Toward Workload-Based Adaptive Automation: The Utility of fNIRS for Measuring Load in Multiple Resources in the Brain

Leanne M. Hirshfield[a,b]*, Christopher Wickens[c], Emily Doherty[a,b], Cara Spencer[a,b], Tom Williams[d], Lucas Hayne[a,b]

[a]Institute of Cognitive Science, University of Colorado, Boulder CO, USA.
[b]Department of Computer Science, University of Colorado, Boulder CO, USA.
[c]Department of Cognitive Psychology, Colorado State University, Fort Collins, CO, USA.
[d]Department of Computer Science, Colorado School of Mines, Golden, CO, USA.

* Corresponding Author: Leanne.Hirshfield@Colorado.edu. Institute of Cognitive Science, CU Boulder, Boulder CO, 80309

Toward Workload-Based Adaptive Automation: The Utility of fNIRS for Measuring Load in Multiple Resources in the Brain

Abstract. We investigate the utility of functional near-infrared spectroscopy (fNIRS) for workload-based adaptive automation through the lens of multiple resource theory. We focus on the criteria of unobtrusiveness, responsiveness, load sensitivity (low vs high load), and load diagnosticity (differentiating types of load). We report a large meta-review, in which we conclude that only a few studies were suitable for evaluating sensitivity and diagnosticity in complex real-world tasks. While these reveal that the fNIRS signal is adequately sensitive to gradations of load level changes (sensitivity), the diagnosticity of fNIRS to different sources of cognitive load remained uncertain. We manipulated mental load of a complex shape sorting task via working memory load (WM) and visual perceptual load (VL), while a secondary auditory task was present throughout. We measured the effect of these manipulations at the group-level using conventional secondary and eyetracking workload measures, as well as hemodynamic response in specific functional regions in the brain, including regions involved in multi-tasking (MT), VL, WM, and auditory load (AL). Our findings revealed that fNIRS is both sensitive and diagnostic to load in complex tasks, with greater sensitivity revealed by deoxyhemoglobin than oxyhemoglobin and the brain regions associated with diagnosticity align with neuroscience literature on perceptual load, WM, and goal-directed multitasking.

Keywords: workload, fNIRS, near-infrared spectroscopy, automation, multiple resource theory

1 Introduction

The concept of adaptive automation (AA) has been discussed frequently in the fields of human-computer interaction (HCI) and human factors, whereby some aspect of automation is changed in real time, based on an inference of human cognitive state made by an automated agent (e.g., Al-Hudhud, Alqahtani, Albaity, Alsaeed, and Al-Turaiki (2019); Dorneich et al. (2015); Rouse (1988); C. D. Wickens, Helton, Hollands, and Banbury (2022)). More specifically, workload-based AA is implemented by using a human operator's cognitive load to define how the agent's response should adapt. For example, if cognitive workload is high then the agent could automate some of the tasks done manually, or at least increase the degree of automation of those tasks

(Onnasch, Wickens, Li, & Manzey, 2014). Attractive as this concept is, AA has proven challenging to implement in practice, and it has been difficult to demonstrate its performance benefits over non-adaptive automation (e.g., Sauer, Kao, and Wastell (2012); see C. D. Wickens et al. (2022) for a summary). A myriad of neurophysiological techniques have been proposed to measure workload objectively and in real time. These techniques require no deliberative response from the human in order for the adaptive agent to form a workload estimate. Most generally, these have been categorized under the purview of physiological or neuroergonomic measures of mental workload (Brouwer, Zander, van Erp, Korteling, & Bronkhorst, 2015; Fairclough, Ewing, Burns, & Kreplin, 2019; Goshvarpour & Goshvarpour, 2023; Saikia, Kuanar, Borthakur, Vinti, & Tendhar, 2021; Shirzadi, Einalou, & Dadgostar, 2020) and include measures such as the power in certain frequency bands (e.g., alpha, theta) as measured by EEG (Gevins & Smith, 2003), pupil diameter (Kaber & Kim, 2011; Recarte & Nunes, 2003), or cardiac parameters (Backs, Lenneman, Wetzel, & Green, 2003). Many of these are covered in depth by H. Ayaz and Dehais (2018) and are summarized by C. D. Wickens et al. (2022).

Of particular interest in the current paper is functional near-infrared spectroscopy (fNIRS), a non-invasive brain blood-flow measurement device that has seen a rapid increase in use across a variety of research domains since its development in the 1990s (Hasan Ayaz et al., 2022; Chance, Zhuang, Chu, Alter, & Lipton, 1993; von Lühmann et al., 2021; M. Yücel et al., 2021; M. A. Yücel, Selb, Huppert, Franceschini, & Boas, 2017; H. Zhao & Cooper, 2018). Aligned with multiple resource theory (Navon & Gopher, 1979; C. D. Wickens, 1980), we focus on the utility of fNIRS for addressing four key measurement challenges that have hampered workload-based AA accomplishments to date, **unobtrusiveness** of the sensors, **sensitivity** to load levels, **diagnosticity** of qualitatively different types of load (e.g., visual vs cognitive vs motor), whose importance, in the context of multiple resources within the brain, will be discussed below, and **temporal responsiveness** suitable for real-time adaptations. These four have been a hallmark of mental workload research for decades, first introduced by Moray (1979), and subsequently formalized by C. Wickens (1984).

One of the largest challenges to the implementation of AA is to obtain an assessment of high mental workload (or reduced residual capacity) in a behaviorally **unobtrusive** fashion (e.g., by avoiding imposing a secondary task, or the requirement that a subjective rating be given in real time). We note the advantages of neurophysiological measures of mental workload in this

regard, in that they are passive, requiring neither vocal nor manual responses to provide workload estimates. In addition to unobtrusiveness, a second criterion imposed on all workload measures is that they are sufficiently **sensitive**. That is, if increases in task load are imposed of differing magnitude, the measure in question will also reflect those proportional differences in a reliable fashion. A third criterion also imposed on effective workload measures is that they be **diagnostic,** in the sense of signaling not only the amount of mental workload, but also the nature of that load, e.g., whether it is visual, auditory, imposed on working memory, or imposed on executive functioning (C. Wickens & Tsang, 2014). This use is not to be confused with a clinical diagnosis of a medical condition. The importance of diagnosticity in workload measures emerged with the development of multiple resource theory (Navon & Gopher, 1979; C. D. Wickens, 1980). This development provided the realization that different "fixes" for a workload-overload situation should depend, to some extent, on *which* resources were overloaded, and not just *that* "resources were overloaded."

Temporal responsiveness, the fourth criterion, is particularly relevant for physiological measures, and applies to any workload measure that is intended for adaptive automation. If changes to automation are to be based on an assessment of momentary capabilities of the human operator, and these capabilities are driven, in part, by dynamic fluctuations of the load imposed by the task (either on all resources, or specific resources), then it is essential that a **fully reliable** workload estimation be provided within a time interval less than the bandwidth (fluctuation rate) of the task demands. If data collection, workload inference, and adapting automation takes too long, then the environment and workload may have already changed, mitigating the need for that adaptation. Alternatively, if an *imperfect* inference of workload is made within less time, then in the case when the inference is wrong (and the degree of automation is lowered or raised, when it should have been raised or lowered respectively), trust in the AA system will rapidly erode. This is particularly relevant for AA, given that adaptive changes to interfaces made by intelligent automation agents should be guided by knowledge of which resource is overloaded.

