods for assessing them, may help to provide insight into the design of future complex systems.

Within the literature, a significant amount of attention has focused on managing levels of MWL during task performance, particularly in relation to MWL imposed by highly demanding or complex tasks, for example, avoiding cognitive overload situations during safety-critical air traffic control tasks (Edwards et al., 2017). For an overview of the MWL construct, the reader is directed to comprehensive reviews such as Sharples and Megaw (2015). Here, we define MWL in line with Sharples and Megaw (2015), where MWL is a property emerging from the relationships among physical and cognitive task demands, operator workload, task performance, and external and internal influences. Influences may include an individual's perceived experience of the task-load associated with the work, additionally affected by factors including temporal demand, individual background experience, and

* Corresponding author.
E-mail addresses: elizabeth.argyle@nottingham.ac.uk (E.M. Argyle), adrian.marinescu@nottingham.ac.uk (A. Marinescu), max.wilson@nottingham.ac.uk (M.L. Wilson), glyn.lawson@nottingham.ac.uk (G. Lawson), sarah.sharples@nottingham.ac.uk (S. Sharples).

https://doi.org/10.1016/j.ijhcs.2020.102522
Received 27 March 2020; Received in revised form 24 July 2020; Accepted 18 August 2020
Available online 22 August 2020

environmental factors (Charles and Nixon, 2019). Assessment of MWL in naturalistic contexts remains challenging, however, in part due to the degree of disruption that results from the use of de facto subjective self-report measures (Marinescu et al., 2018). This is of particular interest in domains where work may involve varying levels of temporal, physical, and cognitive demand, and where subjective measures would prove to be too disruptive or lacking in real-time assessment capability. In digital manufacturing, examples of tasks of this nature include automotive assembly work involving a high degree of product variation (Hermawati et al., 2015) or during work involving human-robot interaction.

Given the limitations of subjective assessment methods, physiological measures have shown promise in their capability to differentiate among varying aspects of cognitive state. Fridman et al. (2018), for example, estimated MWL while driving using deep learning of pupil dilation. Similarly, Svensson and Wilson (2002) demonstrated the use of heart rate measurements in the assessment of task performance and MWL in pilots. Previous research has also identified an association between MWL and patterns of blood oxygenation and deoxygenation (hemodynamic response), in various areas of the brain, through the application of functional near-infrared spectroscopy (fNIRS) (Jobsis, 1977; Villringer et al., 1993), with a particular focus in the domain of human-computer interaction(Causse et al., 2017; Girouard et al., 2009; Maior et al., 2015; Pike et al., 2014; Solovey et al., 2009). Similarly, other research has explored additional indicators of MWL, including breathing rate (Fairclough and Venables, 2006), electrodermal activity (Baldauf et al., 2009), and nose temperature (Marinescu et al., 2018; Or and Duffy, 2007). However, few studies have explored the relationship among demand, fatigue, and physiological response during prolonged work requiring sustained attention. This work sought to address this gap through an investigation into the effects of perceptual load and fatigue on physiological response, including hemodynamic response in the prefrontal cortex (PFC) and middle temporal gyrus, heart rate, breathing rate, and facial skin temperature. In accordance with the aim to inform the design of future digital manufacturing systems, the study employed a task modelled on a quality control work common in certain manufacturing industries, which was developed to manipulate MWL while evoking fatigue and incidents of TUT.

Through studying these factors, we make the following contributions: a novel study of the relationship among task demand, fatigue, and physiological response, and an evaluation of fNIRS and other physiological measures for the assessment of MWL and fatigue.

2. Related work

Below, we begin by reviewing the key concepts relating to motivating our task scenario from digital manufacturing. In particular, we are concerned with increased perceptual load, but also periods of underload which may lead to task-unrelated thoughts. While perceptual demand is being manipulated in our study design (see Section 3), we believe that several aspects of mental workload are affected in this scenario, and so we choose to use this broader construct throughout the article. There are many perspectives that could be taken on such tasks, and so while we explore fatigue in this article, we recognise that further work exists that includes using physiological measures to examine related constructs such as vigilance (Warm and Parasuraman, 2006), sustained attention (Langner and Eickhoff, 2013), stress (Alsuraykh et al., 2018) and cognitive load (Fishburn et al., 2014; Shi et al., 2007).

2.1. Perceptual load, task-unrelated thought, and human performance

Previous studies on TUTs have found that people may experience mind wandering up to fifty percent of the time in daily life, a frequency only slightly moderated by activity (Killingsworth and Gilbert, 2010). Theories surrounding the function of mind wandering include that it may support future-oriented thinking and long-term planning (Schooler et al., 2011; Smallwood and Andrews-Hanna, 2013), or that stimulus-unrelated thoughts may provide a mechanism for individuals to increase their level of cognitive processing, effectively supporting performance on low demand tasks (Mason et al., 2007).

TUTs are commonly associated with mind wandering episodes (hereafter, the terms will be used interchangeably), and occurrence of TUTs have been associated with automaticity and degradation of task performance (Smallwood and Schooler, 2006). Low-demand tasks that require fewer cognitive resources have been associated with more frequent occurrences of TUTs (Forster and Lavie, 2009; Smallwood and Andrews-Hanna, 2013). Task demand can be defined in line with task performance measures (e.g. response time, performance accuracy), interference from a secondary task, and level of cognitive control required to perform a task, among others (Gilbert et al., 2012). In line with the cognitive perspective, Forster and Lavie (2009) identified a relationship between perceptual load and the frequency of TUTs in a visual search task, finding that fewer TUTs in high load conditions as compared to low load conditions. This finding relates to Perceptual Load Theory (Lavie, 1995), which posits that limited attentional resources influence the processing of distractors, and that stimuli that impose a greater perceptual load reduce the ability to attend to stimuli that distract from a primary task.

Previous studies of TUT have assessed mind wandering with measures based on task performance (Durantin et al., 2015) or self-reports and thought sampling (Christoff et al., 2009; Smallwood and Schooler, 2006), but these measures are limited in relation to their subjectivity and reliance on an individual's awareness of the focus of their attention (Smallwood and Schooler, 2006). Research into brain region activation has shed some light on the phenomenon, demonstrating that changes in the direction of attention is related to activation of multiple areas within the brain. The default mode network, or the set of interconnected brain regions commonly deactivated during high demand tasks (Gilbert et al., 2012), has been implicated in the occurrence of mind wandering (Mason et al., 2007), as have areas within the prefrontal cortex (Christoff et al., 2009; Durantin et al., 2015; Mason et al., 2007) and the middle temporal gyrus (Christoff et al., 2009; Dumontheil et al., 2010). Mason et al. (2007) investigated the relationship between the default mode network and TUTs, hypothesising that the magnitude of activations across the default mode network would be directly related to an individual's likelihood of mind wandering. Mason et al. (2007) observed that as the experimental task become more practiced, cognitive processing demands decreased, and both mind wandering and activation within the default mode network increased. Similarly, Durantin et al. (2015) detected higher oxyhemoglobin concentrations in the medial prefrontal cortex during mind wandering episodes evoked during a sustained attention task.

2.2. Assessment of mental workload

Most of the work on MWL focuses the evaluation of various tasks either in laboratory condition or in real life settings. Due to technology limitations, much work involving physiological measures has been performed in laboratory conditions. Some of the most widespread techniques of inducing higher levels of MWL include the n-back task approach (Ayaz et al., 2012; Brouwer et al., 2012) or the Multi-Attribute Task Battery (MATB and MATB-II) (Comstock Jr and Arnegard, 1992; Dell'Agnola et al., 2018; Santiago-Espada et al., 2011). In this paper, we are motivated by the evolving work demands in manufacturing imposed by digital technologies; here, we focus on a sustained attention task which is representative of visual quality control inspection. This task was developed to manipulate MWL and related concepts by way of a perceptual load variation, which has also been investigated by Forster and Lavie (2009).

Current techniques used to assess MWL frequently involve task performance measures or subjective methods in the form of self-

reporting techniques. Performance measures are intended to infer the level of workload based on the performance achieved on either the primary or a secondary task (Young et al., 2015). However, one disadvantage of using primary task performance measures is that they may not be sensitive to variations in workload while there is spare capacity remaining. Using a secondary task can partially address this limitation but may become intrusive during periods of high demand in the primary task (Sharples and Megaw, 2015).

Among the most widely used subjective measures is the NASA Task Load Index (TLX), a multi-dimensional instrument that assesses perceived workload across six scales associated with different aspects of the construct (Hart and Staveland, 1988). Within the NASA-TLX, individuals self-report perceptions associated with the mental, temporal, and physical demand imposed by the task in question, as well as perceptions associated with the effort required to perform the task, individual performance assessment, and frustration levels. NASA-TLX can be used either as a weighted measure, with weights based on individual feedback on the relative importance of each dimension, or as an unweighted measure where each parameter can be considered separately, which allows the researcher to diagnose specific aspects of workload influenced by system design (Hart, 2006). As a multi-dimensional scale, NASA-TLX data is collected *retrospectively* after a task. Alternatives to the NASA-TLX include the Instantaneous Self-Assessment (ISA) tool, a subjective assessment method where an individual rates their perception of MWL *during* tasks on a uni-dimensional five-point scale (Tattersall and Foord, 1996). Compared to the NASA-TLX, the ISA tool is relatively easy to implement during real-time operations with a lesser degree of disruption, but lacks the diagnosticity of the former.

In contrast to performance-based or subjective measures of MWL, psychophysiological techniques can provide insight into an individual's level of arousal, a state closely connected to human physiology (Sharples and Megaw, 2015). Cardiac measures are some of the most widely used in this context. In some studies, heart rate and heart rate variability measures have not been found to be sensitive to variations in workload (Brookings et al., 1996; Casali and Wierwille, 1983), while others have found that heart rate significantly differed in pilots between different flight phases (Svensson and Wilson, 2002), a finding that was consistent even between separate flights (Wilson, 2002). Similarly, Verwey and Veltman (1996) observed that interbeat intervals could marginally distinguish between two workload conditions. Other psychophysiological measures include electrodermal activity (Collet et al., 2014), electrical activity within the brain (Magnusson, 2002) and cerebral hemodynamic response (Jobsis, 1977). In more recent developments, facial thermography has shown promise as technique, nose temperature has been found to be negatively correlated with increases in workload (Kang and Babski-Reeves, 2008; Marinescu et al., 2018; Murai et al., 2008; Or and Duffy, 2007). Eye tracking measures, such as pupil diameter and blink rate, have also been observed to vary with changes in perceptual load and cognitive load (Chen and Epps, 2014).

Psychophysiological indicators are of particular interest due to their minimally disruptive nature, non-invasiveness, and their potential to provide near real-time insight into individual experience (Marinescu et al., 2018). However, despite their relative merits, the mapping between psychophysiological indicators and various cognitive states is not fully understood, and there may be dissociation between different types of MWL measures (Young et al., 2015). In the following sections, we discuss a range of psychophysiological measures and their use in the assessment of MWL and fatigue.

