Workload Alerts—Using Physiological Measures of Mental Workload to Provide Feedback During Tasks

HORIA A. MAIOR, MAX L. WILSON, and SARAH SHARPLES, University of Nottingham

Feedback is valuable for allowing us to improve on tasks. While retrospective feedback can help us improve for next time, feedback 'in action' can allow us to improve the outcome of on-going tasks. In this article, we use data from functional Near InfraRed Spectroscopy to provide participants with feedback about their mental workload levels *during* high-workload tasks. We evaluate the impact of this feedback on task performance and perceived task performance, in comparison to industry standard mid-task self-assessments, and explore participants' perceptions of this feedback. In line with previous work, we confirm that deploying self-reporting methods affect both perceived and actual performance. Conversely, we conclude that our objective concurrent feedback correlated more closely with task demand, supported reflection in action, and did not negatively affect performance. Future work, however, should focus on the *design* of this feedback and the potential behaviour changes that will result.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**;

Additional Key Words and Phrases: Mental workload, fNIRS, physiological computing, feedback, task demand, performance, bio-feedback of workload

ACM Reference format:
Horia A. Maior, Max L. Wilson, and Sarah Sharples. 2018. Workload Alerts—Using Physiological Measures of Mental Workload to Provide Feedback During Tasks. *ACM Trans. Comput.-Hum. Interact.* 25, 2, Article 9 (April 2018), 30 pages.
https://doi.org/10.1145/3173380

1 INTRODUCTION

Although computers are very good at performing repetitive rule-based tasks, humans can excel at 'knowledge-based' tasks that involve decision-making processes, generating new solutions for rapidly changing problems, and developing solutions based on past experience and innovation. Consequently, as technology pervades our everyday life, our own tasks are increasingly 'dominated by mental rather than physical task components' [49]; the human role has moved towards a supervisory and decision-making role for such intelligent systems. This move has the potential to increase the demands on people's mental resources due to the amounts of data being generated and the amount of concurrent tasks and decisions we take every day. Sharples and Megaw [49] describe mental workload as 'the relationship between primary task performance and the resources demanded by the primary task', identifying operator overload as a root cause

This work was supported by the Engineering and Physical Sciences Research Council (Grant numbers EP/G037574/1, EP/M000877/1, and EP/N50970X/1).

Authors' addresses: H. A. Maior, M. L. Wilson, and S. Sharples, School of Computer Science, University of Nottingham, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB; emails: {horia.maior, max.wilson, sarah.sharples}@nottingham.ac.uk.

ACM 1073-0516/2018/04-ART9 $15.00
https://doi.org/10.1145/3173380

of reduced performance – humans have a limited capacity when it comes to the number of things they can manage at any one moment. Assessing, predicting, and managing peaks of mental workload, therefore, can become particularly important, especially in the context of safety-critical domain such as Air Traffic Control (ATC).

In Human Computer Interaction (HCI), we are concerned with understanding users' capabilities and limitations in terms of their mental workload, in order to assess the demands placed upon them while interacting with computer-based systems. While some work focuses on *adapting the task* that users need to perform [1, 64], this work focuses on making participants more aware of their workload so that they can better manage, e.g., non-adaptable tasks. An individual's awareness of workload, therefore, would potentially allow them to regulate their resource allocation to the primary task, therefore avoiding risky high (and significantly low) workload conditions as described above.

There are a variety of subjective and objective methods used for measuring mental workload, including primary and secondary task analysis [40], physiological or psycho-physiological techniques [24, 33, 44, 52], as well as user opinions using subjective techniques [23, 31]. The most commonly used method for mid-task workload monitoring in industry, as a form of secondary task analysis, is Instantaneous Self-Assessment (ISA) and relies on people self-reporting their levels of workload at time intervals. Self-reporting in this way is used successfully, but has also been shown to either increase workload or negatively affect task performance itself [55]. In this article, we evaluate the use of functional Near Infra-Red Spectroscopy (fNIRS) data to give people feedback of their current mental workload levels (MWLs) during tasks, as a form of unobtrusive objective feedback, in comparison to ISA as an intrusive subjective self-assessment. Such objective measures have become more realistic in natural task conditions, as fNIRS, unlike other brain-sensing techniques as follows: (a) particularly resilient to artefacts associated with natural computer usage and (b) sensitive to changes in mental workload [44, 52]. Further, by presenting feedback about current MWLs, we also explore initial insights into how participants monitor their own performance, and respond to the feedback.

In the following sections, we first review related work on working memory, mental workload, and the methods used to measure workload, meta-cognition, and feedback. We then present the workload feedback mechanism developed and the methods used to assess workload during tasks. The article continues by describing a study comparing the impact of workload feedback to a traditional method of asking users to self-assess and report their own mental workload. We then present the results of the study, discuss the findings in terms of what we can learn about feedback of workload in general, and recommendations for further exploring the *design* of feedback and potential user behaviour changes.

2 RELATED WORK

This section presents four key areas of related work: (1) mental workload literature, (2) techniques for measuring workload, (3) measuring workload with brain sensors, including fNIRS, and (4) key literature on reflection and meta-cognition.

2.1 The Concept of Mental Workload

In order to discuss and better understand mental workload, we review (a) literature that underlies the concept and (b) how it relates to task demand and performance.

2.1.1 Background Literature on Mental Workload. At its origins, mental workload is grounded in Baddeley's well-established model of working memory (WM) [7], which includes a central executive process that manages the use of three 'slave' systems: (1) a visuo-spatial sketch pad holding information in an analogue spatial form (e.g., colours, shapes, and maps), (2) a phonological loop holding verbal information in an acoustical form (e.g., numbers and words), and (3) an episodic buffer dedicated to linking verbal and spatial information in chronological order [6]. 'The concept of working memory proposes that a dedicated system maintains and stores information in the short term, and that this system underlies human thought processes' (Baddeley [7]). These concepts are well established, and appear in similar models by Welford [59] and Whiting [60], both models reflecting the same process, however, using slightly different terminology.

One shared characteristic in most models of working memory is the *limited capacity* that we have as human beings, meaning that we can only process a limited amount of information at any one time. The Limited Resource Model (LRM), more recently summarised by Sharples and Megaw [49], highlights the impact of the limited resources on experienced workload and performance. The authors present the concept of mental workload as the relationship between (a) the resources allocated to the primary task, (b) the resources demanded by the primary task, and (c) how performance is 'affected' at different stages of demand. Workload, therefore, is a concept that refers to the amount of resources and necessary 'effort' required by all the processes mentioned above in relation to the demands of a task. Similar notions of limited capacity for a task are also noted by Cognitive Load Theory [41], as a similar concept developed in the context of educational learning materials. Based on this principle, work in HCI has considered adaptive technologies that try to dynamically match the demand of a task with the users current mental workload. Yuksel et al. [64], for example, increased the difficulty of piano music as the participants needed less mental effort to play their current piece.

Although the LRM captures the limitations of our combined mental resources, one limitation is that it does not capture the different subsystems of Baddeley's view of WM. Wickens proposed, therefore, the Multiple Resource Model (MRM) [61], which highlights that verbal and spatial modalities are handled separately at all stages, from perception, to processing (in WM), and during response. In principle, this means that two tasks that utilise different resources (one visual and one verbal) would utilise separate resources and thus mask or confound the experiences of underload and overload. Again, this similar concern is present in the design of learning materials in Cognitive Load Theory [41]. In the MRM, however, the complex interaction described between stages means that any study of workload has to be considerate of which resources are being demanded. In a related HCI work, Maior et al. showed that different physical artefacts affect spatial and verbal mental workloads differently [39]. In this study, we are focusing on a purely spatial task, but the inclusion of self-reporting workload with the ISA scale (described later) might be considered a verbal task. We discuss this complex interaction later in the article.

2.1.2 Mental Workload, Task Demands, and Performance.

To better understand the implications for evaluating mental workload, in relation to the study presented in this article, we use the framework in Figure 1 [49] to formalise the concept and set our hypotheses.

The physical and cognitive task demands reflect the characteristics of a task undertaken by a person, and thus imposed on them cognitively. To measure workload, therefore, we must quantify the work demands the participant faces. As the demands may have different impact on different individuals, it is important to not only capture the externally imposed demands, but also consider measuring the perceived demand.

Operator workload is 'equivalent to measures of operator strain or effort'. A lot of the workload measurement techniques, therefore, simply focus on capturing the operator experience during and after the task via subjective questionnaires like NASA-TLX (Task Load Index) [23], but measures of effort from behaviour indices and the impact of effort on operators' physiology should also be measured.

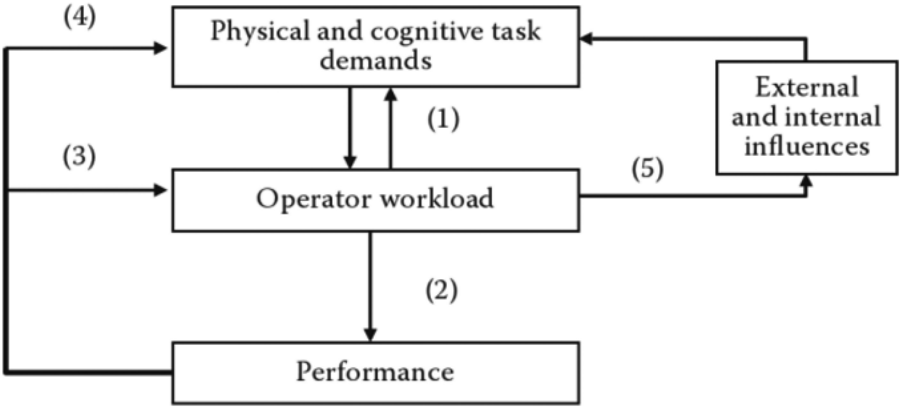

(4)
Physical and cognitive task demands
(1)
External and internal influences
(3)
(5)
Operator workload
(2)
Performance

Fig. 1. A framework for mental workload definition and evaluation [49].

Performance refers to the measures often described in terms of speed and error rates. However, performance measures can become problematic as task complexity increases, but also when closely analysing the relationship between the three mental workload components of demand, workload and performance: 'contrary to what is expected, as task demands increase it is not compulsory to expect an increase in operator workload, or decrease in task performance'.

External and internal influences relate to other factors that can affect workload, such as secondary tasks or interruptions. It is these external influences that we must consider when we change people's environments during tasks or ask them to perform self-assessments as a secondary task.

Sharples and Megaw also identify the five key relationships between these components (numbered in Figure 1):

(1) Operator workload is influenced by how the task is perceived by the operator, and it is not just a simple relationship between demand and workload. It can be seen as a consequence of demand created by not only the task demands, but also by a combination of external and internal influences [43]. Pickup et al. presents workload being influenced by intrinsic factors such as operator skill and amount of training and attitude towards a task. Therefore, these intrinsic factors can influence the strategies the operator can take towards performing a task and the workload perceived by the operator, and indicate that measures of operator workload will not necessarily reflect an objective measure of task demand. In this study, we will investigate the factors that influence operator workload during tasks, by combining objective and subjective measures of workload in contrast with measures of task demands.