It is important to note that in this paper we do not examine AA directly. Instead, we evaluate the promising characteristics of fNIRS that may allow it to serve as a vital component of unobtrusive AA systems, by using group-level statistical analyses to demonstrate the utility of fNIRS for measuring different types of cognitive load, which is a hallmark of multiple resource theory. In particular, fNIRS has not been systematically evaluated in prior literature through our

four key criteria of unobtrusiveness, temporal responsiveness, sensitivity, and diagnosticity. This paper therefore makes three contributions to the workload-based adaptive automation domain: 1) We conducted a meta-review of the literature on fNIRS and workload in the HF and HCI realms, and use this literature to explore temporal responsiveness, unobtrusiveness, sensitivity and diagnosticity. 2) Based on the limited and inconclusive research to date on the topic of fNIRS as a measurement modality for workload-based AA, we describe seven standards of empirical studies (e.g., multiple brain regions measured, use of a complex task, multiple load levels manipulated, use of additional workload measures as manipulation checks, suitable N, investigation of both HbO and HbR) that are needed to advance the field of neuroergonomics with respect to workload-based AA. We describe findings from the small handful of studies in our meta-review that satisfy these standards. 3) We then designed an experiment to empirically evaluate the utility of fNIRS for further examination of sensitivity and diagnosticity. Evaluating the ability to measure different load levels (specificity) in different cognitive resources (diagnosticity) is a crucial step toward realizing the goals of workload-driven AA.

The rest of this paper is organized as follows: First we describe the fNIRS signal in detail. Next, we describe our meta-review and summarize the findings. We then describe our experiment methodology (n = 43) and we present our results and interpretations. We interpret our findings in light of the meta-review findings, comparing and contrasting our results with the prior related work. Finally, we describe limitations of our work and avenues for future work.

2 Brain Measurement and Functional Near-Infrared Spectroscopy

When the brain reacts to a stimulus, neurons send electrical signals down the network of interconnected neurons that are recruited to handle the stimuli. These electrical potentials can be measured with EEG with excellent temporal responsiveness. Unfortunately, EEG has poor signal to noise ratio and spatial resolution (Duan, Liu, & Lian, 2021; Kwon, Shin, & Im, 2020; Putze et al., 2014). These challenges are partially overcome by technologies like fMRI and fNIRS, which can measure the hemodynamic response of blood flow rushing to the area of these electrical potentials to support neuronal activation. While the fMRI represents the gold standard for spatially accurate measurement of the functional human brain, it is not practical for measurements in naturalistic HCI settings. fMRI restricts movement within a scanner and is cost prohibitive, while fNIRS is less expensive, less obtrusive, and offers information highly

correlated to fMRI's BOLD signal (Cui, Bray, Bryant, Glover, & Reiss, 2011; Lai, Ho, Lim, & Ho, 2017). Following increased neural activity is an increase in cerebral blood flow which generally causes an increase in HbO and decrease in HbR (Logothetis, Pauls, Augath, Trinath, and Oeltermann (2001), see Figure 1). fNIRS pulses near infrared light into the brain cortex (in the wavelength range of 650nm – 900 nm) and it measures the blood-flow in the cerebral cortex and the signal may be influenced by systemic physiological factors like respiration and Mayer wave oscillations. Researchers have found that the HbO signal is more affected by these systemic factors (hence contributing "noise" to a workload estimation) than is the HbR signal (Dravida, Noah, Zhang, & Hirsch, 2017; Huppert, Franceschini, & Boas, 2009; Obrig et al., 2000; Q. Zhang, Strangman, & Ganis, 2009; Yiheng Zhang, Brooks, Franceschini, & Boas, 2005). As shown in Figure 1, the hemodynamic response measured by fNIRS (ΔHbO and ΔHbR) is characterized by quick steep peaks in HbO, then HbR, followed by eventual plateaus in both. There are approximately two seconds between HbO and HbR peaks, both happening within approximately eight seconds of stimulus onset (Huppert, Hoge, Diamond, Franceschini, & Boas, 2006). Thanks to rapid developments in biotechnology in recent years, newer fNIRS devices are now portable, wireless, and they offer large numbers of channels across the outer cortex of the brain, allowing for brain measurement in naturalistic settings. Several recent papers provide an excellent overview of recent advances in fNIRS signal processing, analysis techniques, and biotechnology domains (Hasan Ayaz et al., 2022; von Lühmann et al., 2021; M. A. Yücel et al., 2017; H. Zhao & Cooper, 2018).

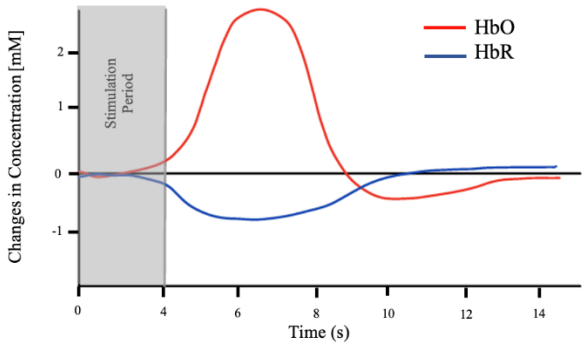


Figure 1: Typical time response of HbO and HbR after stimulus (such as completing a n-back task). HbO peaks between 6-8s following the stimuli and HbR dips at the same time.

3 Meta-Review of fNIRS and Workload Research

A substantial body of work has investigated the use of this device in HCI literature. In this section, we present a meta-review of this work. We included two digital libraries in our review: The ACM digital library (for venues such as ACM Transactions on Computer Human Interaction (TOCHI), ACM Special Interest Group on Computer Human Interaction (SIGCHI)), and Frontiers digital library (for Journals such as Frontiers in Neuroscience, Frontiers in Neuroergonomics). To collect relevant literature across these venues, we used the search terms "fNIRS" OR "NIRS" OR "functional near-infrared spectroscopy" OR "near-infra-red spectroscopy" AND "workload". Of the resulting articles, we reviewed the abstracts and removed duplicate articles and we filtered the articles and only retained papers that manipulated task difficulty level in some way with the goal of measuring workload (this included studies that looked at *rest/no load* vs *task on/load*, as well as studies with more fine-tuned levels of load, such as 0back, 1back, 2back tasks). This resulted in 54 published studies that examined fNIRS as a cognitive workload measure (see Appendix 1). In the following sections, we analyze these studies with respect to our key criteria of interest: unobtrusiveness, temporal responsiveness, sensitivity, and diagnosticity, below.

3.1 Unobtrusiveness

Thanks to rapid developments in biotechnology in recent years, newer fNIRS devices are now portable, wireless, and they offer large numbers of channels across the outer cortex of the brain (Pinti et al., 2020). An unobtrusive device also provides the opportunity to gain access to data while people perform tasks in naturalistic settings, and for longer continuous durations. The 'device type' column in Appendix 1 lists the device and the 'fNIRS Set Up' column lists the number of channels and regions used in each of the studies in our meta-review. We found the most common and median number of channels used was 16, and the mean was 26 channels. There was a single study with two channels, and three studies with over 100 channels. Our review also found the most common device was the OxyplexTS from ISS, followed by various models from NIRx and Hitachi. Several fNIRS companies have developed wireless versions of their technology including (but not limited to) NIRx, Artinis, and Obelab. Recent work has also explored development of highly lightweight, wearable fNIRS systems that can be ergonomically developed for specific operational settings such as team crisis management (Xu, Slagle, Banerjee, Bracken, & Weinger, 2019). fNIRS probes have been further developed to measure different depths within the brain (e.g., short-distance channels). Optodes with less distance

between them penetrate less into the brain, measuring only the superficial layers which in turn can then be used as regressors in analyses such as the oft used general linear model approach (Wyser et al., 2020). Taken together, these advancements in biotech continue to push the envelope of innovation, resulting in highly unobtrusive fNIRS devices, and we expect this trend to continue, with future fNIRS systems being developed to meet the needs of specific use cases.