2.2.1. fNIRS in the assessment of mental workload and fatigue

Within the range of psychophysiological indicators linked to MWL and fatigue assessment within the literature, studies of cerebral hemodynamic response offers intriguing insight into the relationship among cognition, structures, and functions within the brain. Functional near-infrared spectroscopy (fNIRS) offers a promising means of capturing brain activity in the form of blood oxygenation and deoxygenation concentrations. Functional activation of brain areas is typically indicated by an increase in oxygenated blood corresponding to a decreased level of deoxygenated blood (Richter et al., 2009). In comparison to systems like functional magnetic resonance imaging (fMRI), fNIRS offers a more lightweight and portable means of evaluating hemodynamic response, making it suitable for research applications in more naturalistic settings (Parasuraman and Mehta, 2015).

Within the fNIRS literature, findings have been mixed in relation to its ability to detect changes in MWL. While some studies have observed changes in hemodynamic response corresponding to varying levels of MWL (Causse et al., 2017; Foy and Chapman, 2018; Maior et al., 2015; 2018; Solovey et al., 2009), others have argued that fine-scale differences in MWL may be more difficult to detect with fNIRS (Causse et al., 2017; Mandrick et al., 2016). In line with previous research indicating an association between working memory capacity and blood oxygenation using an N-back task, Mandrick et al. (2016) observed a significant difference in the prefrontal cortex (PFC) oxygenation between a 0-back and 1-back task as well as the 0-back and 2-back, but not between a 1-back and 2-back task. While many studies have identified a positive association between fNIRS measures and MWL, further research is needed to address conflicting findings related to the sensitivity of the tool to measuring fine differences in MWL.

In addition, there is evidence to suggest that hemodynamic response may not only provide insight into detection of MWL, but it may also support assessment of human fatigue. Mehta and Parasuraman (2013) observed an increase in oxygenation in the PFC towards the end of a fatigue-inducing task. Finally, to a lesser degree, some studies have explored the relationship between hemodynamic response and episodes of mind wandering. Durantin et al. (2015) investigated variations within PFC hemodynamic response in order to classify episodes of task-related and TUT during a sustained attention task during performance of a Sustained Attention to Response (SART) task. Hemodynamic response findings revealed activation within the dorsomedial PFC alongside high concentration of oxyhemoglobin prior to mind wandering. Although the accuracy of the mind wandering episode classification was only slightly better than random chance, Durantin et al. (2015) recommended investigation of additional factors that may be able to improve the prediction algorithm. However, there were several limitations to their study, including a small sample size used within the classification analysis and the limited relationship between performance measures and mind wandering self-report data.

2.2.2. Facial thermography in the assessment of mental workload

Facial thermography is a technique for measuring skin temperature on the surface of the face. Skin temperature is strongly influenced by the blood flow in the area, which is under sympathetic control of the nervous system. A thermal camera is generally used to collect facial temperatures, having the advantage of obtaining readings non intrusively from the entire visible area of the face, the only downside is that it requires advanced image processing techniques for extracting the data.

Facial thermography has been used in a number of studies to test the effects of various factors on facial temperature. Naemura et al. (1993) investigated the effect of loud noise on nose temperature, reporting 100 dB noises led to a decrease in nose temperatures whereas no effect was noticed for 45 dB noises. Genno et al. (1997) examined the effect of stress on facial skin temperature using thermistors, finding that nose temperature decreased during the stressful condition while forehead temperature remained constant. Or and Duffy (2007) performed a driving study in both a simulator and an on-road driving condition, reporting a significant drop in nose temperature while participants performed mental arithmetic tasks in the simulator. Murai et al. (2008) also reported that nose temperature decreased at the onset of a ship-handling decision making task involving navigating a ship into a port. Kang and Babski-Reeves (2008) found that nose temperature decreased significantly during the learning stages of an alpha-numeric task while

forehead temperature was not affected. Nose temperature has also been reported to be influenced by arousal levels due to time on task, as examined in a 2-hour driving simulator study. Nose temperature was reported to have increased over the first 75 min and then slowly decrease until the end of the study (Diaz-Piedra et al., 2019).

However, like with fNIRS, findings are mixed; Wang et al. (2019) explored the use of low-cost thermal cameras to examine the response of face temperature to workload conditions as estimated by an electroencephalography (EEG) device, concluding that the facial thermography data could not differentiate between variations in workload (Wang et al., 2019). However, this study lacked subjective estimates of workload, workload was estimated solely from exceeding thresholds set based on the EEG data, and findings validating the EEG thresholds were not presented. For these reasons, results of this research are difficult to compare with other studies. While the authors correctly point out the limitations of subjective measures, we would also argue that MWL is a multifaceted concept (Sharples and Megaw, 2015), and that with current understanding of physiological indicators, subjective estimators still offer a valuable means of understanding the experience of workload.

2.2.3. Cardiac and respiration measures in the assessment of mental workload

The rhythm of the heart is modulated by the sinoatrial node, which is influenced by both the sympathetic and parasympathetic branches of the autonomic nervous system (ANS). There is a continuous balance between the two branches of the ANS; the sympathetic activity increases the heart rate while the parasympathetic branch decreases it. Because it is controlled by the ANS, cardiac activity has been considered a good candidate measure for workload. For these reasons, as well as the relative ease of collecting data, cardiac and respiratory measures have a long history of being used in workload assessment, particularly in the aerospace domain.

In early work, Casali and Wierwille (1983) found that mean heart rate, heart rate standard deviation and breathing rate computed over seven minutes time intervals were not sufficiently able to distinguish between three levels of demand manipulated by means of communication call signs, during a flight simulator study conducted on 30 participants. Brookings et al. (1996) assessed heart rate, breathing rate and respiration amplitude during a simulated air traffic control task, finding that heart rate and breathing rate amplitude did not demonstrate significant differences in relation to demand; however, the authors did observe that breathing rate was higher as the complexity of the scenarios increased (Brookings et al., 1996). In a more recent exploratory study, Lehrer et al. (2010) assessed whether cardiac measures could be used to measure workload during a flight simulator study where seven professional pilots performed 18 flight tasks. The MWL for the tasks was rated using the NASA-TLX by both experienced test pilots as well as the pilots performing the task. Flight performance was evaluated by experts on a five-point scale. They found that standard deviation of normal R-R inter-beat intervals (SDNN) was associated with expert ratings of MWL even when the NASA-TLX results of the participants were not, suggesting that the cardiac measures assess something that the NASA-TLX does not (Lehrer et al., 2010).

Similarly, Bonner and Wilson (2002) recorded heart rate data from pilots, copilots and loadmasters during test and evaluation flights, with tasks including aircraft handling and normal flight. While due to the variable nature of the flights, inferential statistics could not be used, but Bonner and Wilson (2002) observed a substantial increase in heart rate for the pilot in control which could play a role in identifying sharing of workload problems inside the cockpit. The authors also observed that increases in heart rate during high workload conditions were reported when the crew members were moving through the aircraft, and as such, removal of these artefacts should be considered (Bonner and Wilson, 2002). In another flight scenario study, Svensson and Wilson (2002) recorded heart rate during 35 simulated combat missions and found that heart rate averaged over two minutes time intervals and significant differences were found between approach and intercept phases.

Despite recent findings offering support to a link between MWL and cardiac response, there is no clear agreement with regard to either cardiac and respiratory response measures and their association with changes in workload; this is most likely because the measures were either not sensitive enough to the changes or because of external factors that can clearly influence heart rate and breathing rate, as reported by Bonner and Wilson (2002).

3. Method

3.1. Research question

Building upon findings within the literature, we explored the mapping of physiological response to variations in demand and fatigue throughout a prolonged visual inspection task requiring sustained attention. In order to identify the efficacy of physiological measures to serve as indicators of MWL, fatigue, and attentional degradation, the work sought to investigate the relationship between perceptual load, fatigue (as inferred from time on task), and physiological response during tasks requiring sustained attention. This question was approached with the underlying assumption being that there was an association between perceptual load, MWL, and frequency of TUT, an assumption that was based on the findings of Forster and Lavie (2009).

3.2. Experimental design

The present study simulated a quality control inspection task that required sustained attention and would vary in MWL and induce fatigue over time, with the aim of also creating opportunities for TUT. In line with the findings of Forster and Lavie (2009), who identified a relationship between perceptual load and the frequency of TUTs, the present study employed perceptual load as an independent variable corresponding to demand, hypothesizing that this would manipulate experienced MWL, which was supported by early pilot testing. The study adopted a two-factor within-subjects approach to investigate the effects of varying perceptual load (low vs. high perceptual load) and time period (10 levels corresponding to block number) on physiological response during the visual search task. Dependent variables included oxyhemoglobin (HbO) and deoxyhemoglobin (HbR) concentrations in the prefrontal cortex (PFC), HbO and HbR concentrations in the middle temporal gyrus, heart rate, breathing rate, and facial skin temperature. Following each block of tasks, participants were asked to provide subjective estimates of their MWL, their level of fatigue, and whether their attention had been on-task or off-task at the time.

The experimental task was modelled after a visual search task used by Forster and Lavie (2009) in an investigation of the influence of perceptual load on mind wandering frequency. During the task, participants were presented with a set of one or more images of cork coasters and were asked to identify whether the set contained a coaster with a target defect type. Target defects were cuts and flattened corners, and participants were asked to press a specific key (“Space” for cuts and “Num Pad 0” for flat corners) to record which defect type was present within each set, and each set contained only one target defect type.

In the low demand condition, participants were shown a single image (Fig. 1 - left), while in the high demand condition, participants were shown the target defect in addition to three other distractor images (Fig. 1 - right), any one of the four images could contain the target defect. Distractor images displayed alternative defect types (scratched surfaces, glue spills, and dents). For both levels of demand, images were displayed at one of six points on the screen, each equidistant from the screen’s centre.