(2) Although there is an expected relationship between operator workload and performance, with higher workload associated with relatively poorer performance, this is not always the case. Sharples and Megaw discuss that even with a highly sensitive performance measure, it is likely to be difficult to determine differences in how hard an individual is working to maintain a good level of performance. User performance will be monitored during this study, and insights into the relationship between workload and performance will be presented.

(3) Feedback, both unconscious and explicit, allows operators to monitor their own performance. This may change the way they perceive a task, but also their decision making, strategy, and attitudes towards a task. Additional to the performance, this study is focused on understanding the impact of workload feedback on operators, therefore feedback is further discussed later in the section.

(4) Performance outcomes may impact task demands. An error that may have occurred due to high demands on the operator, can lead to subsequent task demands to increase, this further increasing the demands placed upon the operator. On the other hand, a good performance may lead to lower demands to the subsequent tasks.
(5) Sharples and Megaw present this relationship, and the whole framework, in the context of work and therefore describe most of the external and internal influences as factors from a workspace perspective. We will consider external factors that will influence the operator's behaviour and experience or perception of workload, but also internal factors, skill, and motivation when drawing hypothesis and conclusions in relation to this study.

2.2 Measuring Workload

Measuring mental workload has been a long-term research challenge, and multiple attempts from different authors have been made to establish the appropriate criteria for it. In this section, we discuss various empirical measures, which can be divided into primary and secondary measures, subjective and psycho-physiological techniques.

2.2.1 Primary and Secondary Task Measures. Primary task measures of workload rely on evaluating the performance of the primary task to estimate changes in operator workload. They have limitations, however, when being used alone: it is difficult to discriminate between levels of effort the operator is going through while the demand changes and performance does not. Consequently, primary task measures should be combined with other workload measures. Secondary task techniques involve the inclusion of an additional task to the primary one, where allocating more resources to the primary task might maintain performance, but reduce their capacity to handle a second task. Gwizdka, for example, used the Stroop test as a secondary task analysis to measure spare cognitive capacity [21]. These secondary measures are used, therefore, in cases where the primary task demands would allow enough available resources for a secondary task to be completed concurrently.

2.2.2 Subjective Measures of Mental Workload. Subjective measures of mental workload attempt to capture users' reflection and perspective on how much effort was expelled in completing the task. They are currently the most popularly used and accepted techniques due to their nature of being easy to administer, cheap, and with high face validity. NASA-TLX, developed by Hart and Staveland [23], is an example of a widely used subjective measure of workload. NASA-TLX is a multi-dimensional tool that uses perceived workload ratings in order to assess a task after performing it [37, 44, 46]. This makes the measure suitable for providing an overview of the task retrospectively, however, it does not provide insight into users workload at a given moment during the task.

Concurrent alternatives to NASA-TLX exist, including ISA, developed by Jordan and Brennen [31] and have been validated as being a reliable workload measure [35, 55]. ISA was derived for use within ATC settings, and it is one of the most frequently used techniques for workload assessment in real-time simulations. ISA measures mental workload using a single five-point unidimensional rating scale, to provide immediate subjective ratings of work demands during the performance of primary work tasks. Users are prompted at regular time intervals during the primary task to rate their current workload levels, where performance in providing ISA ratings can be utilised as a secondary task measure. Naturally, however, any concurrent self-report measure can directly interfere with the performance of the primary task, especially when the operators are highly engaged in the primary task due to high task demands. There are conflicting findings in the existing literature: there are cases of detectable [55] and non-detectable [35] task intrusions. Regardless, however, the measure has been considered preferable to other measures, such as Subjective Workload Assessment Technique [47], and Workload profile [56], as it requires less mental effort and time from the operator due to its unidimensional nature.

2.2.3 Physiological Measures. Research has shown that task demands can induce complex and dynamic processes influencing a variety of physiological change [18]. Therefore, physiological measures can be used to give an objective perspective on mental workload by not relying on subjective scales or performance measures. Different physiological measures, however, capture various aspects of workload [12], therefore consideration should be put in choosing the most appropriate measure for the given task and setting. Wilson and Russell [62] presented different ways in which the physiological signals are known to change with the state of the operator: heart rate increases, for example, as the cognitive demands on the operator increase, and the rate of eye blinking decreases as the visual demands increases. Mental workload changes, in particular, have been obtained by recording: cardiac activity [10, 22, 62], electrodermal response and galvanic skin response [14, 50], eye function (pupil diameter) [9, 32], imaging the brain [8, 52], and even facial skin temperature [53]. As mental workload is a concept, each of these techniques essentially detect the changes in the autonomic nervous systems, which are then used to infer workload levels. Most related to this article, are physiological measures that focus directly on the brain. There are several key approaches to measuring mental workload with brain sensors, including Magnetic Resonance Imaging (MRI), electroencephalography (EEG), and more recently fNIRS. Experiments that have used MRI in a study typically place a mirror above the participant such that they can see a display in another room. Li et al. [36] for example, used real-time fMRI to control the animation speed of a virtual human runner. Normal direct computer use, however, is not practical in an fMRI scanner due to the large magnetic field and requirement that participants remain still, lying down inside it. EEG typically uses between 16 and 64 sensors on the scalp to detect varying electrical charge within the brain. With the introduction of commercially available Bluetooth EEG sensors, like the Emotiv,[1] EEG has become an affordable option for brain sensing [17]. Frey et al. [20] presented a review of the technology for the detection of workload, attention, vigilance, fatigue, error recognition, emotions, engagement, flow, and immersion as being recognisable by EEG. In a different study, Frey et al. [19] presented a framework for mental workload assessment in the context of HCI and user experience. For measurement of mental workload, however, EEG data is highly susceptible to motion artefacts, and so formal laboratory studies typically restrict movement.

fNIRS uses blood oxygenation, rather than electrical levels, for determining the activation of areas in the brain, where more blood flow indicates higher activity. Recent research has shown that fNIRS-related measurements of cortical oxygenation are less affected by body movement compared to techniques that measure the electrical activity of the brain (such as EEG). The reason EEG is more affected from motion artefacts is because of the dominant muscle activity related noise. Therefore, fNIRS could be a more appropriate brain-sensing technology for measuring mental workload in HCI [25, 34, 39, 42]. Because it takes several seconds for blood to flow to the brain [27, 58], fNIRS has been largely discounted for real-time interaction with systems [54]. However, the lightweight nature of the professional fNIRS scanners, described below, and its resilience to movement artefacts have made it popular for evaluation of mental workload in more ecologically valid contexts. Pike et al. presented a more detailed comparison of brain sensing technologies, with a focus on their suitability for measuring mental workload in HCI settings [44].

2.3 Using fNIRS to Measure Mental Workload in HCI

fNIRS is based on the use of near infrared spectroscopy (NIRS), introduced by Jobsis in 1977 [30], which started to be used later for functional brain imaging [13, 57, 58]. fNIRS uses near infrared

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

(NIR) light to measure regional hemodynamic responses associated with neuron behaviour, namely changes in blood volume and cortical oxygenation. This is possible due to the properties of our biological tissue, in our case the skull, that is relatively permeable to electromagnetic (EM) radiation of different frequencies and intensities. Light penetrates the skull well at NIR range, allowing the NIR light to reach different molecules that are known to absorb different wavelengths of EM radiation to different degrees (in this case light). For fNIRS imaging, the concerned molecule is haemoglobin, which is the oxygen carrier for the red blood cells.

fNIRS has been successfully used to measure brain activity in different brain regions, such as prefrontal cortex (PFC) [5], motor cortex [26], and auditory cortex [45]. In relation to mental workload, much research demonstrates that fNIRS can be used to reflect the demands placed upon individuals while performing a task by observing the relative changed in oxygen concentration in the PFC. fNIRS has been used in various tasks, including remotely operating vehicles [5, 16], mental arithmetic [44], *n*-back tasks [5, 16], and other complex cognition tasks such as video games [5, 11, 28].

HCI researchers are typically concerned with using fNIRS to assess workload, although some research works, for example, have used fNIRS to measure anger by detecting frontal asymmetry in the PFC [2]. Measuring mental workload with fNIRS, however, can then be used as an additional channel of information about users during interaction with technology. Recent research has proven fNIRS to be more suitable (compared to other brain sensing techniques) for assessing workload in HCI user studies [1, 39, 52, 63] due to the robustness in noisy environments. Lukanov et al. [38] used fNIRS to assess workload during usability testing of three versions of an insurance claim form. fNIRS has also been used for implicit input. Afergan et al. [1] used fNIRS as implicit input to control and dynamically adjust task difficulty based on user's mental state. Yuksel et al. [64] used the same technology for adapting learning during piano lessons. In this work, rather than designing a technology that adapts the task demands given to the user, we investigate how fNIRS can be used to provide users with feedback of their mental workload during tasks, such that they can better manage their workload.

2.4 Feedback and Reflection

Feedback allows us to review, reflect, and improve our performance, and in this study, we provide people with mid-task feedback about something that is typically hidden from observation: Mental workload. As our study is not focused on the design of feedback itself, but in studying how feedback on mental workload affects actual and perceived performance, we present some core literature in this area to provide a context for reading. Starting from knowledge, Schon et al. [48] presented the following properties of 'knowing':

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

To explain reflection, Schon then discusses the 'feeling' when doing something right or wrong. When one notices doing something good, that let him/her repeat the exact thing that he/she did before that proved successful, one would have a 'feeling' that would allow him/her to do that something again. Schon describes that 'studying the winning habits', makes us think about the 'know-how' that enabled us to win. This process of understanding and thinking about various patterns of actions, while we perform various tasks or after, can be referred to as reflecting on action and, in some cases, *reflecting in action*.

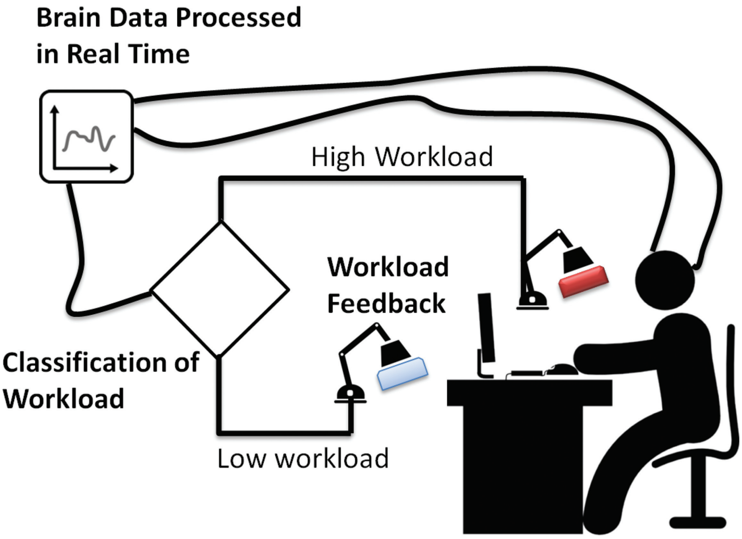


Fig. 2. Mental workload feedback mechanism.