3.2 Temporal Responsiveness

Based on our meta-review, time durations used for fNIRS analysis vary greatly (see 'time window' column in Appendix 1), but the **vast** majority of papers analyze time windows of data that are well over that deemed acceptable for adaptive automation, given the relatively high bandwidth of task demand changes in most applied contexts. In fact, out of 54 papers identified in Appendix 1, only eleven papers include analyses that look at window sizes of ≤ 15 seconds (i.e., for adaptive automation, capable of estimating workload fluctuations of 2 cycles/minute; Girouard et al. (2009); Herff et al. (2014); L. Hirshfield et al. (2011); Nazeer et al. (2020)). Researchers took a thorough look at time windows existing in the literature and reported that a window size of 2-7s following a stimulus led to increased classification accuracy compared to other time windows (R. Liu, Walker, Friedman, Arrington, & Solovey, 2021; Nazeer et al., 2020). Similar bodies of research have proposed a 0-4s window for drowsiness detection using fNIRS (Khan, Liu, Bhutta, & Hong, 2016). Another method for window detection with fNIRS is a moving window method which explores all windows to find the best window for classification. Researchers have used varying window lengths such as 3s (Shin et al., 2017) and 10s (Herff et al., 2014). In recent work, researchers used moving windows of 5s, 10s, and 15s with 1s step size along with their proposed individual-based time window selection (ITWS) algorithm for group-level classification which considers how the best window may vary between participants (R. Liu et al., 2021). They found a 5s window achieved the highest average accuracy (F1 score) and applying their ITWS algorithm to the 5s window achieved the highest performance. These findings throughout the literature suggest that optimal window sizes vary between participants and tasks and suggest that a moving window method may result in the best classification accuracy while using fNIRS. It is important here to realize also that the measure of "performance" (e.g., classification accuracy of a high vs low workload state) has a very stringent criterion in adaptive automation: for example, a classification accuracy of 80% would be

insufficient: a 20% error rate in classifying high vs. low workload could erode trust in the system.

It is also important to emphasize that establishing the feasibility of temporal responsiveness for adaptive automation must be based on individual participant data, rather than group averaging effects. By averaging, the latter data will reveal a smoother and more reliable response curve of the workload measure than any individual response. Yet, by definition, adaptive automation must be based upon the responsiveness of the individual.

Because the underlying hemodynamic response is inherently slow, several researchers have begun to integrate both EEG and fNIRS (Aghajani, Garbey, & Omurtag, 2017; L.-C. Chen, Sandmann, Thorne, Herrmann, & Debener, 2015; Y. Chen et al., 2020; L. M. Hirshfield et al., 2009 ; Pike, Maior, Porcheron, Sharples, & Wilson, 2014; Putze et al., 2014; Yujin Zhang & Zhu, 2020), taking advantage of the better temporal resolution of EEG and better spatial resolution of fNIRS to better quantify the brain's response to stimuli. The two modalities are complementary in nature as EEG measures the electrical response and fNIRS measures the metabolic response to brain activity (Putze et al., 2014).

Research on the temporal responsiveness of fNIRS has not reached a level of maturity where we can confidently say that responsiveness needed for realistic AA is achievable with fNIRS on its own. What we know about the hemodynamic responses measured by both fNIRS and fMRI certainly suggests that responsiveness of fNIRS is slower than needed for realistic AA systems, given their accuracy requirements. But the body of machine learning accomplishments to date on different sized sliding windows, and work on hybrid fNIRS/EEG, suggests a path forward for researchers to continue to explore.

3.3 Sensitivity and Diagnosticity (in Controlled Tasks)

Currently the literature on HbO/HbR sensitivity to measurement of workload is skewed heavily toward studies using tightly controlled psychological tasks (e.g., stroop, n-back tasks) to manipulate load. Furthermore, these studies are skewed toward evaluation of HbO over HbR, with many cognitive load studies using only the HbO data in the analyses. Spotlighting HbO makes sense in light of the strong response where oxygenated blood in the brain floods to regions where neurons are firing, often referred to as "watering the entire garden for the sake of one thirsty flower" (Malonek & Grinvald, 1996). This significant increase in HbO is due to a metabolic increase resulting in a flush of oxygen that exceeds the metabolic needs of the

neurons, resulting in overcompensation (Malonek & Grinvald, 1996). This overly amplified HbO response was helpful for early fNIRS studies, when devices had few channels, while exploring the effect of simple tasks. Undoubtedly, the HbO signal gives a stronger response to neural activation than HbR, and it has shown strong responses in studies with simple stimuli such as n-back tasks (Hasan Ayaz et al., 2012) and Stroop tasks (L. Hirshfield et al., 2011). As described next, when the task becomes more complex, the oversaturation of HbO can result in that measure losing its diagnostic value.

Herff et al. (2014) evaluated both HbO and HbR during an increasingly difficult task demanding working memory (the "n-back" task). In terms of **sensitivity**, both HbO and HbR had steeper slopes during the 3-back test when compared to the 1- and 2- back tests. This suggests that both HbO and HbR have similar sensitivities to WM in a controlled environment with a well calibrated task; but other studies in more complex settings found different patterns in activation between the two measures. One study by Dravida et al. (2017) found HbO a more reliable signal than HbR), in response to increasing mental workload imposed by simple motor tasks. HbR offers an additional benefit of being highly coordinated with fMRI Blood Oxygen Level Dependent (BOLD) signals (Cui et al., 2011; Foy, Runham, & Chapman, 2016; Huppert et al., 2006; MacIntosh, Klassen, & Menon, 2003; Schroeter, Kupka, Mildner, Uludağ, & von Cramon, 2006). Additionally, the HbO signal has shown slow variable drift over a task while the HbR signal did not (Unni, Ihme, Jipp, & Rieger, 2017). Because HbO is more susceptible to drift and systemic artifacts, HbR may be a more reliable measure of workload in complex tasks. Because of the inconclusive results in literature, the current study aimed to compare the relative benefits that HbR has in measuring workload as compared to the HbO signal. Another complication in workload-focused fNIRS research to date that leads to inconclusive sensitivity results concerns the correlation between task performance and cortical activation (Kimberly L Meidenbauer, Choe, Cardenas-Iniguez, Huppert, & Berman, 2021). While some researchers have found increasing brain activation with increasing task difficulty (linear effect), such as the n-back test (Hasan Ayaz et al., 2012; Fishburn, Norr, Medvedev, & Vaidya, 2014; Kuruvilla, Green, Ayaz, & Murman, 2013), others have found that increased task difficulty is not always associated with an increase in HbO and decrease in HbR signals (non-linear effects) (Aghajani et al., 2017; Herff et al., 2014; Mandrick, Chua, Causse, Perrey, & Dehais, 2016). This non-linear activation with task difficulty suggests that participants may reach a maximum level of activation after difficult

tasks (Mandrick et al., 2013), or participants may simply disengage from a task that is too difficult (Causse, Chua, Peysakhovich, Del Campo, & Matton, 2017).

In terms of **diagnosticity**, only one paper, by Putze et al. (2014), took a close exploration of diagnosticity of different types of load with fNIRS in a controlled task setting. Their goal was to differentiate between visual and auditory load and participants were presented with movie and audio clips, i.e., silent movies (no sound; VIS), audiobooks (no video; AUD), and movies with both video and audio (MIX). They measured brain regions associated with visual and auditory processing and found that visual load activated regions in the occipital cortex, while auditory load did not engage that region. While the subject pool was relatively small (n = 12) and the tasks were tightly controlled, these results show promise for fNIRS as a modality for diagnosticity of workload.