Each block included fifty tasks in which the stimuli were presented for 1700 ms, followed by a 1300 ms blank screen during which

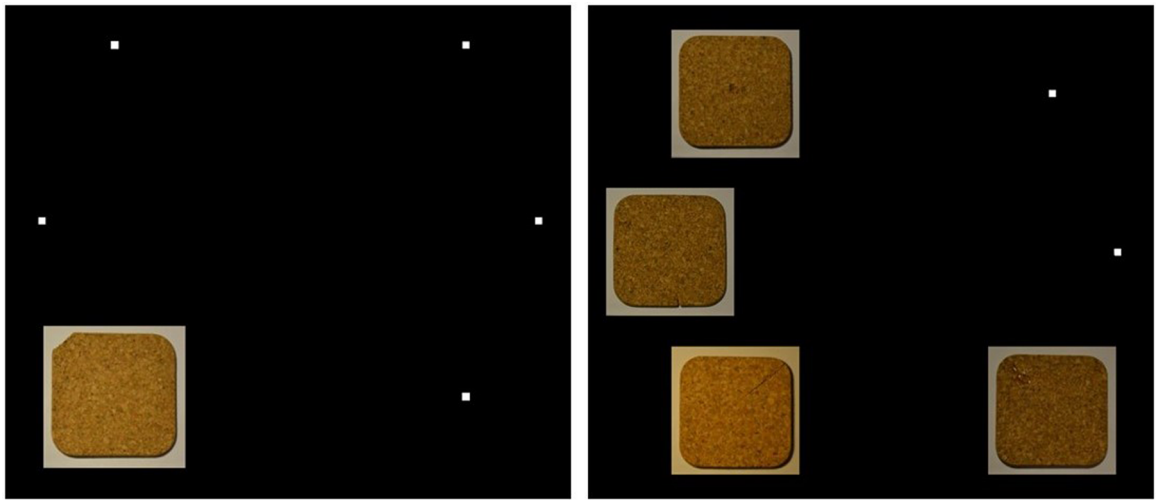

Fig. 1. Task screenshot: low demand condition on the left and high demand condition on the right

participants could respond. A block consisted of either high or low demand tasks. Before each of the task blocks, the participants performed a baseline task during which they were asked to follow a "plus" sign moving on the screen. In between each baseline task and task block the participants were presented with the text: "The task will start in 10 s" on a dark background. The main reason for this was so that participants were not surprised by the sudden start of the task.

3.3. Participants

A total of 36 participants took part in the study (52.8% female and 47.2% male), with a mean average age of 28.2 years ($\sigma = 8.4$). Participants were required to have normal or corrected-to-normal vision and were recruited from the University of Nottingham and surrounding community. Participants were provided with a 15 voucher upon completion of the study. This study was approved by the University of Nottingham's Faculty of Engineering Ethics Committee.

3.4. Equipment

Several sensor systems were used to collect physiological measures from participants, including an Artinis Octamon+ fNIRS device, a Zephyr Bioharness 3, a FLIR thermal camera, and a video camera placed facing the participant.

The Artinis Octamon+ device included 8 optodes split into two groups. One group was placed over the prefrontal cortex and also included a short separation channel (distance to receiver 35 mm and 10 mm for the short separation channel), while the second group was placed over the mid-temporal gyrus. The emittors use 760 nm and 850 nm wavelengths and the fNIRS data was acquired at 10 Hz. The optodes were positioned on a full headcap. Two sizes were available for the headcap with slightly different available positions for the front optodes, as presented in Fig. 2 (right and center). Fig. 2 (left) shows the optode positioning on the side of both the medium and large headcaps. The influence of headcap size on the optode positioning will be revisited in the discussion section.

In addition to the physiological sensors, the materials also included a questionnaire capturing background experience and demographic data, a NASA-TLX questionnaire (Hart and Staveland, 1988), and a 9-point fatigue rating scale. The laboratory setup consisted of a desk and computer workstation which displayed the training and experimental tasks via PsychoPy2 (Peirce et al., 2019), based upon a dataset of images of cork coasters with visible defects.

The Zephyr BioHarness 3, used to collect heart and breathing measures, is a physiological monitoring module attached to a chest strap. The device reports heart rate and breathing rate at 1 Hz and heart rate R-R intervals per R peak detection.

Facial thermography data was collected with the FLIR A65sc, uncooled microbolometer type, thermal camera (45 degree lens) using the ResearchIR software at a mean frame rate of 7.45 FPS. The spectral range of the camera is 7.5-13μm, which is not overlapping with the wavelengths emitted by the fNIRS device.

3.5. Procedure

The researcher introduced the study's objectives and procedure to each participant, and after discussing any questions, the participant provided informed consent and then completed the questionnaire. The participant was then shown how to fit the bioharness, which they completed independently. The participant then completed a training which demonstrated the low and high demand tasks. Performance feedback was provided during the training. After the participant confirmed that they felt comfortable with the task, the researchers worked with the participant to fit the remaining physiological sensors and to configure the thermal and video cameras.

The main body of the experiment began by collecting a set of baseline physiological measures from each participant; in the first two-minute baseline activity, the participant was instructed to maintain a calm state while watching a "+" symbol float around the monitor. The second baseline activity required the participant to sit calmly and reflect upon a positive memory from a vacation, also for two minutes. Following the baseline, participants completed five blocks of the low demand task and five blocks of the high demand task, presented in a fully randomized order. Participants responded to probes between blocks relating to MWL, fatigue, and mind wandering experienced during the previous block. Afterwards, participants provided estimates of their perceived workload in line with the raw NASA-TLX instrument, which assesses MWL across six dimensions (mental demand, physical demand, temporal demand, effort, performance, and frustration).

3.6. Data processing and analysis

Prior to analysis, one participant's data was removed due to a lack of engagement with the task, leading to a sample of 35 participants being included in the study.

3.6.1. fNIRS pre-processing

The pre-processing of the fNIRS data began with a visual inspection of data quality, where channels were excluded if they had high levels of noise or if a heart beat could not be observed in the raw signal. Table 1

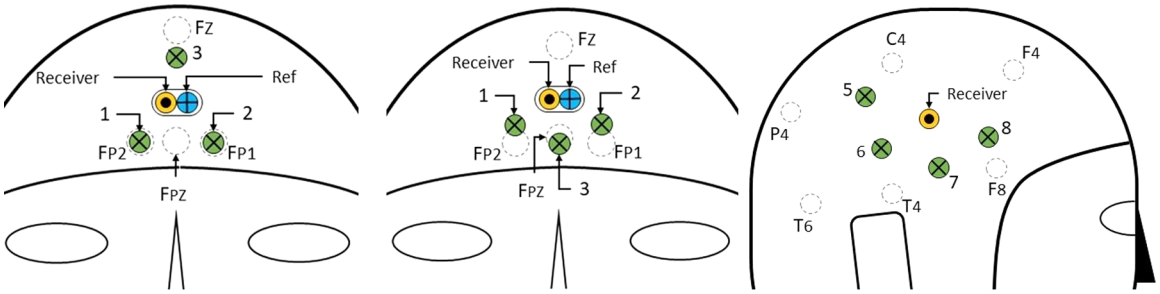

Fig. 2. Frontal positioning of the optodes on the medium size headcap (left) and large size headcap (center). Channel 4 is a short separation channel and is labelled as "Ref" in the figure. Side positioning of the optodes on both the medium and large size headcaps (right). The emitters (green) are marked with an X sign, the short separation emitter (blue) is marked with a + sign while the receiver (yellow) is marked with a black circle in the middle. The dotted circles represent the positions of some 10/20 system locations relative to our optode placement. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Total channels included in the analysis across participants

	PFC				MTG			
Optode	1	2	3	4	5	6	7	8
No. kept (out of 36)	25	20	17	15	13	15	15	16
Percentage kept	69.4%	55.5%	47.2%	41.6%	36.1%	41.6%	41.6%	44.4%

describes the proportion of data included in the analysis for each of the channels. Data from the short separation channel (4) were not included in this analysis. For some participants, where possible, specific blocks of data within the channels were removed, rather than entire channels. The included data were pre-processed using Homer2[1] with a pipeline inspired by Pinti et al. (2019), detailed below, involving conversion from raw data to optical density, a baseline shift and motion artifact correction step, bandpass filtering with a Butterworth filter, conversion of optical density data to concentration, and block averaging across a range of 20 s before the start of each block and 146 s following the start of each block.

1. Raw data was converted to optical density using the 'hmrIntensity2OD' function
2. The optical density data was corrected for baseline shifts and motion artifacts using the 'hmrMotionCorrectPCArecurseCh_SG' function that performs a tPCA on each channel and a Savitzky-Golay smoothing; the parameters used were:
   - tIncMan: all included data was considered, this was a vector of 1's
   - nSV: 0.97, removing the first n components accounting for 97% of the variance
   - maxIter: 100
   - FrameSize_sec: 10
   - turnon: 1
   - the Savitzky-Golay filtered was used with the default polynomial order of 3
3. The resulting data was bandpass filtered using the 'hmrBandpassFilt' function, between 0 and 0.5 Hz on the default Butterworth filter.
   - high pass filter: 0
   - low pass filter: 0.5
   - filter type: the default Butterworth filter
   - filter order: the defaults in Homer2 were used, 3 for the low pass and 5 for the high pass
4. The bandpass filtered data was converted from optical density to concentration using the 'hmrOD2Conc' function with partial path-length factor 6 for each wavelength.
5. The concentration data was block averaged using the 'hmrBlockAvg' function with a range of −20 s (baseline) and 146 s as the block duration

The recovered hemodynamic response function for each of the task blocks was averaged across optodes from each of the two main brain regions analysed. This resulted in having one hemodynamic response function for the prefrontal cortex and one for the middle temporal gyrus; these were averaged over the interval [0,146] seconds for both HbO and HbR. The hemodynamic response function graphs below were generated using the Matlab function shadedErrorBar.[2]

3.6.2. *Facial thermography pre-processing*

The first stage of the analysis involved the extraction of temperatures from various areas of the face, for which generating facial landmarks for every frame in each video was needed. The second stage involved filtering the data, removing task blocks during which the facial landmark tracking under-performed and running the statistics on the remaining data.

For the landmark tracking, the DeepLabCut (Mathis et al., 2018; Nath* et al., 2019) package for markerless pose estimation was used. The workflow included the following steps:

- Create a dataset that will be used for training and testing: a sample of 20 frames per participant were used
- Manually label each of the sampled frames with the 7 facial landmarks shown in Fig. 3
- Split the dataset into training and testing: the default 0.95 split was used
- Choose a pre-trained neural network and refine end-to-end to adapt its weights: the ResNet-50 weights were used to initialize the model and the model was trained for 210000 iterations
- Evaluate performance on the test dataset
- Analyze: use the new weights to estimate the landmark positions in all videos. This generates the coordinates of each landmark in each frame, which were later used to extract the temperatures from the frames containing the thermal data.