> Reflecting-in-action. If common sense recognizes knowing-in-action, it also recognizes that we sometimes think about what we are doing. Phrases like 'thinking on your feet', 'keeping your wits about you,' and 'learning by doing' suggest not only that we can think about doing but that we can think about doing something while doing it. Some of the most interesting examples of this process occur in the midst of a performance. [48]

With this in mind, the study presented in this article investigates the impact of reflecting-in-action on mental workload and the resulting impact on meta-cognition and further on participants performance and perception. Meta-cognition is the state where one reflects upon one's thoughts, i.e., 'thinking about thinking'. Fletcher [18] showed that when in meta-cognitive states, users can monitor their performance, task cues and other states in order to assess their workload throughout the task and 'act' accordingly. One consideration of subjective measures is that rating your own mental workload will also make you aware of this information, potentially having an impact on your meta-cognitive state. The interest of this article in meta-cognition, is the potential for supporting such acts by presenting mental workload feedback, but in a way that does not require task interruption the same way subjective rating tools do.

3 WORKLOAD FEEDBACK SYSTEM, USING FNIRS

While Brain–Computer Interfaces (BCIs) were traditionally focused on users with disabilities, providing them direct control or interface with the outside world, current advances investigate the use of brain as an additional channel of information about healthy users interacting with technology. This 'passive' rather than 'active' channel, sometimes called implicit [51], can act as a complementary source of information about users' state, that can be combined with traditional methods, or used as an input to system, task, or interface. In this section, we present a Workload Feedback System (WFS – see Figure 2) that uses passive BCI to (1) measure, (2) detect, and (3) feedback users' workload during tasks.

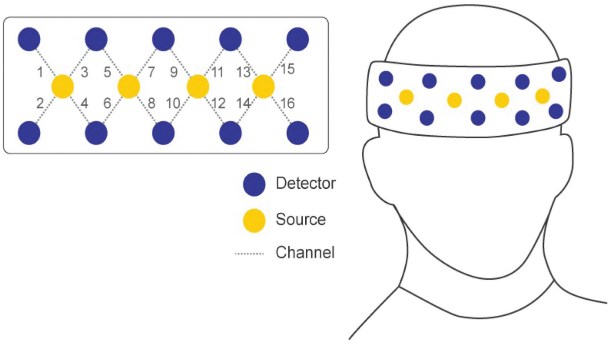


Fig. 3. Sensor layout for the Biopac fNIRS used.

3.1 Monitoring Mental Workload Using fNIRS

The WFS measures brain activity using an fNIRS300 device and the associated Cognitive Optical Brain Imaging (COBI) Studio hardware-integrated software platform provided by Biopac Systems, Inc. The headband shaped sensor, shown in Figure 3 [4], is a 16-channel transducer for continuous NIRS, placed on the forehead. The headband consists of four infrared (IR) emitters operating on a range between 700 and 900nm, and 10 IR detectors. Each channel is defined by the relationship between a IR source and a nearby IR detector pair. Oxygenated (HbO) and deoxygenated (Hb) haemoglobin are both strong absorbers of light, whereas skin, tissue, and bone are mostly transparent to NIR light, this property is typically referred to as the optical window [29]. The tissue is radiated by the light sources and the detectors receive the light after the interaction with the tissue.

For real-time use, COBI Studio requires collecting baseline signal levels that are used in order to calculate oxygenation in real-time, via the Modified Beer–Lambert Law [57]. The resulting data was used as an input for the WFS detection.

3.2 Detection of State

As our fNIRS device provides 16 locational channels of data with two readings per second, an important step before the study could began was identifying localised Hb/HbO changes for each participant. There are three reasons for this step: (1) fNIRS data is highly sensitive to individual differences between participants, (2) the physical placement of the 16 channels varies between participants (based on the shape and size of the forehead), and (3) different forms of workload create changes in Hb/HbO in different regions of the forehead [39]. We used training tasks to identify the most sensitive region, and identified the most valuable channel for the WFS system to focus on; the post-task evaluation, however, utilised recordings from all 16 channels.

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

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

3.3 Feedback Choice

Once the participant state was detected, the WFS produces a binary integer that can be used for changing the state of a feedback mechanism. For the purpose of this study, we specifically designed feedback to be noticeable, but at the same time transparent and in the background of the task, such that a minimum of resources would be used by operators. For our study, the output of the WFS was used to invoke changes in the desk lighting around the participant, using Philips Hue Bulbs (programmable light bulbs) in desk lamps. Initially, the lighting was set as normal white lighting, which would turn red when participants entered states detected by the fNIRS measurements to have high workload and return to white as participants returned to lower levels of mental workload. We discuss this colour choice further, however, in the study methodology below.

4 EXPERIMENT DESIGN

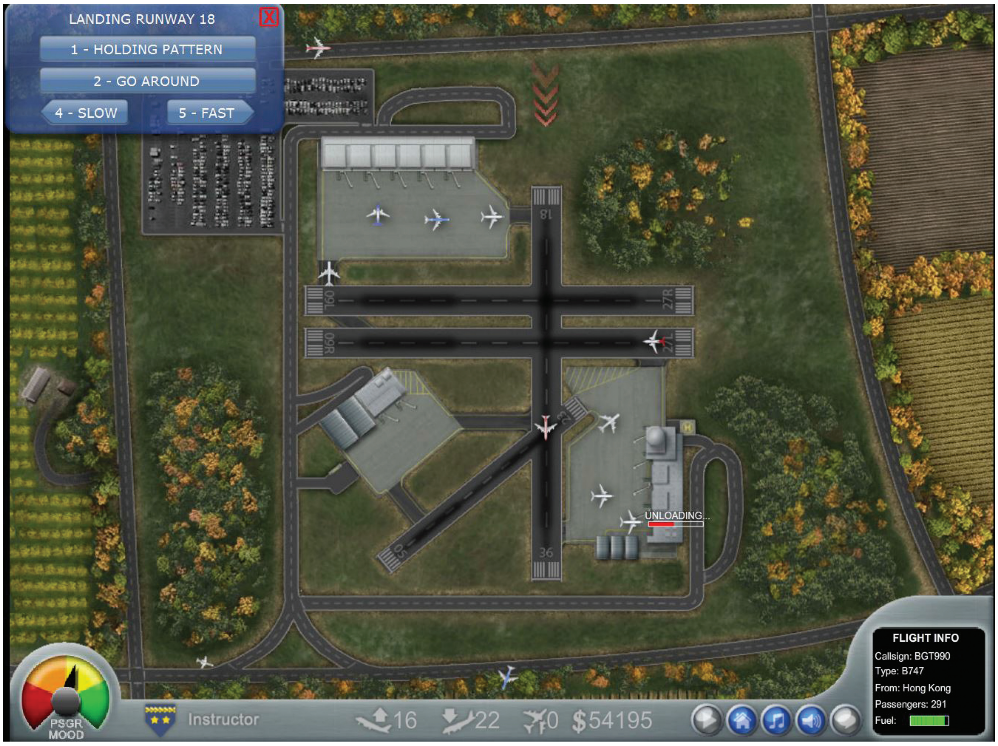
As stated above, our aim was to investigate whether providing people with real-time feedback on their mental workload, now that we can objectively and reliably measure it with fNIRS, could facilitate a form of reflection-in-action during tasks: that participants, in knowing their MWLs, can take action to manage their task or workload. Sharples and Megaw [49], said 'Operator workload or effort is not simply a function of task demands, but is influenced by how the task is perceived by the operator …'. In this case, we are making the operator workload explicit in the model, and examining the impact on both performance, and the demands of the task. Our primary aim, therefore, was to evaluate the effectiveness of using the WFS to aid an individual's self-awareness of current workload, such that they could (a) be more aware of their mental workload and (b) achieve good performance outcomes. As a secondary aim, we wanted to examine these outcomes against one of the widely used techniques for keeping people aware of their workload during tasks: ISAs; ISA, described further below, requires people to self-report their workload at intervals in order to keep them self-aware of it.

4.1 Study Conditions

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

Table 1. Four Main Conditions in the Study

Condition	Includes WFS	Includes ISA
Task only	No	No
Feedback	Yes	No
Feedback + ISA	Yes	Yes
ISA	No	Yes



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

Fig. 4. Airport Madness 4 – Screenshot of participant managing the landing of an aeroplane.

phases, the lights returned to their base colour when workload reduced, and thus could change back and forth multiple times during each task. We include the colour-based independent variable in the results, and examine the implications of colour choice in the discussion.

4.2 Air Traffic Control (ATC) Task

For the experiment, we required a task that (a) increased in difficulty and (b) could be managed by participants taking action in response to feedback. We selected an ATC task, using the commercially available Airport Madness 4 Game,[2] shown in Figure 4, in all task conditions. Participants had to coordinate the landing and departing of as many aeroplanes as possible, without causing incidents (e.g., collision between aeroplanes); the number of aeroplanes increased over time, thus increasing the demand as the task progressed. Planes are managed by clicking on the desired plane

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

icon and selecting an appropriate action – 'Land at runway X', 'Go Around', 'Increase/Decrease speed', 'change direction'. Similar options existed for planes requiring take off e.g., 'line up', 'immediate take off'. These controls allowed participants to use various strategies to reduce their mental workload during moments of high demands by, e.g., sending aeroplanes around, managing all landings on one lane and departures on other. The task interface also presented participants with direct measures of performance (as can be seen in Figure 4), such as the number of landed/departed aeroplanes.

4.3 Participants and Study Protocol

A total of 32 participants were recruited to take part in the experiment. Fifteen participants (9 males and 6 females) with an average age of 25.3 years (SD = 2.31) experienced the white-to-red lighting in Phase 1, and 17 (9 males and 8 females) with an average age of 25.5 years (SD = 8.05) experienced the red-to-white lighting in Phase 2. All participants had normal or corrected vision and reported no history of head trauma or brain damage. Participants were given a £10 voucher as a thank you and remuneration for their contribution to the project. The study protocol below was approved by the School of Computer Science ethics committee.