3.4 Seven Standards for Experiments to Advance Workload-Based AA in Neuroergonomics

The meta-review papers described above involve simple and tightly controlled tasks such as n-back tasks or presentation of video and audio clips to manipulate visual/auditory processing. To further explore sensitivity and diagnosticity for workload-based AA, there is a need to consider more complex study designs and task contexts. Thus, we filter the papers in Appendix 1 in light of what we consider to be seven standards/features that we judge to be important to evaluate (as seen in Table 1), as fNIRS is considered as a measurement tool for workload-based AA.

Table 1: Seven standards for experiments needed to evaluate fNIRS for workload-based AA.

(1) Participants should perform a complex task typical of real-world human-computer interactions.
(2) Workload should be experimentally manipulated in a controlled manner to impose greater or lesser cognitive demands (going beyond just load on/off), in order to evaluate sensitivity of different load levels on a specific resource.
(3) Studies should focus on different specific resources within a multiple resource structure, hence examining diagnosticity
(4) The validity of experiment task manipulations should be assured by including additional workload measures, such as self-report workload, response time, performance, and pupil diameter.
(5) To further examine the diagnosticity of the measures, researchers should measure multiple functional brain regions of interest (ROIs), ideally mapped onto the multiple resources identified in the experimental design, in order to determine if specific ROIs are differentially sensitive to the workload manipulation assumed to be reflected by increased activation there.
(6) Increased activation should be explored via the two different fNIRS measures of HbR and HbO.
(7) Finally, studies should have adequate statistical power, with a suitable N.

Of the studies examined we judged that NONE satisfied all 7 standards listed in Table 1. Table 2 presents the set of studies reviewed (Appendix 1) that adhered to at least three of the standards listed in Table 1, ordered by the number of standards adhered to. We did not rate adherence to the power standard as this could not be represented as a dichotomous variable and

applying any particular N level as a criterion seemed to us to be quite arbitrary. As shown in the list, most studies adhered to no more than 3 standards, and these are described in some detail in the appendix. Only those adhering to 4 or more are described in detail below.

Table 2: Meta-review studies reviewed that adhered to atleast three of the seven standards, ordered by # of standards adhered to.

Author	# standards	complexity	workload manipulation	diff resources	convergent measures	multiple ROI	HBO-HBR	N
Isbilir	3		y		y		y	14
Chu	3	y	y				y	20
Lei	3				y	y		131
Hamann	3	y	y	y				35
Izzetoglu	3	y	y		y			8
Peck	3	y		y	y			16
Solovey	4	y	y		y		y	48
McKendrick	4	y			y	y	y	20
Ayaz	4	y	y		y		y	16
Durantin	4	y	y		y			12
Kerr	5	y	y		y	y	y	7
Putze	5		y	y	y	y	y	12
Gateau	6	y	y	y	y	y	y	28

3.5 Sensitivity and Diagnosticity (as assessed in studies that align with the seven standards)

Given the seven standards from Table 1, many papers were filtered from an in-depth review because they either did not explicitly **manipulate** (feature 2) workload of an ongoing task (instead comparing to a task performed with a resting state; Geng, Liu, Biswal, and Niu (2017)), or compared two qualitatively different tasks (Putze et al., 2014) or had what we judged to be an inadequate sample size (feature 7). Furthermore, many of the remaining studies were not included in our in-depth review because the contributions of specific functional ROIs could not be identified from the article text (feature 5). This included studies that focused on machine learning classifiers of low vs. high workload (Asgher et al., 2019). In some cases, no convergent workload measures were reported (feature 4); or the task was a very basic cognitive task like the

N-back memory task (lacking complexity feature 1), even as such tasks may indeed reflect a cognitive component of a complex real-world task. It is important to note that Afergan et al. (2014) **did** report the success of fNIRS to operate in an adaptive automation system with a complex real-world task (control of multiple unmanned vehicles), but they did not experimentally **manipulate** (feature 2) the complexity of that primary task.

One paper by Hasan Ayaz et al. (2012) manipulated levels of workload during n-back, air traffic control (ATC), and unmanned air vehicles (UAV) tasks. They found that oxygenation increased within the prefrontal cortex (PFC) with increased task difficulty for the ATC experiment. They also found that data communication requires less cognitive resources than voice communications in the ATC simulation. During the UAV task, they found that expertise tends to be associated with lower brain activity in the prefrontal area. These researchers used oxygenation (HbO-HbR) as their fNIRS measure, rather than looking at the two values separately. However, they did not perform ROI analysis and focused only on the PFC with 16 channels. Another paper by E. T. Solovey, Okerlund, Hoef, Davis, and Shaer (2015) investigated how stereo vision and vibrotactile feedback affect user interaction during a spatial task with interactive 3D displays with three levels of difficulty, although they did not systematically compare different workload manipulations (feature 3). They observed difficulty effects on both average HbO and HbR values and vibrotactile feedback on HbO only. While the number of fNIRS channels were limited (10), they had a simulated "real-world task," a large N (48), integrated other subjective measures such as the NASA-TLX, and modified workload levels. Another paper by (Peck, Yuksel, Ottley, Jacob, & Chang, 2013) manipulated the visual/cognitive workload imposed as participants performed a data comparison task (demanding both visual perception and working memory) on either bar graphs (presumed to be low workload) or pie charts (presumed to require high workload). While finding no overall workload effects on fNIRS, subjective ratings, or performance, they did observe individual differences such that those who rated the pie graph to be more difficult, reflected this in the fNIRS measure (HbR) when using the pie graph whereas those who rated the bar graph to be more difficult also showed that same pattern (more oxygenation as reflected by HbR) when using the bar graph. The investigators however did not compare ROIs, nor the two fNIRS measures (HbR and HbO), nor did they include more than the single manipulation (graph type) of mental workload. Another paper by R McKendrick et al. (2016) carried out an applied study in a very real-world context

(natural navigation (feature 1)). They did not explicitly manipulate workload, but did compare navigation supported by a map on either a head mounted display (HMD; which they inappropriately labeled as AR) or on a hand-held display (HHD). The HHD smart phone, by virtue of the smaller map image and the greater requirement for scanning, was assumed to impose greater workload (feature 2), an assumption confirmed by convergent measure of secondary task performance (feature 4). The fNIRS results, measured at left and right PFC ROIs suggested greater sensitivity to this display manipulation of workload while performing an n-back task (and more consistent effects across error and correct trials) for HbR than for HbO [(feature 6) see their Figures 4 and 3 respectively]. Also, these effects appeared to be affected by ROI (feature 5), being more sensitive and consistent for Left Lateral PFC (LLPFC) than for Right Lateral PFC (RLPFC). 20 participants provided them with adequate statistical power (feature 7).

Two aviation studies provide perhaps the closest match of features to the current investigation. In the first study, Durantin, Gagnon, Tremblay, and Dehais (2014) employed a low fidelity desk top flight simulator (feature 1) with workload manipulations (feature 2) along two separate dimensions (feature 3): the dynamics and bandwidth the tracking task whereby the participant followed a target aircraft (perceptual-motor load), and the cognitive complexity of the rule dictating which aircraft to track (cognitive load). Convergent workload measures (feature 4) of subjective ratings and heart rate variability were collected; but only one ROI was measured and only for HbO. While both convergent measures validated the two workload manipulations, the findings for fNIRS were somewhat puzzling. At low cognitive load, the increase in perceptual/motor load did indeed produce increased oxygenation signaled by HbO; but at high cognitive load the reverse effect of increasing perceptual/motor load was observed. The investigators also reported a positive correlation (over participants) between HbO and the level of performance observed, signaling, presumably, the greater cognitive effort required for a participant to perform better. Statistical power was barely adequate (N=12; Feature 7).