Once the temperatures were extracted, the timeseries data was smoothed using a robust local regression approach with a second

[1] https://homer-fnirs.org/

[2] https://github.com/raacampbell/shadedErrorBar

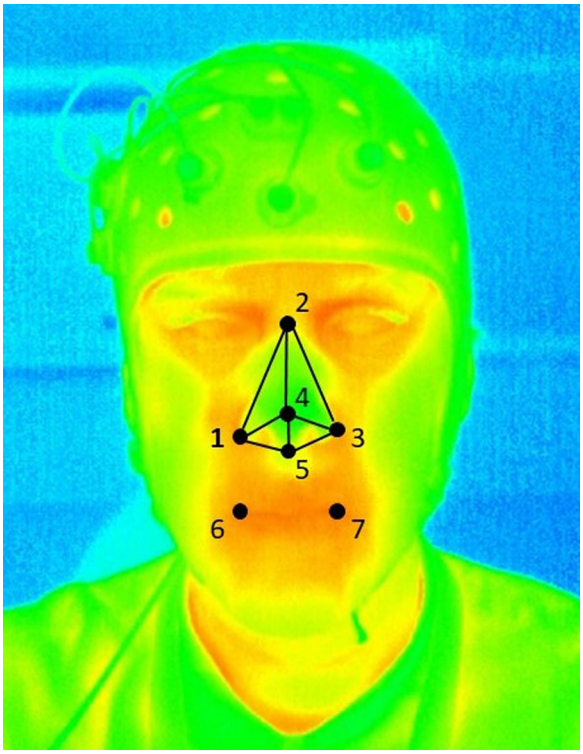

Fig. 3. Facial thermography landmarks

degree polynomial model and a span of 1% of the data. The temperature timeseries for each participant were compared against the block conditions. A few factors influenced the quality of the data, including face orientation for the accuracy of the landmarks and the presence of an automatic non-uniformity correction (NUC) performed by the thermal camera. Data blocks influenced by these factors were removed from the analysis, and about 54.2% of the blocks were kept for the final analysis; these limitations are further explored in the discussion chapter. The final step was extracting the difference between the mean temperatures during the task blocks and the mean temperatures during the baseline before each block.

3.6.3. Heart rate pre-processing

Out of the 35 participants in the study, we have heart rate data for 34 participants as for one, the device did not record the data. Another participant stopped their participantion after 8 blocks of the study, resulting in 338 block being recorded. Heart rate was inspected for outliers. In the heart rate data, values less than 35 beats per minute were considered to be inaccurate and associated with sensor malfunction or poor fit. Values below this threshold were removed from the analysis, resulting in a sample of 335 data points out of 338 recorded being included in the heart rate analysis. Two of the participants had 10%, respectively 20% of the heart rate data removed.

3.6.4. Breathing rate pre-processing

Out of the 35 participants in the study, we have breathing rate data for 34 participants as for one, the device did not record the data. Another participant stopped their participantion after 8 blocks of the study, resulting in 338 block being recorded. Similar to heart rate, breathing rate was inspected for outliers. Out of the total of 338 blocks captured during the study, 1 were associated with poor data quality in the breathing rate sensor, with mean values of 0 breaths per minute. One participant had 10% of the breathing rate data removed. These data points were removed from the final analysis, resulting in a sample of 337 blocks being considered.

3.6.5. Subjective measures

In addition to the physiological measures, subjective data was collected including a questionnaire capturing background experience and demographic data, a NASA-TLX questionnaire Hart and Staveland (1988), and a 9-point fatigue rating scale.

3.6.6. Statistical analyses

All dependent variables were checked for normality and equal variance. Differences in task performance were investigated using a two tailed t-test with a 0.05 significance level, while differences in MWL between demand condition were investigated with a series of Wilcoxon Signed-Rank tests. The association between reports of TUTs and task demand was assessed with a crosstabulation analysis. A series of linear mixed effects models were used to assess the relationship between perceptual demand and block number, and their collective effects on heart rate, breathing rate, nose temperature, and hemodynamic response variables in the prefrontal cortex and middle temporal gyrus.

All linear mixed effects models were run using the Matlab (version R2019b) function fitglme. A random intercept and slope was included for participant, as well as the correlation among perceptual load and block number. The formula for the models is described below using the Wilkinson notation. Block number is denoted by blockNo and the participant by participantNo. The same model was used for each of the physiological measure, so in the notation, "Physiological measure" stands for: heart rate, breathing rate, nose temperatures and all the fNIRS measures reported in the paper.

Physiological measure ~ 1 + blockNo*condition + (1 | participantNo) + (1 + perceptualLoad | participantNo) + (1 + blockNo | participantNo)

3.7. Hypotheses

It was hypothesized that perceptual load would be associated with significantly different blood oxygenation patterns in the two brain regions. It was also hypothesized that increases in perceptual load would be associated with higher, mean heart rate and mean breathing rate, but lower levels of facial skin temperature. We enumerate these below:

H1 Perceived MWL will be higher in the high task demand condition than in the low.
H2 Frequency of TUTs will decrease in the higher demand task condition.
H3 Self-reported fatigue rating will increase over time.
H4 Heart rate will increase in the high task demand condition as compared to the low.
H5 Breathing rate will increase in the high task demand condition as compared to the low.
H6 Hemodynamic response will vary between the low versus high task demand condition.
H7 Nose temperature will decrease in the high task demand condition as compared to the low.
H8 Heart rate will decrease as time progresses.
H9 Breathing rate will vary as time progresses.
H10 Hemodynamic response will vary as an individual progresses through the 10 task blocks.
H11 Nose temperature will vary as as an individual progresses through the 10 task blocks.

4. Results and analysis

Below, we begin by checking that the manipulation of our primary independent variable was successful using task performance data. After that, we first examine H1-H3 relating to the subjective ratings of experience, before moving on to H4-H11 relating to physiological measures.

Table 2
Performance data including error rates and response times for both low and high demand conditions

	Low demand condition		High demand condition	
	Mean	SD	Mean	SD
Error rate [Percentage]	1.1	2.1	3.6	4.3
Response time [s]	0.85	0.094	1.2	0.17

4.1. Task performance

Error rates and response time during each block were used as a performance measure, and Table 2 shows the mean and standard deviation of these measures during the low and high demand conditions. It is important to note that in this and the following analyses, data were not considered to be completely independent as the same participant was represented up to ten times in the analyses, due to sampling data during and following each block. The error rate data was compared using a t-test, showing a significant difference between the low and high demand conditions $t(34) = -6.1$, $p < 0.001$ (95% CI, -0.034 to -0.017), Hedges's $g = -1.05$, CI(-1.48, -0.64). Response time was also significantly different between the low and high demand conditions, as tested using a t-test $t(34) = -21.02$, $p < 0.001$ (95% CI, -0.38 to -0.31), Hedges's $g = -2.89$, CI(-3.63, -2.24). We conclude that the significant drop in both performance measures confirms that perceptual load manipulation was successful.

4.2. Subjective ratings

This section presents the subjective ratings of experience provided by participants during the tasks.

4.2.1. Mental workload

In line with Hypothesis H1, that the variations in perceptual load would reflect differences in MWL, the raw NASA TLX instrument was used to capture subjective ratings of parameters reflecting participant MWL levels. This hypothesis was supported, with a series of Wilcoxon Signed-Rank Sum tests indicating that workload ratings were significantly higher in the high perceptual load condition than in the low perceptual load condition in terms of mental demand ($z = -4.4$, $p < 0.01$, $d = 0.74$), temporal demand ($z = -4.86$, $p < 0.01$, $d = 0.82$), performance ($z = -3.98$, $p < 0.01$, $d = 0.67$), and effort ($z = -4.53$, $p < 0.01$, $d = 0.76$). There was no significant difference between low and high perceptual load conditions in terms of physical demand ($z = -1.94$, $p = 0.05$, $d = 0.33$) and frustration ratings ($z = -0.94$, $p = 0.35$, $d = 0.16$). These results provide further evidence that the chosen perceptual load task successfully created two distinct conditions.

4.2.2. Frequency of task-unrelated thought

A crosstabulation analysis and Chi-square test was used to examine the relationship between perceptual load and the occurrence of TUTs during the task, as subjectively rated by the participants. A significant relationship was identified, finding that participants were less likely to experience TUTs in the high perceptual load condition than in the low demand condition ($\chi 2(1) = 10.30$, $p = 0.0013$, $\varphi = 0.55$). This finding is consistent with findings in the literature and supported the hypothesis that perceptual load would manipulate frequency of TUT episodes (Hypothesis H2).

4.2.3. Relationship between time and fatigue rating

In addition to rating perceived MWL and attentional direction, participants also provided a subjective estimate of their level fatigue in between task blocks. It was hypothesised that fatigue would increase over time, so a one-tailed Wilcoxon Signed-Rank test was performed to test this assumption. Findings provided support for this hypothesis, identifying a significantly higher level of self-reported fatigue following the final block than following the first block ($z = -4$, $p < 0.001$, $d = 0.67$). These results show that the participants perceived a higher level of fatigue at the end of the task as compared to the beginning, indicating that as the variable block number increases, so does the level of fatigue. In line with these findings, physiological data was analysed with respect to block number as opposed to the fatigue rating.

4.3. Physiological measures

In this section, we focus on the physiological measures taken during the study. We examine whether they are affected by the two perceptual demand levels, and consequently how they relate to differences experienced in MWL and fatigue.

4.3.1. Heart rate

A Kolmogorov-Smirnov test indicated that heart rate did not follow a normal distribution, $D(335) = 0.96$, $p < 0.001$. In support of Hypothesis H8, a linear mixed effects model indicated that in relation to the heart rate task-baseline difference, block number (time) had a small but significant effect, with heart rate difference from baseline decreasing as time progressed ($\beta = -0.17$, 95% CI $= (-0.33, -0.0056)$, $p = 0.042$). Task demand did not have a significant effect on heart rate task-baseline difference ($\beta = 1.05$, 95% CI $= (-0.24, 2.35)$, $p = 0.11$), a finding which failed to support Hypothesis H4. Likewise, the task demand-block number interaction was not significant ($\beta = -0.18$, 95% CI $= (-0.39, 0.031)$, $p = 0.094$). These results indicate that heart rate difference from baseline decreased as the task progressed while the same measure was not influenced significantly by changes in demand.

4.3.2. Breathing rate

A Kolmogorov-Smirnov test indicated that breathing rate was not normally distributed, $D(337) = 0.34$, $p < 0.001$. A linear mixed effects model indicated that in relation to the breathing rate task-baseline difference, block number had a small but significant effect, with breathing rate difference from baseline increasing as time progressed ($\beta = 0.19$, 95% CI $= (0.062, 0.33)$, $p = 0.0041$). Task demand also had a significant effect, with breathing rate difference from baseline increasing by a factor of 1.51 in the high demand versus low demand condition ($\beta = 1.51$, 95% CI $= (0.50, 2.53)$, $p = 0.0036$). Similarly, the task demand-block number interaction was also significant, ($\beta = -0.17$, 95% CI $= (-0.34, -0.0038)$, $p = 0.045$). The increasing breathing rate trend over time is shown in Fig. 4, while the increase in breathing rate with task demand is presented in Fig. 5. These results indicate that breathing rate difference from baseline increased as the task progressed as well as the fact that the same measure was influenced by changes in demand, increasing in the higher demand condition as compared to the lower demand one.