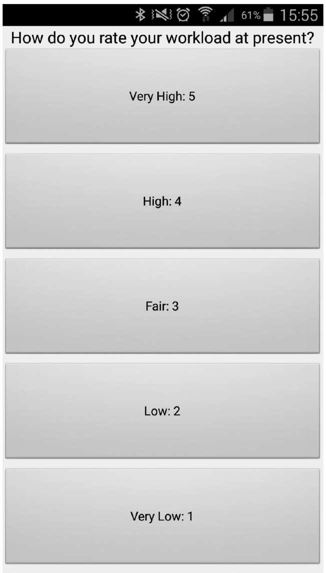
After gaining informed consent, participants began with a task familiarisation tutorial. All participants watched the same recorded video that introduced all the interactions with the video, and then were given the opportunity to practice the task until they felt confident in the game play; participants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

Participants completed each of the four study conditions, which were counterbalanced using Latin-square design to account for learning effects. In each condition, they were required to perform the study task from scratch for a period of 7 minutes. If they were to cause more than three major incidents within a condition, the game would automatically stop and the study condition would end (this however was not common). Seven minutes was enough to see numerous workload changes in the lights, but keep the full length of participation, including training, calibration, four tasks and between-task rest periods, to approximately o1 hour. After each condition, participants filled in a questionnaire to record perceived performance, before moving onto the next condition. After all four conditions, the study finished with a short interview, where participants had the chance to discuss the study experience and the way they perceived the WFS.

4.4 Measures of Dependent Variables

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

4.4.1 fNIRS Data. Although the WFS was only used during the feedback conditions, participants wore the fNIRS sensor in all conditions. While the WFS system used a single channel in the most sensitive region to monitor workload during tasks, comprehensive fNIRS data (HbO and HO) was recorded from all channels for the duration of all conditions. fNIRS data was further processed for post-experiment analysis using additional filters and feature extraction techniques. fNIRS data was processed using fnirSoft, the Comprehensive Signal Processing, Analysis, and Visualisation Platform for Optical Brain Imaging [3]. A low-pass filter with cutoff frequencies of 0.2Hz was used in order to remove high-frequency noise, physiological artefacts such as heartbeats and motion-derived artefacts. Additionally, we applied the Correlation-Based Signal Improvement (CBSI) filter [15], a technique designed for fNIRS technology in order to improve detection of workload. We named the resulting data OXY. In analysing the OXY data, we also considered the delay associated



How do you rate your workload at present?

Fig. 5. Instantaneous Self-Assessment (ISA) recorder app.

with the hemodynamic response [57], using various techniques including: averages across blocks of data, omitting the first few seconds of the trials when processing, and simply delaying the trial data [42, 44].

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

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

4.5 Study Hypotheses

To better understand the relationship between Feedback, ISA, workload, and performance, we state our hypotheses based on an adapted version of Sharples and Megaw's Framework for Mental Workload Measurement [49], where we controlled the external influences. Essentially, as shown

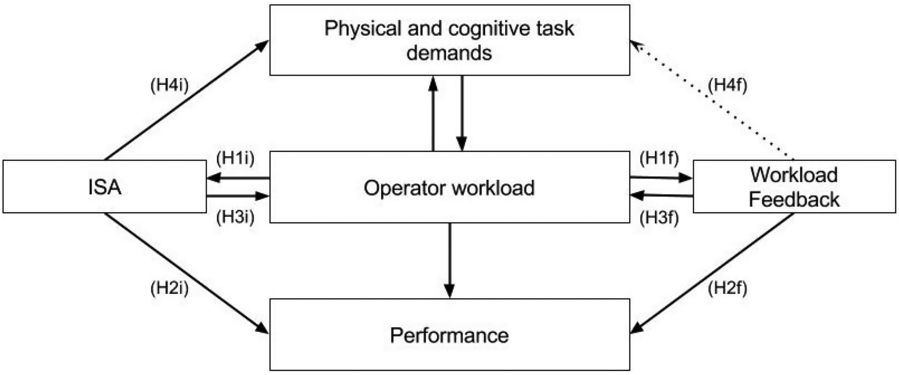


Fig. 6. Framework for Mental Workload Measurement: The relationship between MWL Feedback, ISA, performance, and workload (adapted from [49]).

in Figure 6, our two primary independent variables are shown as alternatives to the external influences boxes.

- H1 – *Variation in task demand will create measurable differences in mental workload.*
  As a baseline, as it is generally expected that increased task demands will generate increased levels of workload, we therefore expect that changes in task demands should be observable in both the ISA ratings and the objective measures of workload. We expect ISA ratings (H1i in Figure 6) to increase with increased task demand, and fNIRS measures to correlate (H2f), either positively or negatively, with task demand.
- H2 – *Participants' performance will be affected, positively or negatively, when made aware of their mental workload.*
  Ideally, participants may perform better because they are more aware of their workload without having to self-report using ISA, but may also have decreased performance if the feedback affects their ability to focus. We expect a lower performance in the presence of ISA (H2i) because of the activity involved in self-reporting. However, we do not expect a negative impact on performance from the WFS lighting changes (H2f).
- H3 – *Participants' perception of performance will be affected positively or negatively, when made aware of their mental workload.*
  As operators monitor their own performance, unconsciously or explicitly, having their workload levels presented during tasks should allow operators better reflect on performance. Aside from actual performance, participant's perception of their performance might be affected as they are made aware of their workload levels – increased workload could create a sense of poorer performance or higher performance for both independent variables (H3i and H3f).
- H4 – *Participants' perception and management of the task demands will be affected, positively or negatively, when made aware of their mental workload.*
  We expect that, given feedback-in-action, participants will think about their state whilst performing the task. Participants may also then take action to manage and manipulate the demands of the task, in order to maintain their workload levels to a particular point. In the case of ISA, H4i in Figure 6 highlights that there is a direct connection from ISA to the task demands, as participants have to do extra work to report their workload levels. As the WFS does not require additional effort from participants during tasks, we expect that their task

demands will not change, however, being presented with feedback of their workload levels more explicitly during tasks, participants perception of the task, and the decisions during the task may be affected (H4f).

5 RESULTS

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

5.1 H1 – Variation in Task Demand will Create Measurable Differences in Mental Workload

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

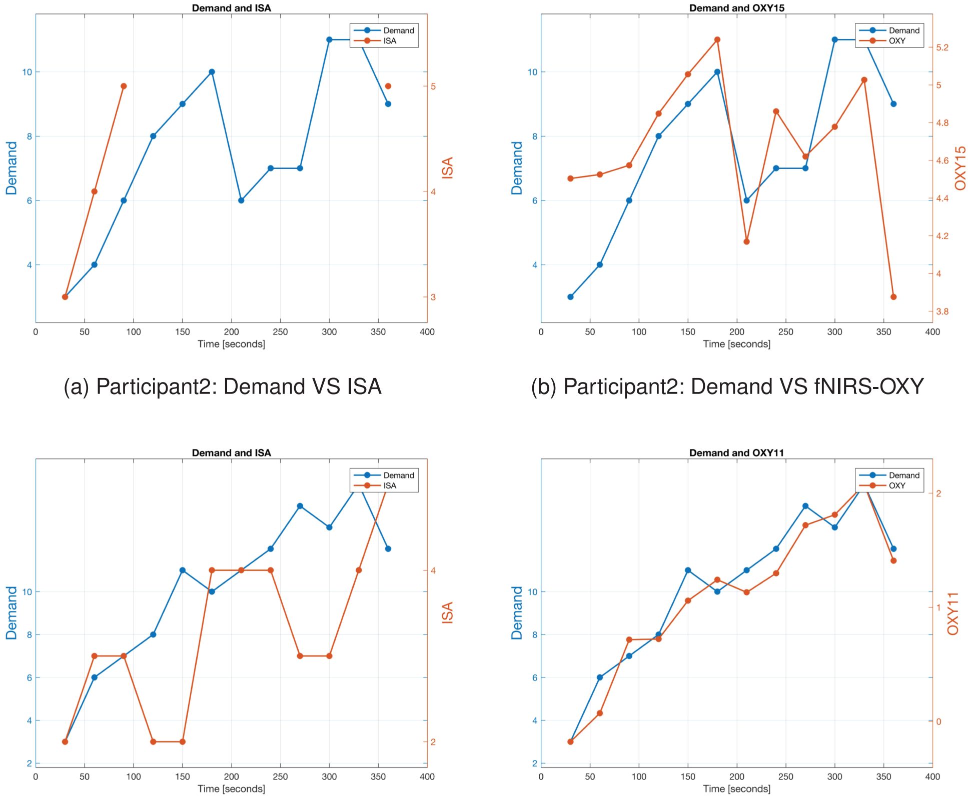
5.1.1 Subjective Ratings from ISA. We found strong correlations between demand and ISA measures, with examples shown in Figure 7, further showing the hypothesis H1i in Figure 6. The average correlation value across all participants, between ISA and demand (measured every 30 seconds) was $r = .68$ with the maximum value of $r = .899$ ($p = .006$) for Participant 4 in Phase 2. This correlation was strong for some participants, where P1's ISA correlation with task demand in Phase 2 during the ISA condition, for example, was $r = .808, p = .003$, and $r = .751, p = .003$ in the Feedback+ISA condition. There were, however, several cases across participants, where ISA did not reflect well its relationship to task demand, such as when participants were either too busy or to focused on the task and thus did not respond to the ISA questions (Figure 7(a)). This range of correlations highlights one of the known limitations of using mid-task self-assessment scales, as they rely on operators rating their workload during tasks.

5.1.2 Objective Measures from fNIRS. As described in the previous subsection, ISA was not always able to reflect the user's state (see Figure 7(a) and (c)), and mainly because it relies upon users subjectively reporting how they *feel*. For the same conditions and the same participants, Figure 7(b) and (d) shows how OXY correlates more objectively with the task demand. The average correlation value across participants, between fNIRS OXY and demand (measured every 30 seconds), was $r = .81$ with the maximum value of $r = .973$ ($p = .001$) for Participant 6 in Phase 2. This shows how fNIRS could be used to assess workload without relying on a subjective measure such as ISA.

5.1.3 Summary of H1 Results. Based on these results, we are able reject the null hypothesis and accept H1, as both subjective and objective measures provide evidence that participants' mental workload was associated with task demand. We conclude, however, that our objective measure (hypothesis H1f in Figure 6) was able to provide stronger and more consistent evidence of increased workload than subjective ratings (hypothesis H1i).

5.2 H2 – Participants' Performance will be Affected, Positively or Negatively, When Made Aware of Their Mental Workload

Considering all participants across both phases, a series of two-way repeated measure ANOVAs showed no statistical significance in the three performance measures (Total Departures, Total



(a) Participant2: Demand VS ISA

(b) Participant2: Demand VS fNIRS-OXY

(c) Participant6: Demand VS ISA

(d) Participant6: Demand VS fNIRS-OXY

Fig. 7. Emphasising the value and limitations of ISA (when participants fail to self-report values of their workload (a), it becomes hard to understand what happened during the task).