The second study by Gateau, Ayaz, and Dehais (2018) involved aircraft piloting (feature 1). Manipulations of workload (feature 2) were imposed along two separate dimensions (feature 3): the working memory demands of air traffic control (ATC) communications (cognitive load) and whether fNIRS was recorded in a flight simulator or in the actual aircraft during flight, with working memory manipulated via flight parameter instructions given to participants. The

convergent measure of the increasing multi-tasking workload from simulator to flight is based on task analysis, and that from increasing ATC communications load was validated by embedded secondary task performance (communications errors; feature 4). Four prefrontal ROIs (feature 5) were assessed for both HbO and HbR (feature 6). Their data signaled a strong positive effect of communications load on HbO. While the increasing load from simulator to the aircraft had no significant main effect on HbO, the two workload manipulations did interact, in conjunction with ROI to suggest that the increased cognitive load had a much greater effect on HbO in flight than on the simulator; however, this enhanced effect was only observed at one ROI which appears, from their figure to be the left medial prefrontal cortex, perhaps an ROI related specifically to multi-tasking capabilities. The investigators collected data on HbR but did not report it; only stating that HbR did not show the workload X ROI interaction described above, and thereby suggesting HbR to be less sensitive. Statistical power was adequate (N=28; feature 7), but this concern was mitigated using trained pilots as participants.

The research conducted to date suggests a complex interplay between task complexity, practice effects, and human performance, which all have an effect on fNIRS measures of HbO and HbR taken from the outer cortex of the brain. The literature outlined above suggests that assessing the utility of fNIRS for sensitivity measurements of workload is a complex problem, with more empirical work needed. In terms of diagnosticity of the fNIRS signal in response to different types of load (auditory, visual, memory, etc.), very little research has been performed. Therefore, more work is needed to evaluate the diagnosticity of fNIRS for its suitability for AA.

4 Methods

Noting the lack of research from the meta-review addressing the sensitivity and diagnosticity of fNIRS, we designed an experiment to address the following three research questions:

RQ1: Is fNIRS sensitive to Working memory (WM) manipulations (high/low) in a complex task environment?

RQ2: Is fNIRS sensitive to Visual Load (VL) manipulations (high/low) in a complex task environment?

RQ3: Is fNIRS diagnostic to the type of load, specifically when considering VL versus WM?

RQ4: Are HbO and HbR differentially sensitive to (and therefore diagnostic of) these two different workload manipulations?

To address these questions, we designed a testbed environment to enable us to explore sensitivity and diagnosticity in a study that checked off the seven standards in Table 1, where workload was (2) **experimentally manipulated** in a controlled manner to impose greater or lesser cognitive demands on **(3) different specific resources** within a multiple resource structure, and the validity of our manipulations was assured by (**4) examining additional workload** measures, to assure the sensitivity of our measure to demand manipulations that were also reflected in those other subjective, secondary task, and physiological measures. Simple measures of *performance* on the task whose workload is manipulated are inadequate for assessing mental workload (C. Wickens & Tsang, 2014). Hence, we validate our workload manipulations against three conventional and well-established workload measures: secondary task performance, subjective ratings and the physiological measure of pupil diameter. To examine the diagnosticity of our measure, we clearly identified **(5) multiple functional regions of interest (ROIs)**, to determine if different ROIs were differentially sensitive to the workload manipulation assumed to be reflected by increased activation there. Within each ROI we also explicitly (6) **compared the two different fNIRS measures**, HbR and HbO. Finally, our study had a large amount of (**7) statistical power,** given its high N.

4.1 Testbed

The task is a shape sorting task (Fig 2) that involves sitting in front of a large monitor while wearing headphones. The task was based on a previous task used by our team that facilitates clean manipulation of multiple types of workload while aligning with dimensions of task domains in which we foresee adaptive automation being employed (N. Tran et al., 2021). Specifically, in other lines of our work we have been motivated by future scenarios in which robotic teammates collaborating with humans in mixed reality environments will adaptively select between communication strategies based on level and type of cognitive load. Inspired by this vision, we implemented a mixed reality interaction task inspired by pick-and-place tasks common to current industrial human-robot collaborative environments. This mixed reality domain was then used to prototype workload manipulations that leveraged the structure of the mixed reality task environment. Finally, the overall structure of this mixed reality testbed was used to inform the design of the 2D testbed used in this work, which aligned with the domain-

based design objectives of the mixed reality testbed while avoiding hardware constraints and sources of noise that were not necessary for the purposes of the present experiment.

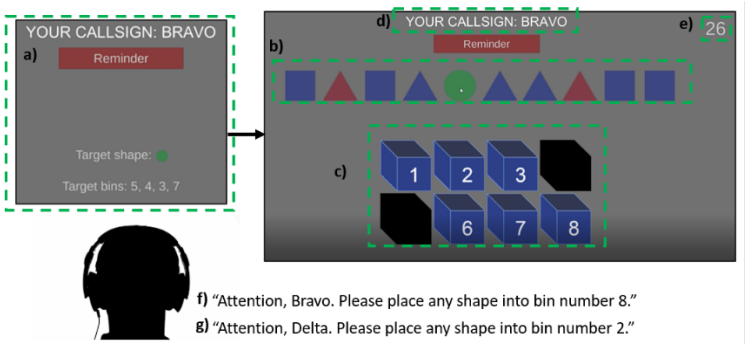


Figure 2: The shape sorting testbed. The a) instruction screen directs the participant on the primary task target shape and target bins. Participants then sort the correct target shape out of a list of possible shapes (b) into the correct numbered bins (c). Participants are assigned a callsign (d) and a secondary auditory task is presented through the right side of the headphones, where the information can either be ignored (g) or where it must be attended to (f). Each task session lasts 45 seconds with time being counted down (e), before filling out surveys and beginning a new task, with new updated instructions.

Primary Task. At the beginning of each task, the participants are shown an instructions screen (Fig 2a). They are instructed to search for a specific target colored shape (e.g., green circle) and to place that shape into bins with specific numbers (e.g. bins 3,4,5,7). When participants have read the instructions (Fig 2a) they click on a 'begin task' button. At that time, the instructions screen is replaced by the task screen (Fig 2b-e). When the task begins, the participant has 45 seconds to identify every target shape and to sort it into a valid bin, while the shapes are continually refreshed (swapping one shape out for another) every 4 seconds. To ensure target shapes appear often, a timer of 6 seconds is also included in the task. If the target shape is not present on the screen when the time elapsed, the target shape is swapped for one of the distractor shapes during the next refresh cycle. The shapes and target bins are shown in Fig 2b and 1c, respectively. While the participants complete the tasks, some bins are randomly blacked out for a few seconds at a time, so that shapes cannot be placed into them (2c). Shapes are sporadically blacked out in such a way that there is always at least one bin accessible to the participant for placement into a target bin. For example, if the target bins are 5, 4, 3, 7 (Fig. 2a) there will never be a time when all four of those bins will be inaccessible. This is done to ensure that participants keep all bin numbers in working memory throughout each task. A timer on the screen (Fig 2e) counts down from 45 seconds while the participants sort the shapes.