4.3.3. Cerebral hemodynamics in the pfc and middle temporal gyrus

A Kolmogorov-Smirnov test indicated that the fNIRS data for both the PFC and MTG did not follow a normal distribution, $D(347) = 0.49$, $p < 0.001$. Table 3 shows the results of the linear mixed effects model for hemodynamic respnse (HbO and HbR) in both the prefrontal cortex and mid temporal gyrus. It can be observed that both independent variables and their interaction had no significant effect on HbO or HbR concentration in both the PFC and MTG areas. Findings failed to provide support for either Hypotheses H6 or H10, indicating that the hemodynamic response was not influenced by either the change in demand or changes in fatigue. To illustrate, the hemodynamic response curves comparing low versus high demand levels are shown in Fig. 6.

4.3.4. Nose temperature

A Kolmogorov-Smirnov test indicated that nose temperature relative to baseline did not follow a normal distribution, $D(194) = 0.3$,

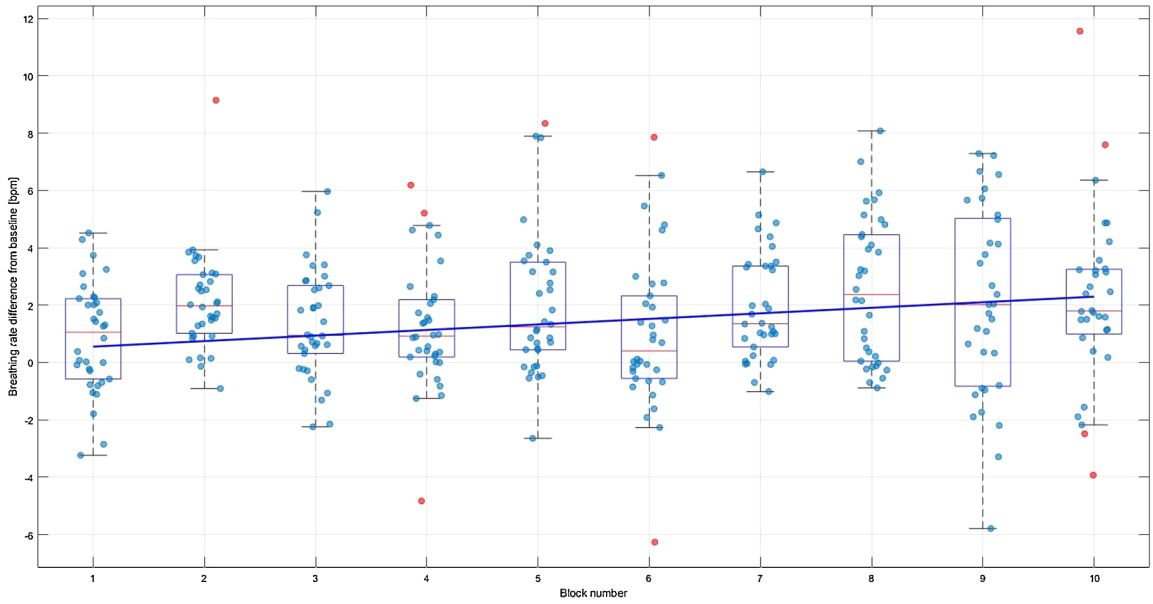

Fig. 4. The effect of time on breathing rate was significant, the data showing a small increase in breathing rate compared to baseline as time progressed

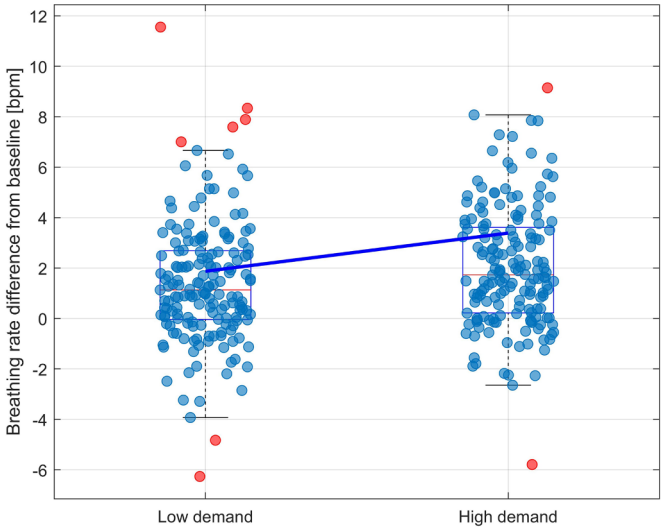

Fig. 5. Demand level had a significant effect on breathing rate difference from baseline, showing an increase in the higher demand condition

$p < 0.001$. A linear mixed effects model indicated that in relation to nose temperature task-baseline difference, block number (time) did not have a significant effect on nose temperature ($\beta = -0.01$, 95% CI = $(-0.03, 0.02)$, $p = 0.65$), a finding which did not support Hypothesis H11. However, Hypothesis H7 was supported; task demand had a significant effect, with nose temperature decreasing by a factor of -0.25 in the high demand versus low demand condition as compared to the baseline ($\beta = -0.25$, 95% CI = $(-0.45, -0.04)$, $p = 0.02$) (Fig. 7). Task demand-block number interaction was not significant, ($\beta = 0.03$, 95% CI = $(-0.005, 0.05)$, $p = 0.1$). These results indicate that nose temperature difference from baseline was not influenced by changes in fatigue while the same measure was influenced by the level of demand.

5. Discussion

During a prolonged visual inspection task, we explored the relationship among task demand, fatigue (represented by block number), and their collective effect on task performance, MWL, TUT, and physiological response. The MWL ratings as well as task performance measures reflected the expected behavior of our primary independent variable manipulations. The difference in the performance data, for both response time and error rate, indicated that responses to the NASA-TLX scale reflected workload differences across multiple dimensions, confirming the hypothesis that the contrasting levels of perceptual load would manipulate the perceived experience of MWL.

5.1. Mental workload and task-unrelated thought in a visual inspection tasks

The prolonged visual inspection task with two levels of demand was used as a means of manipulating the level of MWL experienced by the participants. Task performance as well as subjective MWL ratings confirmed the difference between the two levels succeeded in manipulating the level of perceived MWL.

In addition to confirming the hypothesis that perceptual load would manipulate MWL, findings also revealed that participants were less likely to experience TUTs during the high perceptual load condition as

Table 3
Summary of the linear mixed effects models for both the Pre-frontal Cortex (PFC) and Middle Temporal Gyrus (MTG)

	Block Number				Task Demand				Interaction			
	β	95% CI		*p*	*β*	95% CI		*p*	*β*	95% CI		*p*
		Lower	*Upper*			*Lower*	*Upper*			*Lower*	*Upper*	
PFC HbO	0.015	−0.03	0.06	0.49	−0.11	−0.46	0.24	0.52	0.02	−0.04	0.08	0.47
PFC HbR	0.01	−0.01	0.03	0.47	−0.05	−0.21	0.12	0.6	0	−0.03	0.02	0.77
MTG HbO	0.032	−0.023	0.087	0.25	0.2	−0.245	0.645	0.38	−0.021	−0.096	0.053	0.57
MTG HbR	−0.01	−0.04	0.02	0.56	−0.09	−0.33	0.15	0.48	0.02	−0.02	0.06	0.29

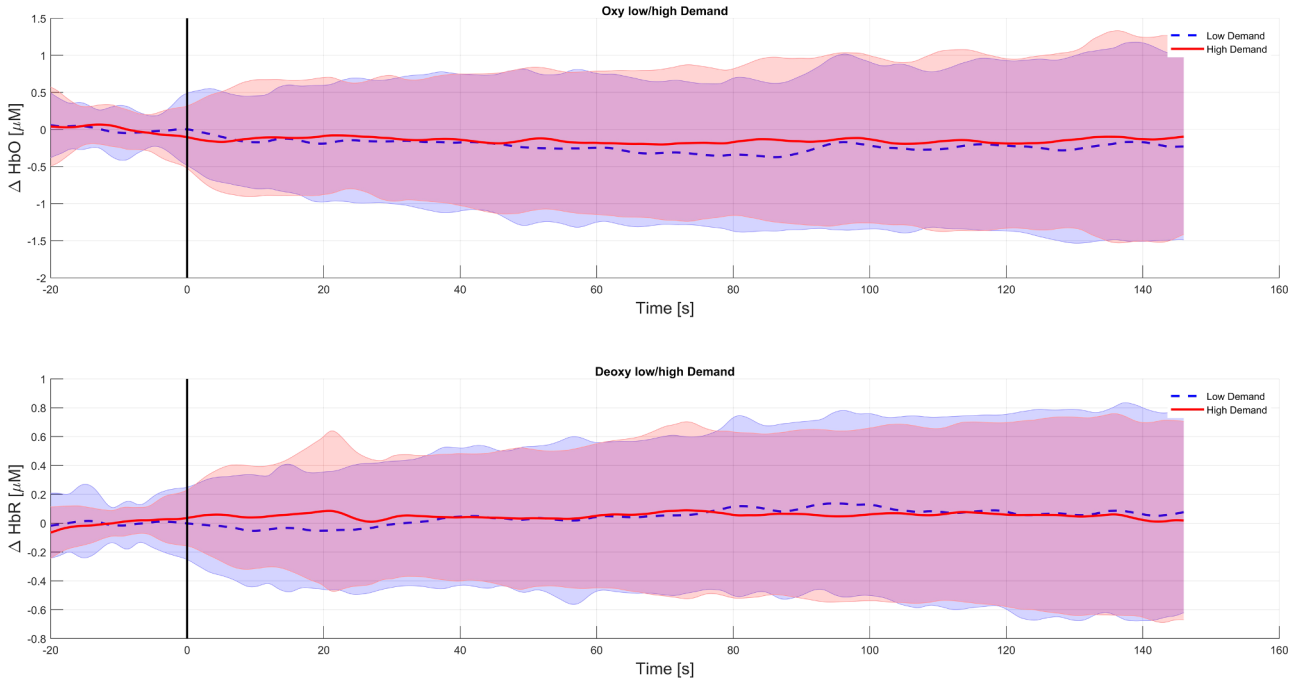

Fig. 6. Hemodynamic response for both HbO and HbR in the low demand condition as compared to the high demand condition. The shaded areas represent the standard deviation

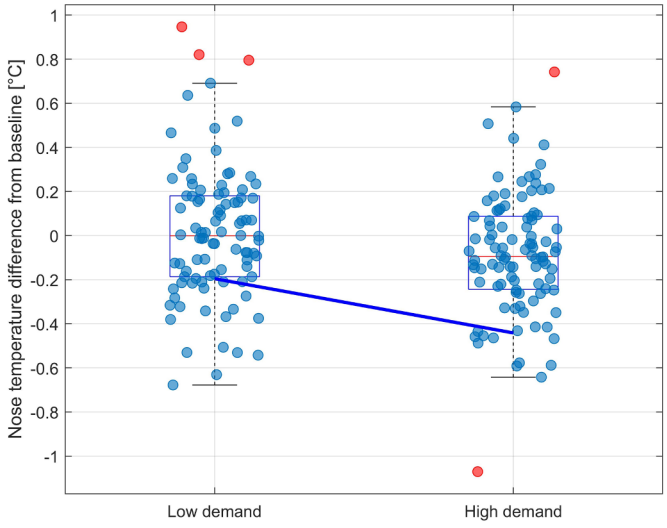

Fig. 7. Demand level had a significant effect on nose temperature difference from baseline

compared to the low perceptual load one, supporting Hypothesis H1. These findings are similar to those described by Forster and Lavie (2009), who found that engagement with internal distractors was a function of a task's perceptual load.