Landings, and Total Performance) between the four conditions. Total Performance, shown in Figure 8 for example, was different between the two phases. Similarly, the number of departures[3] and landings varied between phases, especially during the presence of both Feedback + ISA (see Figure 9). Participants performed slightly worse in Task Only condition in Phase 2 (the mean Total Departures was 10.2 for Phase 1 and 11.13 for Phase 2, and the mean Total Landings was 18.15 for Phase 1 and 17.36 for Phase 2), and slightly better in the Feedback + ISA condition in the same study, compared to Phase 1 (the mean Total Departures was 11.1 for Phase 1 and 9.93 for Phase 2, and the mean Total Landings was 16.46 for Phase 1 and 18.43 for Phase 2).

5.2.1 Phase 1 Performance Data. To consider the performance in Phase 1 (white light changed to red in periods of high workload as detected via fNIRS), we examine the Total Landings and Total Departures data over the four conditions in *Phase 1*, shown in the blue bars of Figure 9. For all the cases, the Total Departures, the Total Landings, and the Total Performance measures, performance appeared to decrease in the presence of ISA, suggesting that ISA might have a negative effect over the average participants' performance, and therefore affecting hypothesis H2i in Figure 6. This is

[3] There was one outlier in the total departures and the total performance data, which had a studentised residual value greater than ±3. The outlier was removed from the analysis.

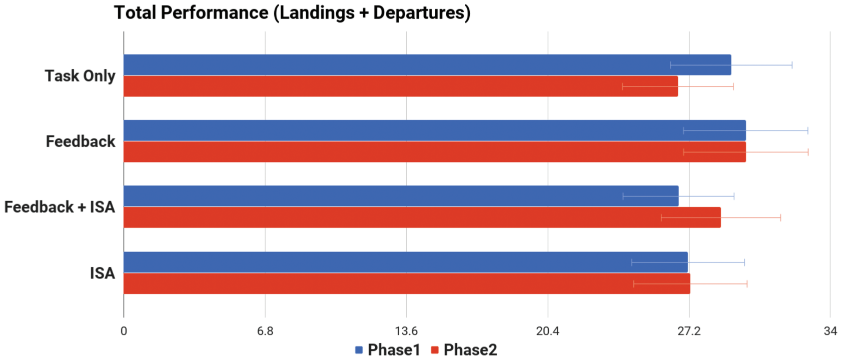
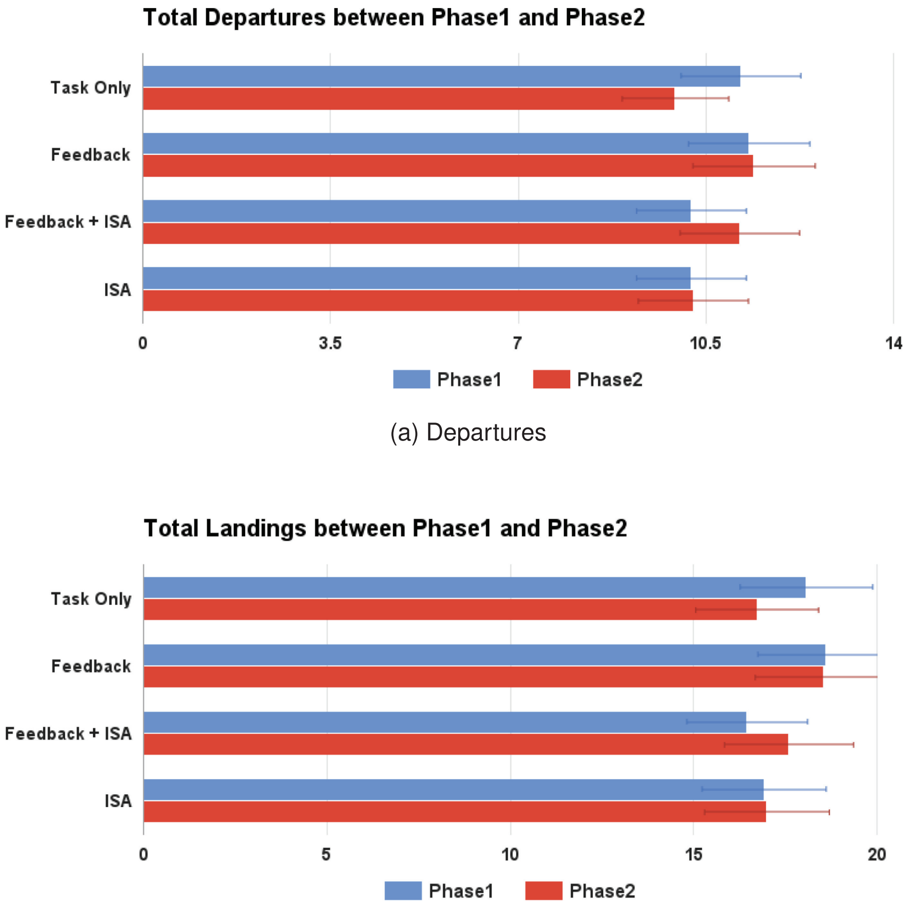


Fig. 8. Total Performance difference between Phase 1 and Phase 2 – mean scores across conditions.



(a) Departures

(b) Landings

Fig. 9. Performance data across the two phases.

not the case in the presence of Feedback, therefore, hypothesis H2f presents workload feedback having no explicit negative effects on performance.

There was no statistical significance found in the Total Departures with feedback impact, $F(1, 12) = .055, p = .819$, and ISA impact, $F(1, 12) = 2.476$, $p = .142$ as assessed by a two-way repeated measure ANOVA. There was also no statistically significant two-way interaction either between Feedback and ISA effect, $F(1, 12) = .014, p = .907$. For the Total Landings, the presence of feedback showed no significant impact, $F(1, 12) = 1.147$, $p = .305$; however, ISA significantly reduced performance, $F(1, 12) = 5.637$, $p = .035$ as assessed by a two-way repeated measure ANOVA. These results indicate that participants who responded to the ISA scale during task performance performed less well on the task; hence, hypothesis H2i in Figure 6 suggests a negative impact of ISA use on performance. There was no statistically significant two-way interaction between Feedback and ISA effect, $F(1, 12) = .014, p = .907$. A two-way repeated measure ANOVA additionally showed the impact of ISA on the Total Performance measure, $F(1, 12) = 5.368$, $p = .039$, and neither the effect of the presence of Feedback was found, $F(1, 12) = .675$, $p = .427$, nor the two-way interaction between the two, $F(1, 12) = .007$, $p = .937$. From this, we conclude that objective feedback provided via the change in lighting colour had no explicit negative impact on performance scores, and thus all the significant differences are caused by the deployment of ISA.

5.2.2 Phase 2 Performance Data. In *Phase 2* (red light changed to white in periods of high workload as detected via fNIRS), looking at the red bars in Figure 9, the negative effect of ISA was not found to be significant. Instead, the graph suggests a higher average performance during the presence of Feedback, after the colour change via lighting. This may suggest that workload feedback, and hypothesis H2f in Figure 6, may have a positive impact on performance.

Feedback showed no significant impact on performance with Total Departures, $F(1, 14) = .008$, $p = .932$, Total Landings, $F(1, 13) = 0.127$, $p = .727$, and Total Performance, $F(1, 13) = 0.072$, $p = .793$. ISA had no longer significant impact on performance, with Total Departures, $F(1, 14) = .229$, $p = .639$, Total Landings, $F(1, 13) = .011$, $p = .919$, and Total Performance, $F(1, 13) = 0.064$, $p = .804$. There was a statistically significant two-way interaction between ISA and Feedback effect on all performance measures: Total Departures, $F(1, 14) = 7.565$, $p = .015$; Total Landings, $F(1, 13) = 6.475$, $p = .024$; and Total Performance, $F(1, 13) = 9.388$, $p = .009$. The average performance across participants was lower in the absence of Feedback and ISA, compared to all other conditions as shown in Figure 9.

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

5.3 H3 – Participants' Perception of Their Performance and Workload will be Affected, Positively or Negatively, When Made Aware of Their Mental Workload

5.3.1 Perceived Performance Scores. A 5-point rating scale was used to capture participants' subjective perception of performance after each condition. Across both phases, a Friedman test was conducted to understand the within-subjects effects between all levels of the two factors: Feedback and ISA on *perceived* performance scores with Feedback–NoFeedback × ISA–NoISA conditions. Results showed a statistically significant difference between conditions, $N = 30, X^{(2)} = 9.072, p = .05$ (Figure 10). Pairwise comparisons were performed with a Bonferroni correction for multiple comparisons. There was a significant difference in the perceived performance scores between Feedback+ISA and Task alone, $p = 0.05$, and between Feedback+ISA and Feedback conditions, $p = 0.05$. Figure 10 shows how participants' perceived performance was significantly lower when ISA was

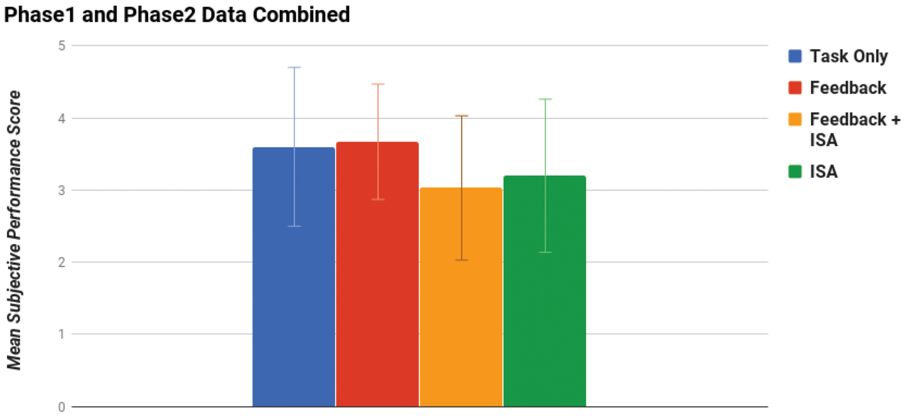


Fig. 10. Combined Phase 1 and Phase 2 mean perceived performance scores across conditions.