Secondary Auditory Task. While participants do the continuous task and search for their specific target shape, they are simultaneously monitoring auditory information being played through headphones. Each participant is assigned a callsign (Fig 2d), which is Bravo. Two types of auditory information are randomly played through the headphones on average every 15 seconds. Distractor audio information is played at times (Fig 2g) when the callsign does not match the participant's callsign. They are told that they can ignore this information. A target auditory task (Fig 2f) uses the participant's actual callsign of Bravo, and a request is made to the participant to place an additional shape (of a different shape/color than the primary task target shape) into any bin. When this target callsign of Bravo is used, the participant must quickly sort the secondary task shape, and then return to the primary task.

Working Memory and Visual Perceptual Load Manipulations

Working Memory Load (WM) is either low or high, depending on the number of bins needing to be remembered. Low WM has 2 bins, while high WM has 4 bins to be memorized. Visual Perceptual Load (VL) is either low or high, depending on the similarity metric between the target shape and the rest of the shapes available at the top of the screen (see Lavie's foundational work on visual perceptual load for more detail; Lavie (1995); Lavie, Hirst, de Fockert, and Viding (2004)). High Similarity (high VL) is defined by a sort distractor object sharing one property with the sort target object in terms of their shape or color. For example, a green circle and a green square are considered similar as they both share the same color feature. A red circle and a blue square are considered dissimilar as they share neither of the color or shape features. There are a total of nine different sort objects defined by the combination of Color = [red, green, blue] and Shape = [circle, square, triangle]. Once a target shape is selected, the distractors are selected from a subset of that total, which matches the VL for that condition, resulting in four distractor objects for each type of perceptual load. Figure 3 shows an example of low VL (top) and high VL (bottom). In low similarity (Low VL) the target shares no features in common with the distractors.

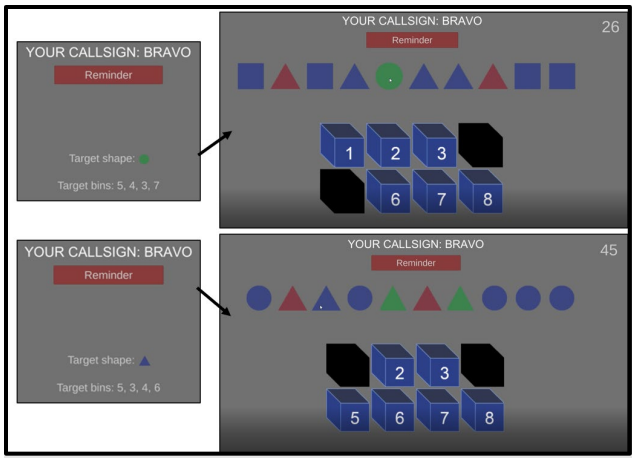


Figure 3: Left (instructions before a task begins). Top right: An example of a search task where the VL is low because target and distractors share no features in common. Bottom right: an example of a search task where the VL is high: 1 feature is shared. See (Nhan Tran et al., 2021) for an example of this task implemented in a mixed reality context.

4.2 Experiment Protocol and Procedures (IRB Protocol #19-0436)

In its simplest form, the control task in the testbed involved sorting of shapes in such a way that it elicited low levels of WM, VL, and AL ($L^{wm}L^{vl}L^{al}$), which is our control task. From there, the testbed was configured to experimentally manipulate WM and VL between low and high levels, while keeping AL low. Thus, our experiment had four conditions where load levels of WM and VL were modulated between low and high, while AL was maintained at a continuous low level:

- $L^{wm}L^{vl}$ L^{AL}(control)
- $L^{wm}H^{vl}$ L^{AL} (VL modulated to high)
- $H^{wm}L^{vl}$ L^{AL} (WM modulated to high)
- $H^{wm}H^{vl}$ L^{AL} (Both WM and VL modulated to high)

We note that since the auditory load (AL) secondary task was always set to a low load level throughout all conditions, we omit that redundant item in our results section (e.g., $L^{wm}L^{vl}$ L^{AL} becomes $L^{wm}L^{vl}$). Equipped with high-density fNIRS, we identified four regions of interest (ROIs) to measure in the brain, enabling us to specify the type of load experienced by our participants (**diagnosticity**). These four ROIs included brain regions associated with WM, VL, and AL, as well as a critical multitasking (MT) region that becomes engaged during complex multitasking scenarios, where users coordinate their short- and long-term goals and intentions with the immediate constraints of the task environment (Tomasi, Chang, Caparelli, & Ernst, 2007).

43 participants completed this experiment (51% female, median age = 23 years). All participants were recruited from a population consisting of staff, faculty, and students at a large university in the Western United States. After providing informed consent, participants were equipped with the neurophysiological sensors (described below), and earbud headphones were placed into each ear to deliver the auditory secondary task. Next all participants went through a tutorial to learn how to complete the task. They did an example task and had the opportunity to ask questions from the researcher before beginning the experiment. There were four conditions in the experiment, with each combination of WM (high/low) x VL (high/low). Participants then completed 24 continuous 45-second-long tasks, with the four conditions described above presented in a randomized block design order, while the AL secondary task continued at a low load level, but continuous pace, throughout the experiment. After each trial, they completed the mental demand item from the NASA-TLX, workload rating scale (Hart & Staveland, 1988).

As shown in Figure 4 (left), participants were equipped with a desk-mounted eye tracker (Tobii 4c), and functional near-infrared spectroscopy (NIRx Sport 2) with a custom montage designed to measure regions of interest (ROIs). The montage included 42 measurement channels, as shown in Figure 7. We selected the montage to cover regions of the brain including the frontal, visual, and auditory cortical regions that have been implicated in prior cognitive load research on working memory, visual load, and auditory working memory load (Crottaz-Herbette, Anagnoson, & Menon, 2004; Muller-Plath, 2008; Putze et al., 2014; Suh et al., 2019; Tomasi et al., 2007).

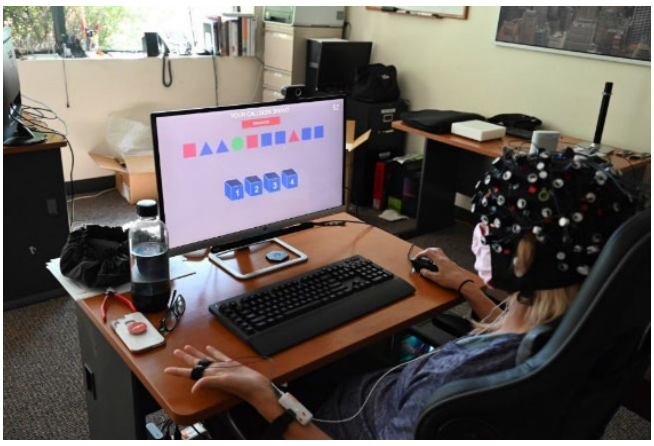


Figure 4: Sensor set-up included a Tobii 4c eye tracker and a NIRx Sport2 fNIRS device.

5 Results and Interpretation

5.1 Conventional Workload Measures

Of the 43 participants in the study, data from one participant were discarded because the behavioral data suggested that they did not participate in the task (e.g., they did not move shapes into bins at all). Thus, there were 42 participants used in the resulting analysis. Our conventional workload measures include self-report mental workload, secondary task accuracy and response time, and measures from eye tracking to further assess workload.