It is worth noting, though, that mind wandering is a challenging phenomenon to observe. Although it has been demonstrated that TUT frequency can be manipulated by varying perceptual load (Forster and Lavie, 2009) and working memory load (Teasdale et al., 1993), it is difficult to assess reliably whether an individual's attention is directed towards task-related or TUTs. Equally, because the phenomenon is not directly activated by a manipulation in the protocol, it is difficult for participants to report how long internally-directed thoughts were, or even when they began and ended (such that we can mark those times in the data). Considering these challenges, it is interesting to find a significant result that confirms previous research indicating a relationship between perceptual load and TUTs. In addition, it is intriguing that the analysis of breathing rate and nose temperature revealed a significant difference between perceptual load conditions, albeit a difference accounting for a small amount of variance.

5.2. Effects of task demand and fatigue on physiological measures

Although participants' perceived differences in demand between the two conditions, demand did not produce a significant effect on the task-baseline difference for either heart rate or for the concentrations of HbO and HbR in both the prefrontal cortex and mid-temporal gyrus. This was contrary to Hypotheses H4 (heart rate) and H6 (fNIRS) and contrasts with literature that has described an association between increasing oxygenation in the PFC under higher demand conditions (Richter et al., 2009). Nevertheless, findings supported Hypotheses H5 and H7, with results indicating that as perceptual load increased, participants experienced a significant increase in the breathing rate task-baseline difference and a significant decrease the in nose temperature task-baseline difference. In the task-baseline difference for breathing rate analysis, breathing rate increased by a factor of 1.51 in the high demand condition as compared to the low demand condition. This finding aligns with those of Brookings et al. (1996) who observed a significant increase in breathing rate for pilots during increasingly complex flight tasks. Task-baseline difference for nose temperature decreased by a factor of -0.25 in the high demand compared to the low demand condition.

In the present work, findings demonstrated that task-baseline difference in HbO and HbR concentrations did not effectively distinguish between the significantly different levels of perceived MWL, evoked by manipulating perceptual load within the task. While the subjective data (both NASA-TLX responses and task performance data) reflected the expected changes of our primary independent variable manipulation, findings failed to reflect differences in the activation in the prefrontal cortex. In this case, the difference in the performance data, for both response time and error rate, indicated that responses to the NASA-TLX scale reflected workload differences, confirming the hypothesis that the contrasting levels of perceptual load would manipulate the perceived experience of MWL. This, in turn, indicates that although participants

perceived differences, the fNIRS measures did not detect a significant change in the prefrontal cortex or middle temporal gyrus (although we would not, perhaps, expect activation in the middle temporal gyrus based upon the existing literature).

In light of these findings, it is possible that, despite evidence within the literature implicating the PFC and middle temporal gyrus in mind wandering (Durantin et al., 2015), the prefrontal cortex was not necessarily involved in the visual inspection task. It is possible that activation was more prominent elsewhere in the brain, or across multiple brain areas, such as the default mode network (Christoff et al., 2009; Durantin et al., 2015). However, the prefrontal cortex, in particular, has been shown to experience increased oxygenation when an individual is exposed to higher workload (Maior et al., 2018), and given that the current work identified a significant effect of task demand on perceived MWL, one could make the assumption that similar trends would also be observed in the PFC. This finding raises several intriguing questions for future research, particularly in relation to exploring the mapping between the embodied experience of workload as reported subjectively by individuals and the response of physiological systems during these experiences. For those interested in the relationship between subjective ratings of MWL and brain-activation, future work should explore the degree to which subjective measures capture experiences associated with activation in multiple regions of the brain, rather than specifically in prefrontal cortex, where active executive thought typically occurs. Although some work already exists that compares how e.g. visual and verbal tasks differently affect cognitive concepts, task performance, and physiological measures (Klingner et al., 2011; Yekhshatyan and Lee, 2012), it would be of great value to the human factors and human-computer interaction research communities to investigate the specific relationship among different types of tasks, regions of brain activation, and subjective ratings of MWL.

Another possibility explaining the fNIRS results may be that the measures were not sufficiently sensitive to measure the difference between these two levels of task difficulty, and that fNIRS is more suitable for detecting larger differences. While we chose a sustained attention task, similar difficulties in using physiological responses to recognise small differences in MWL have been seen in other task manipulations. Mandrick et al. (2016), for example, did not find differences in fNIRS measures between a 1-back and 2-back task, and other papers often use 1-back and 3-back as baselines for easy and difficult conditions, respectively. While Fishburn et al. (2014), for example, observed a linear increase in activation for n-back tasks, both Herff et al. (2014) and Ayaz et al. (2012) found very little difference between 0- to 2-back tasks, but clear differences in comparison with 3-back. For the future of our research, and indeed any work that is interested in using brain data to evaluate the workload involved in everyday work tasks, that the sensitivity of fNIRS to small task demand manipulations might pose a notable challenge. Such future work may need to complement small variations of MWL, with comparable data from more difficult conditions that clearly manipulate MWL.

Several significant findings were also observed in relation to fatigue during the visual inspection task. As fatigue, inferred from time on task, increased the heart rate task-baseline difference slowed significantly, supporting Hypothesis H8. Time on task was expressed in terms of block number [1-10], spanning approximately 45–50 min; although block time lengths were held constant, participants spent different amounts of time completing the subjective reports. Interestingly, we had also hypothesised that breathing rate would vary with fatigue (H9), as inferred from time on task, and a significant effect was observed in the positive direction. However, the data did not support all hypotheses, particularly in relation to hemodynamic response (H10) and nose temperature (H11). Similar to variations in task demand, no significant difference was identified as time progressed in the task-baseline differences in concentrations of HbO and HbR, neither for the PFC nor the middle temporal gyrus.

There were several aspects of the experimental design and analysis that may have limited the generalizability of our conclusions. In terms of experimental design, the study focused on a perceptual load task requiring sustained attention, performed in a stationary, computer-based scenario. One of the reasons behind this choice was to minimise the potential occurrence of motion artefacts in the fNIRS data. Thus, while the results provide insight into the physiological response to task demand variations in this type of situation, further research is needed to understand the degree to which this response generalizes to more dynamic work environments.

In terms of the analysis, it is important to note that the choice of pre-processing pipelines for the physiological data influences later interpretations of the data; for example, in fNIRS data, choice of filtering parameters may result in varying levels of motion artefacts left in the final data set. In the current work, the fNIRS pre-processing pipeline was based on one established within the literature (Pinti et al., 2019), but within the broader research community, there is as of yet no agreement as to best practice or a single recommended pipeline.

5.3. Open challenges for using physiological data for operator state assessment

As highlighted in this paper, we encountered multiple challenges in collecting physiological data during the study, which led to the removal of some data from the analysis. This section aims to clarify the reasons for removing data from the analysis as well as provide some guidance for other researchers willing to apply similar techniques; this may be of particular value as physiological measures are increasingly of interest for use in representative environments and even in situ.

Compared to the other measures, capturing heart rate and breathing rate was relatively straightforward, with challenges mainly relating to inaccurate readings caused by the chest strap losing contact with the skin. One way to address this is to provide participants with the right size for the strap as well as provide clear instructions on how this should be fitted.

Facial thermography data presented multiple challenges including the depth of field for the far and near limits between which the participant would be in focus, big variations in temperature due to the non-uniformity correction (NUC) and accuracy of facial landmark tracking. The depth of field is important as the accuracy of the temperature readings can not be guaranteed when the participant is out of focus. The main way of dealing with this is making sure to choose the right lens for the application: one that provides enough space between the near and far limits of the field of view to allow for the participant's movements. This needs to be balanced against the horizontal and vertical fields of view; the larger the difference between the near and far field of view, the larger the vertical and horizontal fields of view. This, in turn, may cause the image of the participant's face to be very small making landmark tracking difficult.

Thermal cameras apply a NUC in real time in order to update the offset correction coefficients, resulting in a better quality measurement; such occurrences can influence the final results of the statistics. Our approach was to remove all blocks that had an NUC during the baseline or task of that block. An alternative approach could have been to use a calibrated black body device set to a temperature similar to that of the human skin; these types of devices can be used to correct temperature measurements by the estimating the variation recorded by the camera in the area of the stable temperature black body device, however, it is not always practical to use this type of device, which may limit applications in live manufacturing environments.

Facial landmark tracking can be time-consuming to implement, and as an added challenge, facial landmark algorithms trained on visual images may not always perform as well on thermal images. For these reasons, our approach was to label a number of thermal images manually and perform transfer learning using DeepLabCut (Mathis et al., 2018; Nath* et al., 2019). Even so, tracking accuracy decreased at times and affected blocks were also removed. The neural

network training stage offers a lot of room for improvement, as labeling of additional images for training data and adjusting the parameters for training the neural network may offer opportunities to improve tracking accuracy.

In terms of the advances of fNIRS technology, even though great progress has been made with regards to their portability, as well as with the available data analysis methods in open source packages like Homer2, there are still challenges to overcome for both sustained data collection and less controlled task contexts. For example, optode positioning might vary slightly between participants due to different head sizes. For our data collection, this was reflected in the varied locations of the pre-determined sensor holes in the medium and large headcaps. Natural head-shape differences between people, though, mean that sensor placement also varies within the same headcap. fNIRS as a measure, however, is based on light scattering within a 2-3cm radius, and so the variation of placement has only a limited affect on activation of a cluster of sensors. In our experience, positioning optodes over the mid temporal gyrus was especially challenging due to the presence of hair, and in some cases data had to be discarded due to poor signal quality. In practical terms, recording changes in hairless areas, such as the forehead over the prefrontal cortex, is easier to implement. In addition, and perhaps most importantly for measurement of sustained tasks, wearing a headcap for a long duration (about 45 min) proved to be uncomfortable for some participants. The portability and fit of modern fNIRS devices would, in practice, need to be considered in the context of actual comfort and any negative impact on e.g. ability to concentrate.