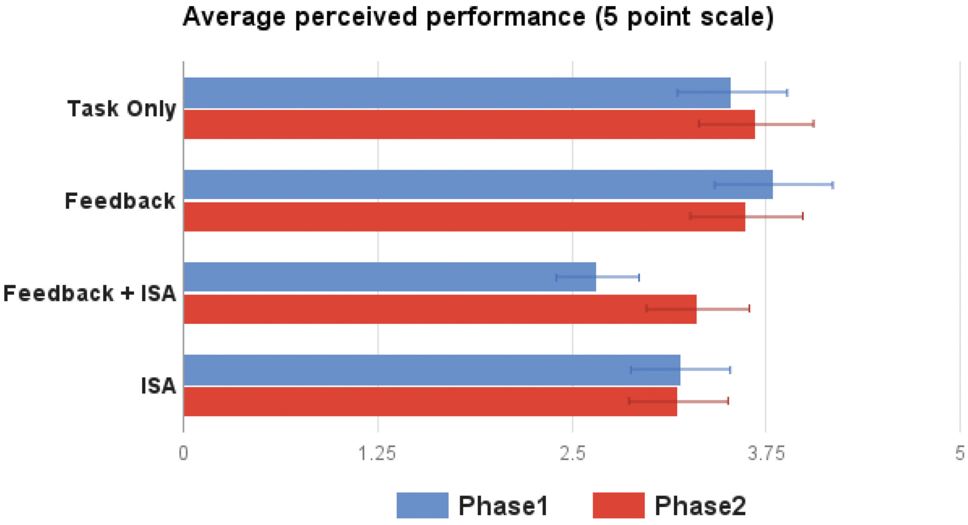


Fig. 11. Mean perceived performance scores across conditions.

present. These results show how ISA significantly reduces participants' perceived performance (hypotheses H2i and H3i), while the presence of feedback has no negative impact (hypotheses H2f and H3f).

For each of the two phases, the subjective performance scale generally showed lower perception of performance during the presence of ISA as shown in Figure 11, this being somewhat expected. This is directly related to hypotheses H4i, H2i, and H3i in Figure 6. It is interesting to observe how the perception of performance increased during the presence of ISA for Phase 2 compared to Phase 1, this effect being significant in the Feedback+ISA condition. This suggests an impact caused by the feedback type. To investigate the within-subjects effects between all levels of the two factors Feedback and ISA on *perceived* performance scores, a Friedman test was conducted with Feedback–NoFeedback × ISA–NoISA conditions for each phase separately.

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

5.3.3 Phase 2 Perceived Performance. Similar to the performance data, the subjective perception of performance data increased for the Feedback+ISA condition in Phase 2 (Figure 11). The Friedman test revealed no more significance in the perceived performance data between conditions, the presents of ISA having no longer significant impact, $X^{(2)} = 1.38, p = .71, df = 3$.

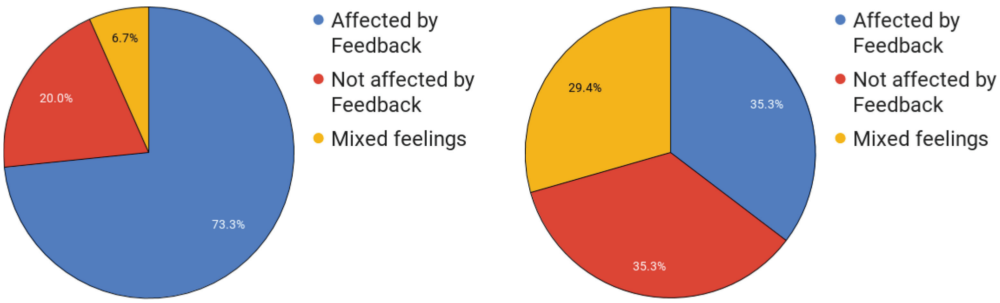
5.3.4 Comparing Phases. A Mann–Whitney *U*-test was conducted for each of the four conditions between the two studies to determine if there were any significant differences in perceived performance scores between Phase 1 and Phase 2. The test showed statistical significance between the two studies in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

5.3.5 Summary of H3 Results. We found that the deployment of ISA significantly reduced participants' perceived performance (hypotheses H2i and H3i), while the presence of feedback had no negative impact (hypotheses H2f and H3f in Figure 6). However, as with actual performance, it is interesting to note that the mean ISA score (across participants) revealed higher perceived workload in the presence of our WFS (see Figure 11). We again conclude that the impact of deploying ISA was exaggerated by additional demands, since participants' perception of workload was affected by the presence of ISA, but not by WFS alone. Overall, we have enough evidence to reject the null hypothesis and accept H3 for ISA, but not in the case of the WFS.

5.4 H4 – Participants' Perception and Management of the Task Demand will be Affected, Positively or Negatively, When Made Aware of Their Mental Workload

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

5.4.1 Impact of Feedback on Participants. In Phase 1, 11/15 of participants reported feedback affected them, 3/15 reported feedback did not affect them in any way, and just 1/15 was not sure about it (see Figure 12(a)). In Phase 2, only 6/17 reported feedback affected them, 6/17 reported feedback did not affect them, and 4/17 were not sure about it (see Figure 12(b)). This finding suggests that one particular feedback type (Phase 1) had more of an effect on participants, or the case where feedback in Phase 2 was more transparent, hence, not directly affecting participants' perception. However, it does not reflect the type of effects (negative or positive) the feedback had.



(a) Phase 1

(b) Phase 2

Fig. 12. The impact of different feedback on participants.

Participants were affected in different ways by the feedback. Some participants described feedback as a good indicator of ‘how much’ is going on during the task; P11 (Phase 1) gave an example of how he/she used the feedback during the task: ‘helpful in that you knew there was a lot going on and you are concentrating, which meant you had to pay extra attention to the details. When the lights went red, it meant that, yeah, I am probably concentrating a lot, that means I am probably miss some of the smaller aircrafts, and I would try and go back and have a look around to see if I was missing any small planes’. P9 (Phase 1) had similar feelings ‘They did help me to figure out how I was feeling, and how I was going. Sometimes when it went bright red, I thought to myself, I need to be calm, and think for a second what was going on in the game again. It made me take a second and just relax, that is all’.

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

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

5.4.2 Validity, Accuracy, and Delay. Participants also questioned the validity and correctness of the feedback. There were mixed views on what the feedback was informing, and how it worked. P29 (Phase 2) said ‘I noticed that when I would get calmed they would change to red. But sometimes I thought that I was working and they would still be red’, P2 (Phase 1) said ‘…most of the times it was accurate…’, and P4 (Phase 1) said ‘I was impressed though, they seemed what was not immediately responsive, but accurate. They seemed to change red when I was under a higher workload. When I was under more stress I said: “crap…another plane is coming in, they will crash!”. It did seem to pick up on that quite a lot which was cool’. These comments imply that,

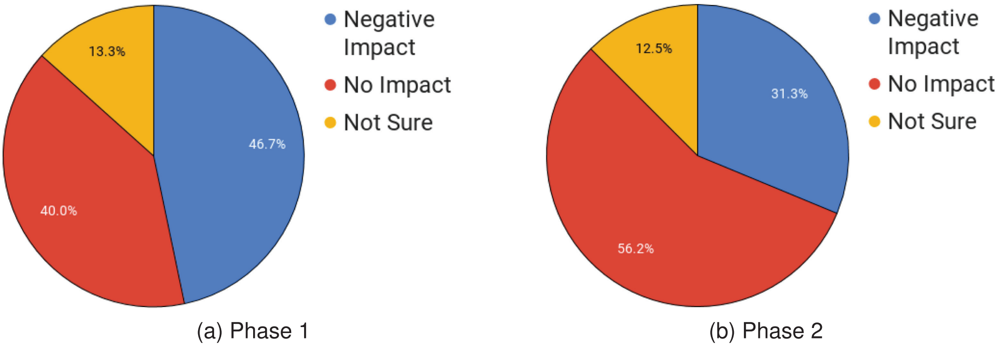


Fig. 13. The impact of ISA on participants perception.

despite being able to perceive the delay caused by the hemodynamic response [57], participants found the WFS to be mostly accurate to their current perceived workload.

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

5.4.4 Impact of ISA on Participants. During the interview, participants were also asked thoughts about ISA, its use, whether they think ISA had any impact on their performance and whether it was ambient or distracting by nature. We found 7/15 participants in Phase 1 and 5/17 in Phase 2 believed their performance was worse because of it, 6/15 in Phase 1 and 9/17 in Phase 2 believed it had no effects and the rest were not sure about it (see Figure 13(a) and (b)). One participant reported that ISA had no impact on them, however, in the presence of both, feedback and ISA, it made him/her think: P16 (Phase 2) saying: '"No. It did not bother me that much. But at times it made me think, during the condition with feedback as well, especially when there was a discrepancy between the two'. The majority of participants who reported ISA having a negative impact on them, also reported losing focus when having to answer ISA questionnaire during the task. P26 (Phase 2) said 'It definitely made it worse … because it takes you out of the action, and then it takes a little while to figure out where you were'. P31 (Phase 2) reported that ISA had a continuous negative impact, 'Sometimes I would hear the notification in my subconscious, and I did not pay attention because I was very concentrated on the task'.

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

5.4.5 Summary of H4 Results. This hypothesis was mostly examined through the post-experiment interview and participants' opinions, since we did not have any objective measurements of how people reflected on the feedback mid-task. It is clear that both ISA and feedback had an impact on participants' management of the task. Although the opinions were divided, and participants were 'affected' differently by the feedback and ISA, the findings were very constructive and informative. We found insights to confirm that participants noticed the feedback and considered it at a meta-cognitive level during the task, and therefore have enough evidence to reject the null hypothesis in case of H4. In comparison to ISA, P24 (Phase 2) said 'I think the lights are more effective, because the cellphone app (ISA) just made me feel more busy. They lights show me when I am busy, where ISA made me feel busy'. In Section 6, we consider what future work may do to investigate this finding in more detail.

6 DISCUSSION

This article described an alternative way to raise self-awareness of metal workload, through providing workload feedback based upon a concurrent objective measure, and our results showed that it did so without negatively affecting performance (as with ISA). We expected that if users are alerted that they are approaching a drop/dip in performance because of high/low workload, then they might be able to take action to avoid it. Table 2 summarises the results of the study that relate to the hypotheses visualised in Figure 6.

6.1 Impact of ISA

In line with the findings of a previous work [35, 55], we found that self-reporting workload using ISA significantly impacted both the actual performance and perceived performance. Participants landed significantly fewer planes in the presence of ISA compared to all other conditions. As expected with ISA, we saw many participants miss ISA entries when they had high workload, and were often surprised when we showed them gaps in their self-reporting. P13, who missed several ISA responses, said: 'I did not know if there was a time limit I had to answer. I do not think I missed any', but said ISA was 'just another button I had to press … it became a mechanical task'. Conversely, P14 said 'It was annoying! I was ok, go away, go away … It is like an alarm in the morning' and P15 said 'Annoying..your phone app workload questionnaire is really annoying … I did not notice the phone sometimes when I was concentrating on the task, so completely ignoring ISA. Sometimes people get easily distracted and for this kind of task it can be dangerous …'. Even though the general feeling was against ISA, some participants' perception of ISA was not that bad; P3 said 'I do not think that ISA had an impact on performance …'. These sentiments were generally observable in the data, and so our findings match the consensus of prior research into both its validity for measuring workload (since it had strong correlation scores) and the interference it has on the primary task.