The response time (RT) data were recorded for each participants' response to a task. In both the primary and secondary tasks, the testbed recorded the number of milliseconds between when a target shape was presented on the screen (as was the case in the primary task) or when a target audio prompt was delivered (as was the case in the secondary task) to when the participant selected the target shape for sorting. Average sorting accuracies were calculated for both the primary and secondary tasks as the number of correct sorts divided by the number of total possible sorts. The number of total possible sorts is the sum of the number of correct sorts, incorrect sorts, and missed sorts. Because the RT data were skewed, prior to analysis the data were transformed via the inverse transform.

The results of the primary independent variables were analyzed using a 2 x 2 ANOVA, for each dependent measure. Specifically, the independent variables were (load level high|low) x (load type WM|VL). Results of primary task performance are shown in Figure 5. The left side of Figure 5 shows the data for primary task accuracy. The ANOVA revealed a significant increase in accuracy associated with increasing WM ($F=17.075$; $df = 1$; $p<.001$; eta squared = 0.016), no effect of VL ($F=0.066$), and a non-significant interaction between the two sources of load ($F=3.774$; $p=.0523$ eta squared = 0.003), seen in Figure 5, whereby the increasing accuracy with higher WM was attenuated at higher levels of VL.

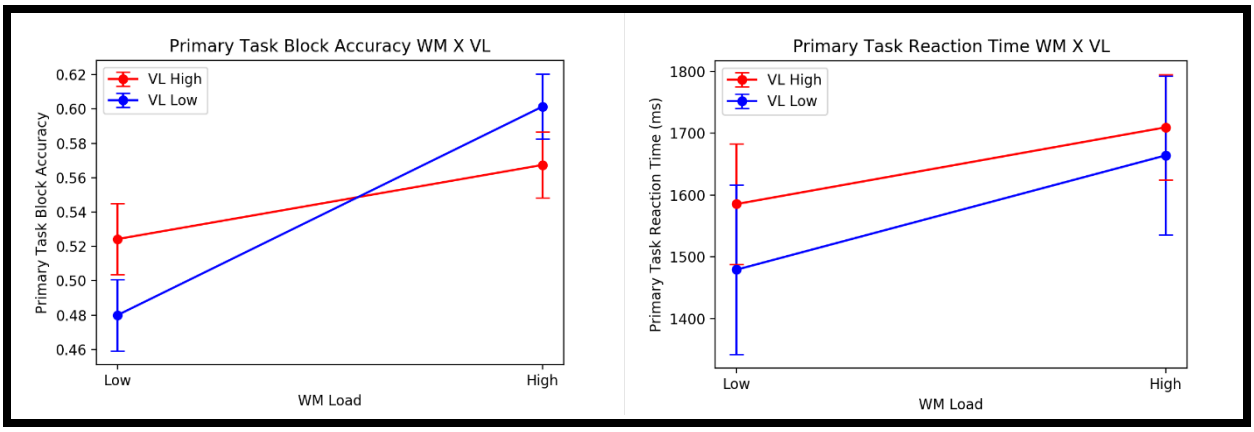


Figure 5: Left: Effect of WM and VL on primary task accuracy. Right: The effects of WM and VL on primary task RT. The error bars represent the unbiased one standard error as imple
mented by the Pandas sem function.

The right side of Figure 5 depicts primary task RT, where the ANOVA revealed highly significant increases in RT associated with both increasing WM (F=8.224, P<.01, eta squared = 0.013) and VL (F= 25.294, p<.001, eta squared = 0.042). The former effect, coupled with the data in Figure 6 suggests that the influence of WM produced a speed accuracy tradeoff. Higher WM produced more accurate responding (Fig 5, left) but at the cost of considerably slower processing (Fig 5, right). There was no interaction between the two variables.

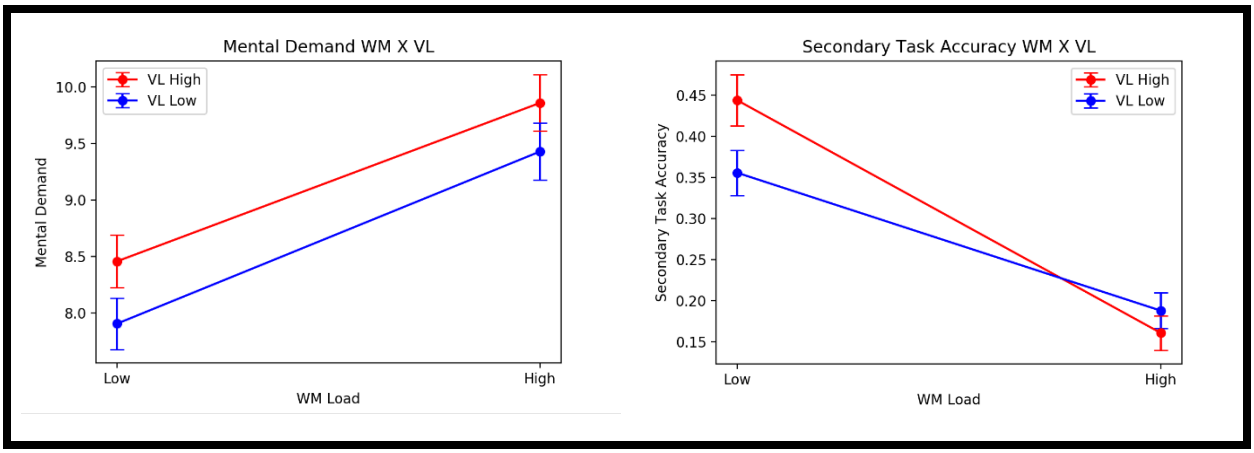


Figure 6. Left: Effects of WM and VL on self-report mental demand. Right: Effects of WM and VL on secondary task accuracy. The error bars represent the unbiased one standard error as implemented by Pandas 'sem' function.

Figure 6 depicts the effects of WM and VL on subjective mental workload as assessed by the mental demand sub-scale from the NASA-TLX (Hart & Staveland, 1988). There was a significant increase in mental load imposed by increases in both WM (F=36.868, p<0.01, eta squared = 0.034) and of VL (F= 4.183, p<.05, eta squared = 0.003) with no interaction. The right

side of Figure 6 shows the effect of VL and WM on secondary task accuracy. There was a significant decrease in secondary task accuracy associated with increasing WM ($F = 78.65$, $p<0.01$, eta squared = 0.069). The significant interaction of WM with VL ($F = 5.705$, $p<0.05$, eta squared = 0.005) signaled that the accuracy decrease imposed by WM was amplified at high levels of VL. There was no main effect of VL. The only effect observed on secondary task RT was the slowing caused by increasing WM load ($F = 7.424$, $p<0.01$, eta squared = 0.0208).

The eye tracking analysis was carried out on the data of a 36-participant subset of the 42 described above. Unfortunately, the eyetracking acquisition computer did not function properly during six data collections, resulting in data loss of six participants. The eye tracking features were generated using Tobii Pro Lab software. For each sample, Pro Lab reports the pupil diameter for each eye, the mental workload measure of interest in the current analysis. A mean value was computed of the samples corresponding to the time the task was undergone to obtain a value for each eye.

The analysis revealed that both left and right pupil diameter increased with increased WM load ($F=9.325$, $p<0.01$, eta squared = 0.0136) and left pupil ($F=9.459$, $p<0.01$, eta squared = 0.013). This aligns with prior eye tracking research which has repeatedly found pupil diameter to be a reliable measure of workload (Duchowski et al., 2018; Lohani, Payne, & Strayer, 2019), especially in visual attention tasks. The VL manipulation had no effect on pupil diameter.