Finally, filtering fNIRS data remains extremely challenging, especially in the presence of motion artifacts. The fNIRS community continues to discuss what data processing pipelines are broadly acceptable, and indeed required or recommended. Different approaches are better at removing e.g. motion artifacts and baseline shifts in data (created by headcaps shifting slightly on the head). Beyond being difficult, and still under debate within the research community, these challenges would become more critical in the contexts that our project wishes to study in working manufacturing conditions.

5.4. Implications for theory and practice

Overall, this work has provided insight into the utility of physiological measures for assessing cognitive state, and as such, has several implications from both a theoretical as well as a practical perspective. From a practical perspective, findings can inform future research into the assessment of fatigue and demand levels for humans at work, factors that, as we have demonstrated, closely related to MWL and attentional degradation. The current findings indicate that, among the measures considered here, breathing rate and nose temperature may be most effective for assessing demand, whereas heart rate and breathing rate may be most effective for assessing fatigue. This is of particular relevance to industries where safety and quality are critical to individual and organisational effectiveness. Indeed, our work in physiological assessment of operator state has been motivated by challenges arising within the manufacturing industry, where certain tasks and work environments impose variable demands and constraints upon human workers. While the current work focused on an abstracted quality control task involving visual inspection, we hypothesise that these findings may generalise to other predominantly visual tasks of similar natures, and may even overlap with similar challenges in other safety-critical domains, such as pilot state monitoring in aviation. Further research is needed to identify and exploit appropriate mappings between physiological measures and human factors constructs, and such work may have greater implications on the design of future work systems, particularly those incorporating human-automation interaction.

In addition to practical implications, this work has revealed several overarching questions related to the use of physiological assessment of human factors constructs and phenomena such as MWL, fatigue, and mind wandering. Although the literature contains evidence supporting a relationship between workload and physiological measures such as HbO concentration (Causse et al., 2017; Foy and Chapman, 2018; Maior et al., 2018) and heart rate (Bonner and Wilson, 2002), our work could not replicate these findings in the context of a visual inspection task. With regards to the fNIRS data, while it is possible that visual inspection and coping with variations in perceptual load were processed in an area of the brain not considered within this work, this offers a partial explanation, at best. Oxygenation concentration within the PFC has been shown to be associated with variations in MWL during a variety of tasks, but in the current work, significant differences in subjective perceptions of MWL between the two task demand conditions did not translate into clear differences in hemodynamic response. This leads us to recommend further investigation into novel methods for assessment of human factors constructs. Why some measures are more sensitive in certain situations remains an open question, and continued research is needed to explore the efficacy of different measures under varying work conditions. Furthermore, based on our findings, we argue that, while subjective measures of MWL have well-documented limitations (Marinescu et al., 2018; Sharples and Megaw, 2015), they still offer value in terms of understanding a highly complex phenomenon. It is generally accepted that MWL is a function of not only physical and cognitive demands, but also external influences and an individual's background experience and perceptions of the work at hand (Charles and Nixon, 2019; Sharples and Megaw, 2015). We challenge the assumption that physiological indicators inherently provide a more accurate and objective way to assess such constructs given their innate complexities, but we also encourage further exploration and critical evaluation of their use in order to further the scientific debate on human-centred sensing and its applications for supporting individuals at work.

Lastly, lessons learned during this research have provided several insights into effective experimental design and data collection protocols that would perhaps be most interesting for others running studies with physiological measures over prolonged periods of time. One typical challenge is identifying acceptable pre-task baseline conditions; because parasympathetic and circadian rhythm changes over time, this means that baseline recordings also change over time and can also be affected by on sedentary behaviour, tiredness, time of day, and time since calorie intake. Specific to fNIRS protocols, drift is often algorithmically removed from the data, such that increases and decreases are produced by brain activation rather than longitudinal rhythms in oxygenation.

6. Conclusions

This research contributes to the understanding of the relationship between perceptual load, MWL, fatigue, task unrelated tasks, their effects on physiological response as well as facilitating a discussion about the sensitivity of physiological measures and the challenges in applying them in human factors studies. The subjective rating results obtained in our study showed that participants perceived a significant difference in terms of MWL level between the two demand conditions. Performance data also revealed significant differences between the two demand conditions in terms of error rate and response times, indicating to the fact that demand was well manipulated. Findings also showed that the frequency of TUTs was lower in the high demand condition.

The physiological data results indicate that demand did not have a significant effect on HbO and HbR concentrations in either the PFC or MTG areas nor on heart rate. Breathing rate and nose temperature were the two physiological measures that presented significant differences between the two demand conditions, with breathing rate increasing and nose temperature slightly decreasing. In terms of fatigue having an effect on physiological measures, as time progressed, breathing rate showed an increase as compared to the baseline, heart rate decreased

relative to the baseline, while no effect was observed on nose temperature or changes in HbO or HbR concentrations in the PFC or MTG areas.

This result is important as it potentially indicates that fNIRS might not be sensitive to such small changes in workload. Sensing brain activity using fNIRS seems to have high face validity as it returns a measure directly related to changes that occur in the brain. Nevertheless, it is still a recent field of research and a lot of progress still needs to be made in understanding this measure. Researchers planning to use fNIRS in more naturalistic human factors studies should consider the limitations on the study design. Although our findings contrast with some of the literature, our work suggests that, for our task, breathing rate and nose temperature may be most effective out of the measures we considered for assessing demand, whereas heart rate and breathing rate may be most effective for estimating fatigue. Physiological measures may vary in terms of their level of sensitivity to fine variations in human factors constructs such as MWL and fatigue, and this should be considered when determining whether and how to implement human-centred sensing.

CRediT authorship contribution statement

Elizabeth M. Argyle: Conceptualization, Methodology, Software, Formal analysis, Data curation, Investigation, Writing - original draft, Writing - review & editing. **Adrian Marinescu:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Visualization, Investigation, Writing - original draft, Writing - review & editing. **Max L. Wilson:** Conceptualization, Methodology, Writing - review & editing, Resources, Supervision, Project administration, Funding acquisition. **Glyn Lawson:** Conceptualization, Writing - review & editing, Resources, Supervision, Project administration, Funding acquisition. **Sarah Sharples:** Conceptualization, Writing - review & editing, Resources, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/R032718/1]. The authors would like to thank all their collaborators on the DigiTOP ("Digital Toolkit for optimisation of operators and technology in manufacturing partnerships") project for their constructive feedback and suggestions during the development of this study. We also wish to thank Dr Robert Houghton for his guidance in the early stages of the project and for sharing the stimuli used in the experimental task.

References

Alsuraykh, N.H., Maior, H.A., Wilson, M.L., Tennent, P., Sharples, S., 2018. How stress affects functional near-infrared spectroscopy (fNIRS) measurements of mental workload. Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1-6. https://doi.org/10.1145/3170427.3188646.

Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglu, K., Willems, B., Onaral, B., 2012. Optical brain monitoring for operator training and mental workload assessment. Neuroimage 59 (1), 36–47. https://doi.org/10.1016/j.neuroimage.2011.06.023.

Baldauf, D., Burgard, E., Wittmann, M., 2009. Time perception as a workload measure in simulated car driving. Appl. Ergon. 40 (5), 929–935.

Bonner, M.A., Wilson, G.F., 2002. Heart rate measures of flight test and evaluation. Int. J. Aviat. Psychol. 12 (1), 63–77.

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

Brouwer, A.-M., Hogervorst, M.A., Van Erp, J.B., Heffelaar, T., Zimmerman, P.H., Oostenveld, R., 2012. Estimating workload using EEG spectral power and ERPs in the n-back task. J. Neural Eng. 9 (4), 045008.

Casali, J.G., Wierwille, W.W., 1983. A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. Hum. Factors 25 (6), 623–641.

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., Matton, N., 2017. Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. Sci. Rep. 7 (1), 5222. https://doi.org/10.1038/s41598-017-05378-x.

Charles, R.L., Nixon, J., 2019. Measuring mental workload using physiological measures: a systematic review. Appl. Ergon. 74, 221–232.

Chen, S., Epps, J., 2014. Using task-induced pupil diameter and blink rate to infer cognitive load. Hum.–Comput. Interact. 29 (4), 390–413.

Christoff, K., Gordon, A.M., Smallwood, J., Smith, R., Schooler, J.W., 2009. Experience sampling during fMRI reveals default network and executive system contributions to mind wandering. Proc. Natl. Acad. Sci. 106 (21), 8719–8724.

Collet, C., Salvia, E., Petit-Boulanger, C., 2014. Measuring workload with electrodermal activity during common braking actions. Ergonomics 57 (6), 886–896. https://doi.org/10.1080/00140139.2014.899627.

Comstock Jr., J. R., Arnegard, R. J., 1992. The multi-attribute task battery for human operator workload and strategic behavior research.

Dell'Agnola, F., Cammoun, L., Atienza, D., 2018. Physiological characterization of need for assistance in rescue missions with drones. 2018 IEEE International Conference on Consumer Electronics (ICCE). IEEE, pp. 1–6.

Diaz-Piedra, C., Gomez-Milan, E., Di Stasi, L.L., 2019. Nasal skin temperature reveals changes in arousal levels due to time on task: an experimental thermal infrared imaging study. Appl. Ergon. 81, 102870.

Dumontheil, I., Gilbert, S.J., Frith, C.D., Burgess, P.W., 2010. Recruitment of lateral rostral prefrontal cortex in spontaneous and task-related thoughts. Q. J. Exp. Psychol. 63 (9), 1740–1756.

Durantin, G., Dehais, F., Delorme, A., 2015. Characterization of mind wandering using fNIRS. Front. Syst. Neurosci. 9, 45. https://doi.org/10.3389/fnsys.2015.00045.

Edwards, T., Gabets, C., Mercer, J., Bienert, N., 2017. Task demand variation in air traffic control: implications for workload, fatigue, and performance. Advances in Human Aspects of Transportation. Springer, pp. 91–102.

Edwards, T., Sharples, S., Wilson, J.R., Kirwan, B., 2012. Factor interaction influences on human performance in air traffic control: the need for a multifactorial model. Work 41 (Supplement 1), 159–166.

Fairclough, S.H., Venables, L., 2006. Prediction of subjective states from psychophysiology: a multivariate approach. Biol. Psychol. 71 (1), 100–110.

Fishburn, F.A., Norr, M.E., Medvedev, A.V., Vaidya, C.J., 2014. Sensitivity of fNIRS to cognitive state and load. Front. Hum. Neurosci. 8, 76. https://doi.org/10.3389/fnhum.2014.00076.

Forster, S., Lavie, N., 2009. Harnessing the wandering mind: the role of perceptual load. Cognition 111 (3), 345–355. https://doi.org/10.1016/j.cognition.2009.02.006.

Foy, H.J., Chapman, P., 2018. Mental workload is reflected in driver behaviour, physiology, eye movements and prefrontal cortex activation. Appl. Ergon. 73, 90–99. https://doi.org/10.1016/j.apergo.2018.06.006.