6.2 Impact of Objective Feedback

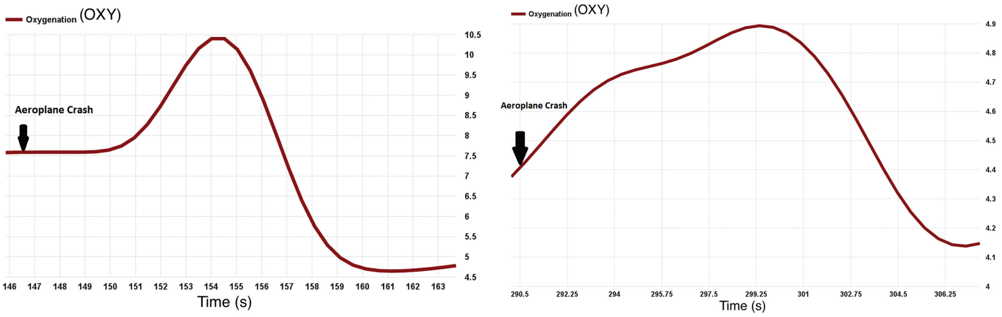
The aim of our study was to investigate whether presenting users with real-time mental workload feedback, would make them aware of their load without notably reducing either actual or perceived performance. This was presented in contrast to ISA measure, which requires the user to reflect upon their mental workload and take action to report it. In our results, feedback did not affect

Table 2. Summary of Key Findings, by Hypothesis

Hypotheses	Expected effects	Results
H1	Participants' workload generated by the task demands would have measurable effects with ISA and fNIRS	—We found both fNIRS and ISA measures sensitive to task demands (in our case the number of aeroplanes to control during an ATC game) —We found high correlation coefficients between fNIRS and demand, and we showed how it can be used to assess workload without relying on participants' ability to self-report during the task (see Figure 7)
H2 and H3	ISA would have a negative impact on performance. Feedback would have no explicit negative impact on performance. Workload and perception of performance would increase or decrease in the presence of both ISA and our WFS	—We presented evidence supporting ISA's negative impact on both performance and perceived performance measures —In contrast to ISA, we found mental workload feedback having no explicit negative impact —Figure 11 suggests similar or slightly improved performance with the WFS and workload feedback
H4	In contrast to our WFS, ISA will create additional physical and cognitive task demands	—Although we found no direct evidence, the performance and perceived performance results suggested a negative impact in the presence of ISA, most likely due to the additional required resources. This was not the case with our WFS —During the interview, participants had mixed feelings about the impact of ISA; views were divided into participants affected by ISA, participants who considered ISA having no negative impact (describing it 'easy' and 'fast'), and participants who ignored ISA when concentrating on the task

actual and perceived performance in a negative way. Although not significant within our sample, feedback appeared to slightly improve actual performance and participants perceived that they performed slightly better (Feedback Condition – see Figure 11). The findings alone, however, do not tell us whether participants noticed the feedback, understood their workload, and took action to reduce them.

In interviews, some participants indicated that they did take note of the feedback: 'When the lights become red, it works as a reminder to take a big breath and relax [. . .] it is like a warning . . . ' (P15 Phase1). For some, this was positive, with P14 (Phase 1) saying 'I really liked it! The whole



(a) Example 1

(b) Example 2

Fig. 14. Participant oxygenation levels measured with fNIRS after a 'crash' event.

experience … If I would get another chance I would do it again'. Some, however, were frustrated that they couldn't do much about it: 'It is actually affecting me. When the feedback is red, I try to relax. To try to make it white. But it did not work, because I felt even more concentrated … because I was looking at the planes and to the lights as well, so it added up really to my concentration' (P14). Other participants felt that feedback had no use as they already know when they are busy and when not: P1 (Phase 1) said 'I felt like: why are you showing me red if I already know that I am busy?" even though he said later that 'I am usually really bad at judging my own workload'. These insights confirm that participants noticed the feedback and considered it at a meta-cognitive level during the task. For some participants the feedback, however, perhaps increased the sense of anxiety (especially in the first phase), when participants were not able to take action to change it. Because of this, P15 (Phase1) went further to suggest that although the objective measure of their workload was useful, they would have preferred to see it afterwards, rather than during the task: 'I would like to add that it would be much more interesting for me to have a feedback to reflect on but not a concurrent one. So maybe record it and reflect on it later on'. This may be an interesting area of future work, as a mental workload parallel to life-logging and tracking daily fitness activity – a form of *Mental Workload Fitness* tracker. Further, more participants suggested various levels of feedback would be much more useful 'rather than a cut off point … a gradual transition in a way' (P1 Phase 1).

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

6.3 Future Work

6.3.1 Continuously Assessing Workload. Beyond the challenges of real-time processing of data, there are a number of interesting events that can occur when continuously assessing workload – we noted that mental workload fluctuated noticeably when aeroplanes crashed, and informally participants noted feeling stressed. It is interesting to consider what participants do in these situations, and what this might look like in workload data. We wanted to investigate what happened when participants failed to monitor and control all aeroplanes on the screen, and two or more aeroplanes ended up in a colliding. Figure 14(a) and (b) shows, for example, measurable changes in fNIRS OXY signal after such a fail, and its impact on Feedback. It is clearly important to consider whether this is cause or affect, but based on informal secondary analyses of our data, we saw many of these dramatic shifts in mental workload around fail scenarios.

In future work, we would like to more directly evaluate and, associate these reactions to events, whilst still giving people reliable feedback about their mental workload. Such future work may also focus on using other measures of workload, as emotional reactions are typically more observable through other physical reactions. We used an fNIRS device, which has been shown to be suitable for HCI user study evaluations, but more commercially available devices like the NeuroSky[4] EEG device might be more suitable for day to day feedback. Similarly, even less invasive measures of mental workload could be taken from Heart Rate Variability [24] through smart watches, remotely detected by pupil dilation [33] or facial skin temperature [53] with cameras. Many of these other measures also better detect emotional responses, and perhaps concepts like stress and anxiety, and so might better serve future work on recording both mental workload and emotional response.

6.3.2 Designing Different Feedback Types. In general, the binary feedback of workload was alerting users of a high workload when a sudden increase in oxygenation was measured using fNIRS. In the same way, a sudden decrease would cause a low workload alert after. The changes were visible to participants, such that they could monitor and use their workload feedback presented by the WFS. However, future work could first examine more granular forms of feedback, as noted qualitatively by participants. It was interesting, however, to first informally observe, and then analytically find differences between the choice of lighting feedback in two phases in the study. This post-hoc independent variable in our analysis revealed interesting results that imply that confirm Sharples and Megaw's [49] description that workload is closely affected by way in which participants experience that workload. In Phase1, red colour was used for feeding back high workload states, and white colour for low workload states, and in Phase 2 the colours were swapped. Having white light to alert of high workload made some participants feel 'right' being on the white colour rather than red making them feel they are not working enough; P29 (Phase 2) '…when they were red, I thought I am not working enough. When they were white it felt right, it felt that I was paying a lot of attention, it was in the right track'. On the other side, having colour red to alert of high workload generated pressure when 'being' on red, P9 (Phase 1) reported that 'When the lights become red, it works as a reminder to take a big breath and relax'. It would be extremely interesting in future work to artificially manipulate changes in feedback, and to observe changes in workload in a similar way to when participants experienced crashes (like in Figure 14). Such an analysis would help us to separately examine the impact of mental workload created by feedback and mental workload created by task demand. Future work should also, therefore, explore the design and mode of feedback, as well as the granularity of feedback.

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

7 CONCLUSIONS

In this article, we considered whether brain sensing techniques, which are increasingly becoming commercially available, could be used to give people concurrent feedback about their mental workload levels. Although existing techniques, like the ISA tool, are designed to help people to report and reflect on their current mental workload levels, they also often have a negative impact

[4] http://neurosky.com/.

on the primary task at hand. We hoped that, with objectively measuring and providing concurrent feedback during tasks, participants would be able to reflect on the mental workload levels, without the associated performance drops. In order to capture and understand these effects, we have adapted the framework for mental workload evaluation (presented in Figure 1). We ‘controlled’ the ‘external factors’ presented in the relationship 5 of the framework (in our case the presence of ISA and the presence of Feedback). Our results first confirmed both approaches to measuring mental workload during tasks, accurately correlating the measures with task demands. We then confirmed prior research findings that self-reporting techniques had an impact on both actual and perceived performance, as well as increasing the task demands on the participants. Our results, however, showed no such drops in performance were found with our mental workload feedback system. Using the framework, we confirmed the existing relationships between the physical and cognitive task demands, and the operator workload (see relationship 1 in Figure 1), as well as the direct connection between operator workload and performance (see relationship 2 in Figure 1). Further, our interviews confirmed that feedback led participants to reflect on their mental workload during tasks, but that the choice of feedback (using red lights to warn them of high mental workload) created a negative stressor to their experience. This effect was removed after changing the choice of colour in our feedback mechanism. The results suggested that participants do use the feedback of workload, therefore showing the relevance of relationship 3 in the same framework. Relationship 4 was not directly studied in this article; however, the future work section presents the authors interest in associating various physiological reactions to events (e.g., such as task failure). We conclude that objectively measured concurrent feedback of mental workload can help people to understand and actively manage their behaviour during tasks, but without the negative effects on performance created by self-reporting techniques. Such personal insight would be important for safety critical tasks like ATC, but has the potential to more broadly help us to understand and manage our own mental workload across the many tasks that fill our lives. Future work, however, should focus on the design of mental workload feedback mechanisms, to provide people with insight without increasing their workload and stress levels.

ACKNOWLEDGMENTS

The authors greatly thank Alexandru Ghitulescu for the support with the development of the Workload Feedback System, Hyosun Kwon for all the effort with Figure 3, and Christopher Ellis for the development of the fNIRS data processing script (available at: https://github.com/Christopher51773/FNIRS_DATA_PROCESSING).

Data access statement: Consent was not gained from participants to release the study data online, and so a dataset is not openly available.

REFERENCES

[1] Daniel Afergan, Evan M. Peck, Erin T. Solovey, Andrew Jenkins, Samuel W. Hincks, Eli T. Brown, Remco Chang, and Robert J. K. Jacob. 2014. Dynamic difficulty using brain metrics of workload. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3797–3806.

[2] Gabor Aranyi, Fred Charles, and Marc Cavazza. 2015. Anger-based BCI using fNIRS neurofeedback. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, 511–521.

[3] Hasan Ayaz. 2010. Functional Near Infrared Spectroscopy Based Brain Computer Interface (Doctoral Dissertation). Retrieved March 19, 2018 from http://www.hasanayaz.com/docs/2010_Ayaz_PhD_Sum.pdf.