Fridman, L., Reimer, B., Mehler, B., Freeman, W.T., 2018. Cognitive load estimation in the wild. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, pp. 652. https://doi.org/10.1145/3173574.3174226.

Genno, H., Ishikawa, K., Kanbara, O., Kikumoto, M., Fujiwara, Y., Suzuki, R., Osumi, M., 1997. Using facial skin temperature to objectively evaluate sensations. Int. J. Ind. Ergon. 19 (2), 161–171.

Gilbert, S., Bird, G., Frith, C., Burgess, P., 2012. Does "task difficulty" explain "task-induced deactivation?". Front. Psychol. 3, 125.

Girouard, A., Solovey, E.T., Hirshfield, L.M., Chauncey, K., Sassaroli, A., Fantini, S., Jacob, R.J., 2009. Distinguishing difficulty levels with non-invasive brain activity measurements. Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part I. Springer-Verlag, Berlin, Heidelberg, pp. 440–452. https://doi.org/10.1007/978-3-642-03655-2_50.

Hart, S.G., 2006. Nasa-task load index (NASA-TLX); 20 years later. Proceedings of the human factors and ergonomics society annual meeting. vol. 50. Sage publications Sage CA: Los Angeles, CA, pp. 904–908. https://doi.org/10.1177/154193120605000909.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., Schultz, T., 2014. Mental workload during n-back task-quantified in the prefrontal cortex using fNIRS. Front. Hum. Neurosci. 7, 935. https://doi.org/10.3389/fnhum.2013.00935.

Hermawati, S., Lawson, G., D'Cruz, M., Arlt, F., Apold, J., Andersson, L., Lövgren, M.G., Malmsköld, L., 2015. Understanding the complex needs of automotive training at final assembly lines. Appl. Ergon. 46, 144–157.

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

Kang, J., Babski-Reeves, K., 2008. Detecting mental workload fluctuation during learning of a novel task using thermography. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 52. SAGE Publications Sage CA: Los Angeles, CA, pp. 1527–1531.

Killingsworth, M.A., Gilbert, D.T., 2010. A wandering mind is an unhappy mind. Science 330 (6006). https://doi.org/10.1126/science.1192439. 932–932

Klingner, J., Tversky, B., Hanrahan, P., 2011. Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. Psychophysiology 48 (3), 323–332.

Langner, R., Eickhoff, S.B., 2013. Sustaining attention to simple tasks: a meta-analytic review of the neural mechanisms of vigilant attention. Psychol. Bull. 139 (4), 870.

Lavie, N., 1995. Perceptual load as a necessary condition for selective attention. J. Exp. Psychol. 21 (3), 451.
Lehrer, P., Karavidas, M., Lu, S.-E., Vaschillo, E., Vaschillo, B., Cheng, A., 2010. Cardiac data increase association between self-report and both expert ratings of task load and task performance in flight simulator tasks: an exploratory study. Int. J. Psychophysiol. 76 (2), 80–87.
Magnusson, S., 2002. Similarities and differences in psychophysiological reactions between simulated and real air-to-ground missions. Int. J. Aviat. Psychol. 12 (1), 49–61.
Maior, H.A., Pike, M., Sharples, S., Wilson, M.L., 2015. Examining the reliability of using fNIRS in realistic HCI settings for spatial and verbal tasks. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 3039–3042. https://doi.org/10.1145/2702123.2702315.
Maior, H.A., Wilson, M.L., Sharples, S., 2018. Workload alerts & using physiological measures of mental workload to provide feedback during tasks. ACM Trans. Comput.-Hum. Interact. 25 (2), 1–30. https://doi.org/10.1145/3173380.
Mandrick, K., Peysakhovich, V., Remy, F., Lepron, E., Causse, M., 2016. Neural and psychophysiological correlates of human performance under stress and high mental workload. Biol. Psychol. 121, 62–73. https://doi.org/10.1016/j.biopsycho.2016.10.002.
Marinescu, A.C., Sharples, S., Ritchie, A.C., Sanchez Lopez, T., McDowell, M., Morvan, H.P., 2018. Physiological parameter response to variation of mental workload. Hum. Factors 60 (1), 31–56. https://doi.org/10.1177/0018720817733101.
Mason, M.F., Norton, M.I., Van Horn, J.D., Wegner, D.M., Grafton, S.T., Macrae, C.N., 2007. Wandering minds: the default network and stimulus-independent thought. Science 315 (5810), 393–395. https://doi.org/10.1126/science.1131295.
Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M., 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat. Neurosci.
Mehta, R.K., Parasuraman, R., 2013. Effects of mental fatigue on the development of physical fatigue: a neuroergonomic approach. Hum. Factors 56 (4), 645–656. https://doi.org/10.1177/0018720813507279.
Murai, K., Hayashi, Y., Okazaki, T., Stone, L.C., Mitomo, N., 2008. Evaluation of ship navigator's mental workload using nasal temperature and heart rate variability. 2008 IEEE International Conference on Systems, Man and Cybernetics. IEEE, pp. 1528–1533.
Naemura, A., Tsuda, K., Suzuki, N., 1993. Effects of loud noise on nasal skin temperature. Jap. J.Psychol. 64 (1), 51–54.
Nath*, T., Mathis*, A., Chen, A.C., Patel, A., Bethge, M., Mathis, M.W., 2019. Using deeplabcut for 3d markerless pose estimation across species and behaviors. Nat. Protoc.
Naweed, A., 2013. Psychological factors for driver distraction and inattention in the australian and new zealand rail industry. Accid. Anal. Prev. 60, 193–204. https://doi.org/10.1016/j.aap.2013.08.022.
Or, C.K., Duffy, V.G., 2007. Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement. Occup. Ergon. 7 (2), 83–94.
Parasuraman, R., Mehta, R., 2015. Neuroergonomic methods for the evaluation of physical and cognitive work. Eval. Hum. Work 609–640.
Peirce, J.W., Gray, J.R., Simpson, S., MacAskill, M.R., Hochenberger, R., Sogo, H., Kastman, E., Lindelov, J., 2019. PsychoPy2: experiments in behavior made easy. Behav. Res. Methods 51 (1), 195–203. https://doi.org/10.3758/s13428-018-01193-y.
Pike, M.F., Maior, H.A., Porcheron, M., Sharples, S.C., Wilson, M.L., 2014. Measuring the effect of think aloud protocols on workload using fNIRS. Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems. ACM, New York, NY, USA, pp. 3807–3816. https://doi.org/10.1145/2556288.2556974.
Pinti, P., Scholkmann, F., Hamilton, A., Burgess, P., Tachtsidis, I., 2019. Current status and issues regarding pre-processing of fNIRS neuroimaging data: an investigation of diverse signal filtering methods within a general linear model framework. Front. Hum. Neurosci. 12, 505.
Richter, M.M., Zierhut, K.C., Dresler, T., Plichta, M.M., Ehlis, A.-C., Reiss, K., Pekrun, R., Fallgatter, A.J., 2009. Changes in cortical blood oxygenation during arithmetical tasks measured by near-infrared spectroscopy. J. Neural Transm. 116 (3), 267–273.
Santiago-Espada, Y., Myer, R. R., Latorella, K. A., Comstock Jr, J. R., 2011. The multi-attribute task battery ii (MATB-ii) software for human performance and workload research: a user's guide.
Schooler, J.W., Smallwood, J., Christoff, K., Handy, T.C., Reichle, E.D., Sayette, M.A., 2011. Meta-awareness, perceptual decoupling and the wandering mind. Trends Cogn. Sci. 15 (7), 319–326. https://doi.org/10.1016/j.tics.2011.05.006.
Sharples, S., Megaw, T., 2015. The Definition and Measurement of Human Workload, fourth. CRC Press, Boca Raton, FL, USA.
Shi, Y., Ruiz, N., Taib, R., Choi, E., Chen, F., 2007. Galvanic skin response (GSR) as an index of cognitive load. CHI'07 Extended Abstracts on Human Factors in Computing Systems. pp. 2651–2656.
Smallwood, J., Andrews-Hanna, J., 2013. Not all minds that wander are lost: the importance of a balanced perspective on the mind-wandering state. Front. Psychol. 4, 441. https://doi.org/10.3389/fpsyg.2013.00441.
Smallwood, J., Schooler, J., 2006. The restless mind. Psychol. Bull. 132 (6), 946–958. https://doi.org/10.1037/0033-2909.132.6.946.
Solovey, E.T., Girouard, A., Chauncey, K., Hirshfield, L.M., Sassaroli, A., Zheng, F., Fantini, S., Jacob, R.J., 2009. Using fNIRS brain sensing in realistic HCI settings: experiments and guidelines. Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology. ACM, New York, NY, USA, pp. 157–166. https://doi.org/10.1145/1622176.1622207.
Svensson, E.A., Wilson, G.F., 2002. Psychological and psychophysiological models of pilot performance for systems development and mission evaluation. Int. J. Aviat. Psychol. 12 (1), 95–110.
Tattersall, A.J., Foord, P.S., 1996. An experimental evaluation of instantaneous self-assessment as a measure of workload. Ergonomics 39 (5), 740–748. https://doi.org/10.1080/00140139608964495.
Teasdale, J.D., Proctor, L., Lloyd, C.A., Baddeley, A.D., 1993. Working memory and stimulus-independent thought: effects of memory load and presentation rate. Eur. J. Cogn. Psychol. 5 (4), 417–433.
Verwey, W.B., Veltman, H.A., 1996. Detecting short periods of elevated workload: acomparison of nine workload assessment techniques. J. Exp. Psychol. 2 (3), 270.
Villringer, A., Planck, J., Hock, C., Schleinkofer, L., Dirnagl, U., 1993. Near infrared spectroscopy (NIRS): a new tool to study hemodynamic changes during activation of brain function in human adults. Neurosci. Lett. 154 (1–2), 101–104.
Wang, X., Li, D., Menassa, C.C., Kamat, V.R., 2019. Can infrared facial thermography disclose mental workload in indoor thermal environments? Proceedings of the 1st ACM International Workshop on Urban Building Energy Sensing, Controls, Big Data Analysis, and Visualization. ACM, pp. 87–96.
Warm, J. S., Parasuraman, R., 2006. Cerebral hemodynamics and vigilance.
Wilson, G.F., 2002. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. Int. J. Aviat. Psychol. 12 (1), 3–18.
Yekhshatyan, L., Lee, J.D., 2012. Changes in the correlation between eye and steering movements indicate driver distraction. IEEE Trans. Intell. Transp.Syst. 14 (1), 136–145.
Young, M.S., Brookhuis, K.A., Wickens, C.D., Hancock, P.A., 2015. State of science: mental workload in ergonomics. Ergonomics 58 (1), 1–17. https://doi.org/10.1080/00140139.2014.956151.
Zhang, Y., Kumada, T., 2017. Relationship between workload and mind-wandering in simulated driving. PloS One 12 (5).