[4] Hasan Ayaz and Banu Onaral. 2005. *Analytical Software and Stimulus-presentation Platform to Utilize, Visualize and Analyze Near-infrared Spectroscopy Measures*. Drexel University.

[5] Hasan Ayaz, Patricia A. Shewokis, Scott Bunce, Kurtulus Izzetoglu, Ben Willems, and Banu Onaral. 2012. Optical brain monitoring for operator training and mental workload assessment. *NeuroImage* 59, 1 (2012), 36–47.

[6] Alan Baddeley. 2000. The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences* 4, 11 (2000), 417–423.

[7] Alan Baddeley. 2003. Working memory: Looking back and looking forward. *Nature Reviews Neuroscience* 4, 10 (2003), 829–839.

[8] Michela Balconi, Elisabetta Grippa, and Maria Elide Vanutelli. 2015. What hemodynamic (fNIRS), electrophysiological (EEG) and autonomic integrated measures can tell us about emotional processing. *Brain and Cognition* 95 (2015), 67–76.

[9] Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin* 91, 2 (1982), 276.

[10] George E. Billman. 2011. Heart rate variability – A historical perspective. *Frontiers in Physiology* 2 (2011), 86.

[11] Scott C. Bunce, Kurtulus Izzetoglu, Hasan Ayaz, Patricia Shewokis, Meltem Izzetoglu, Kambiz Pourrezaei, and Banu Onaral. 2011. Implementation of fNIRS for monitoring levels of expertise and mental workload. In *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems*. Springer, 13–22.

[12] Brad Cain. 2007. *A Review of the Mental Workload Literature*. Technical Report. Defence Research and Development Toronto, Canada, OMB No. 0704-0188.

[13] Britton Chance, Endla Anday, Shoko Nioka, Shuoming Zhou, Long Hong, Katherine Worden, C. Li, T. Murray, Y. Ovetsky, D. Pidikiti, and others. 1998. A novel method for fast imaging of brain function, non-invasively, with light. *Optics Express* 2, 10 (1998), 411–423.

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

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

[16] Gautier Durantin, J.-F. Gagnon, Sébastien Tremblay, and Frédéric Dehais. 2014. Using near infrared spectroscopy and heart rate variability to detect mental overload. *Behavioural Brain Research* 259 (2014), 16–23.

[17] Matthieu Duvinage, Thierry Castermans, Thierry Dutoit, M. Petieau, T. Hoellinger, C. D. Saedeleer, K. Seetharaman, and G. Cheron. 2012. A P300-based quantitative comparison between the Emotiv Epoc headset and a medical EEG device. In *Proceedings of the IASTED International Congress on BioMedical Engineering*.

[18] Kingsley Fletcher. 2015. *The Dynamic Effects of Task Demands on Resource Availability, Resource Allocation and Metacognitive States*. The University of Queensland.

[19] Jérémy Frey, Maxime Daniel, Julien Castet, Martin Hachet, and Fabien Lotte. 2016. Framework for electroencephalography-based evaluation of user experience. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 2283–2294.

[20] Jérémy Frey, Christian Mühl, Fabien Lotte, and Martin Hachet. 2013. Review of the use of electroencephalography as an evaluation method for human-computer interaction. arXiv:1311.2222.

[21] Jacek Gwizdka. 2010. Using stroop task to assess cognitive load. In *Proceedings of the 28th Annual European Conference on Cognitive Ergonomics*. ACM, 219–222.

[22] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*. ACM, 301–310.

[23] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology* 52 (1988), 139–183.

[24] Javier Hernandez, Daniel McDuff, and Rosalind W. Picard. 2015. BioWatch: Estimation of heart and breathing rates from wrist motions. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 169–176.

[25] Leanne M. Hirshfield, Erin Treacy Solovey, Audrey Girouard, James Kebinger, Robert J. K. Jacob, Angelo Sassaroli, and Sergio Fantini. 2009. Brain measurement for usability testing and adaptive interfaces: An example of uncovering syntactic workload with functional near infrared spectroscopy. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*. ACM, 2185–2194.

[26] Christina Hirth, Hellmuth Obrig, Kersten Villringer, Andeas Thiel, Johannes Bernarding, Werner Mühlnickel, Herta Flor, Ulrich Dirnagl, and Arno Villringer. 1996. Non-invasive functional mapping of the human motor cortex using near-infrared spectroscopy. *Neuroreport* 7, 12 (1996), 1977–1981.

[27] T. J. Huppert, R. D. Hoge, S. G. Diamond, Maria Angela Franceschini, and David A. Boas. 2006. A temporal comparison of BOLD, ASL, and NIRS hemodynamic responses to motor stimuli in adult humans. *NeuroImage* 29, 2 (2006), 368–382.

[28] Kurtulus Izzetoglu, Scott Bunce, Banu Onaral, Kambiz Pourrezaei, and Britton Chance. 2004. Functional optical brain imaging using near-infrared during cognitive tasks. *International Journal of Human-Computer Interaction* 17, 2 (2004), 211–227.

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

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

[31] C. S. Jordan and S. D. Brennen. 1992. Instantaneous self-assessment of workload technique (ISA). Retrieved March 19, 2018 from http://www.skybrary.aero/bookshelf/books/1963.pdf.

[32] Daniel Kahneman. 1973. *Attention and Effort*, Vol. 1063. Prentice-Hall, Englewood Cliffs, NJ.

[33] Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. 2008. Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*. ACM, 69–72.

[34] M. Leanne and J. Robert. 2009. Using brain measurement to evaluate reality based interactions. In *Proceedings of CHI'09 Workshop: Challenges in the Evaluation of Usability and User Experience in Reality Based Interaction*. 19–20.

[35] Andrew Leggatt. 2005. Validation of the ISA (instantaneous self assessment) subjective workload tool. In *Contemporary Ergonomics 2005: Proceedings of the International Conference on Contemporary Ergonomics (CE'05)*, 5–7 April 2005, Hatfield, UK. CRC Press, 74.

[36] Xiaofei Li, Lele Xu, Li Yao, and Xiaojie Zhao. 2013. A novel HCI system based on real-time fMRI using motor imagery interaction. In *Foundations of Augmented Cognition*. Springer, 703–708.

[37] Manja Lohse, Reinier Rothuis, Jorge Gallego-Pérez, Daphne E. Karreman, and Vanessa Evers. 2014. Robot gestures make difficult tasks easier: The impact of gestures on perceived workload and task performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1459–1466.

[38] Kristiyan Lukanov, Horia A. Maior, and Max L. Wilson. 2016. Using fNIRS in usability testing: Understanding the effect of web form layout on mental workload. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, 4011–4016.

[39] Horia A. Maior, Matthew Pike, Sarah Sharples, and Max L. Wilson. 2015. Examining the reliability of using fNIRS in realistic HCI settings for spatial and verbal tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*, Vol. 15. 3807–3816.

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

[41] Fred Paas, Alexander Renkl, and John Sweller. 2003. Cognitive load theory and instructional design: Recent developments. *Educational Psychologist* 38, 1 (2003), 1–4.

[42] Evan M. Peck, Beste F. Yuksel, Alvitta Ottley, Robert J. K. Jacob, and Remco Chang. 2013. Using fNIRS brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 13)*. ACM.

[43] Laura Pickup, John R. Wilson, Beverley J. Norris, Lucy Mitchell, and Ged Morrisroe. 2005. The integrated workload scale (IWS): A new self-report tool to assess railway signaller workload. *Applied Ergonomics* 36, 6 (2005), 681–693.

[44] Matthew F. Pike, Horia A. Maior, Martin Porcheron, Sarah C. Sharples, and Max L. Wilson. 2014. Measuring the effect of think aloud protocols on workload using fNIRS. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3807–3816.

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

[46] Peter Rasche, Alexander Mertens, Christopher Schlick, and Pilsung Choe. 2015. The effect of tactile feedback on mental workload during the interaction with a smartphone. In *Cross-Cultural Design Methods, Practice and Impact*. Springer, 198–208.

[47] Gary B. Reid and Thomas E. Nygren. 1988. The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in Psychology* 52 (1988), 185–218.

[48] Donald A. Schön. 1983. *The Reflective Practitioner: How Professionals Think in Action*, Vol. 5126. Basic Books.

[49] Sarah Sharples and Ted Megaw. 2015. Definition and mesurement of human workload. In *Evaluation of Human Work*, John R. Wilson and Sarah Sharples (Eds.). CRC Press.

[50] Yu Shi, Natalie Ruiz, Ronnie Taib, Eric Choi, and Fang Chen. 2007. Galvanic skin response (GSR) as an index of cognitive load. In *Proceedings of the CHI'07 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2651–2656.

[51] Erin Solovey, Daniel Afergan, Evan M. Peck, Samuel W. Hincks, and Robert J. K. Jacob. 2015. Designing implicit interfaces for physiological computing: Guidelines and lessons learned using fNIRS. *ACM Transactions on Computer-Human Interaction* 21, 6 (2015), 35.

[52] Erin Treacy Solovey, Audrey Girouard, Krysta Chauncey, Leanne M. Hirshfield, Angelo Sassaroli, Feng Zheng, Sergio Fantini, and Robert J. K. Jacob. 2009. Using fNIRS brain sensing in realistic HCI settings: Experiments and guidelines. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology (UIST'09)*. ACM, 157–166.

[53] John Stemberger, Robert S. Allison, and Thomas Schnell. 2010. Thermal imaging as a way to classify cognitive workload. In *Proceedings of the IEEE Canadian Conference on Computer and Robot Vision*. 231–238.

[54] M Strait, C. Canning, and M. Scheutz. 2013. Limitations of NIRS-based BCI for realistic applications in human-computer interaction. In *Proceedings of the BCI Meeting*. 6–7.

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

[56] Pamela S. Tsang and Velma L. Velazquez. 1996. Diagnosticity and multidimensional subjective workload ratings. *Ergonomics* 39, 3 (1996), 358–381.

[57] Arno Villringer and Britton Chance. 1997. Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neurosciences* 20, 10 (1997), 435–442.

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

[59] Alan Traviss Welford. 1968. *Fundamentals of Skill*. London, Methuen.

[60] Harold Thomas Anthony Whiting. 1969. *Acquiring Ball Skill: A Psychological Interpretation*. Lea & Febiger.

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

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

[63] Beste F. Yuksel, Daniel Afergan, Evan M. Peck, Garth Griffin, Lane Harrison, Nick W. B. Chen, Remco Chang, and Robert J. K. Jacob. 2015. BRAAHMS: A novel adaptive musical interface based on users' cognitive state. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME'15).*

[64] Beste F. Yuksel, Kurt B. Oleson, Lane Harrison, Evan M. Peck, Daniel Afergan, Remco Chang, and Robert J. K. Jacob. 2016. Learn piano with BACh: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'16).* ACM, 5372–5384.

Received October 2016; revised November 2017; accepted November 2017