In summary, the results from the conventional workload measures described above are conclusive: The degrading effects of increasing both WM and VL were quite pronounced; on response time of the primary task, and three conventional measures of workload (secondary task, subjective ratings and, for WM load, pupil diameter (C. Wickens & Tsang, 2014)). Furthermore, in general, when effect sizes are compared between the two manipulations, there was a considerably greater load imposed by higher WM than by the higher VL. Indeed, increasing VL had no effect on either pupil diameter or performance of the auditory secondary task. The only puzzling and unexpected effect in these data was the actual **increase** in primary task performance accuracy associated with increasing WM (Figure 5 left). We interpret this effect in terms of a strategic speed-accuracy tradeoff (C. D. Wickens et al., 2022) in which participants, encountering the need to retain more information in working memory when load is higher, are increasingly cautious and take significantly more time to carefully rehearse the item. Consequently, they are substantially slowed in their response, but accuracy is improved. The

mental workload imposed by this greater rehearsal processing is clearly expressed in the three conventional workload measures: secondary task RT, NASA-TLX, and pupil diameter. The dissociation between primary task performance and workload measures has been frequently observed (Yeh & Wickens, 1988).

5.2 fNIRS Workload Measures

In this section we look at the main effects of our independent variables, increasing visual and working memory load), as we did for the conventional measures. We look at these main effects on the fNIRS data, using two common techniques for defining regions of interest (ROIs). These are average-across channels ROI analysis and channel-specific ROI analysis, as detailed next.

Preprocessing Pipeline. All fNIRS preprocessing was conducted in NIRS AnalyzIR Toolbox in MATLAB (H. Santosa, Zhai, Fishburn, & Huppert, 2018). First, the raw voltages were down sampled to 4Hz and converted to optical density. The Modified Beer Lambert Law (Jacques, 2013; Strangman, Franceschini, & Boas, 2003) was then applied to convert optical density signals to oxygenated and deoxygenated hemoglobin concentrations using a canonical HRF basis set, which has been shown as the best performing basis set for longer task durations (Hendrik Santosa, Fishburn, Zhai, & Huppert, 2019). We then applied motion correction to the hemoglobin signals using the NIRS Toolbox's autoregressor function, which adds the NIRx accelerometer data as auxiliary data into the regressors in the generalized linear model (GLM) function. The GLM was applied with the default parameters, using an autoregressive, iteratively reweighted least-squares model (AR-IRLS) with pre-whitening to correct for serially correlated errors and motion present in the fNIRS signal (J. W. Barker, Aarabi, & Huppert, 2013).

Following preprocessing, we took the resulting ΔHbO and ΔHbR timeseries data and calculated subject level (first-level) statistics using a mixed-effects model (Kimberly L. Meidenbauer, Choe, Cardenas-Iniguez, Huppert, & Berman, 2020). The resulting first-level model contains the subject level regression coefficients as well as their corresponding error-covariance matrices per subject. We then used the subject level results to conduct t-tests at the group level (Kimberly L. Meidenbauer et al., 2020).

5.2.1 Sensitivity and Diagnosticity via Average-Across Channels ROI Analyses

Results from this analysis were then used for group-level contrasts between individual load levels (IVs) within specific pre-defined ROIs. We spatially register our fNIRS channels (C. Holmes et al., 1998; H. Santosa et al., 2018) onto anatomical brain regions in LONI space, with accompanying Brodmann Area (BA) labels (Jacobs, 2011; Shattuck et al., 2008). We used that information to identify four functional ROIs: WM, VL, AL, and multitasking based on the following literature: Tomasi et al. (2007) used fMRI to measure brain activation patterns during two sets of tasks with graded levels of cognitive load; including verbal working memory (WM) and visual attention (VA) tasks. They specifically outlined networks where WM and VA were activated during these tasks. Based on their ROI analysis, these researchers found that for both tasks, increased task difficulty resulted in increased BOLD responses in the parietal, occipital, and fusiform gyri, which relates to sensitivity. They also found that increased load increased the BOLD response in the inferior, medial, and middle frontal gyrus (BA 9) more strongly during WM tasks than VA tasks. Finally, they found only two regions to be activated uniquely to the VA task: the postcentral gyrus and superior occipital gyrus, which relates to diagnosticity. This information was used to define the VL and WM regions shown in Figure 7.

We expected to see interaction effects between our WM and VL condition combinations, so we also identified a multitasking ROI, where our goal was to focus on functional brain regions responsible for goal-directed multitasking. Since multitasking has been found to engage Brodmann Area 10 in the frontopolar region, we used that region to define our multi-tasking ROI (Mansouri, Koechlin, Rosa, & Buckley, 2017). Tomasi et al. (2007) found that increased WM caused greater activation in a frontoparietal network and was more pronounced for WM than VA tasks. Nevertheless, both types of tasks caused activation of this interconnected network. The frontoparietal network has previously been closely tied to the control of working memory (Wallis, Stokes, Cousijn, Woolrich, & Nobre, 2015). This is further evidence that the MT ROI has been implicated in both WM and VA tasks but shows more significant results for WM tasks. Lastly, we defined an Auditory Load (AL) region of interest (Figure 7), based on prior fMRI work on auditory perceptual load, to measure the effects of the WM and VL load manipulations on the auditory processing of the secondary auditory task.

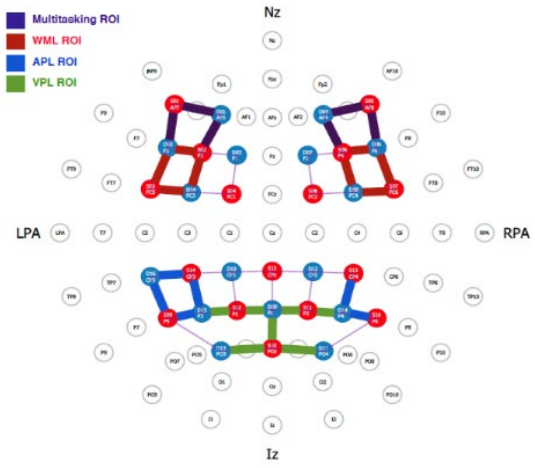


Figure 7: Regions of interest overlaid over the 42-channel fNIRS montage. Red circles represent light sources, blue circles represent light detectors. Red, green, blue, and purple lines represent a channel of measured data that falls into the WM, VL, AL, and MT ROIs, respectively. In the schematic picture, Nz represents the nasion, Iz represents the inion, LPA and RPA represent the left and right pre-auricular regions, respectively (used in standard EEG 10-20 landmarking).

We then ran contrast statements that corresponded to WM main effects (high WM – low WM) and VL main effects (high VL – low VL). Furthermore, for the HbO and HbR data, we performed ROI analysis using the ROI average function in NIRS toolbox, which averages the contrast statistics over the specified ROIs (more on this in H. Santosa et al. (2018); Zhai, Santosa, and Huppert (2020)). The result is a set of beta (ß) values and t-values for each contrast. ß-values are the resulting coefficients from the GLM and tell how well the data fit the expected hemodynamic response (canonical, in this case), which is a rise in HbO and decrease in HbR. While ß-values can be difficult to interpret on their own, they can be compared statistically through t-tests. The resulting t-values represent the results from the above two contrasts between conditions. Table 3 presents the effect sizes of the workload manipulations on fNIRS, showing HbO on the left half of the table and HbR on the right half and Figure 8 overlays the values from Table 3 over our ROIs on the brain. Since we employed an auditory secondary task to assess workload, we were also interested in the sensitivity of the auditory ROI (AL ROI) to the manipulations of both visual and WM, as will be discussed below.

Table 3: HbO and HbR ß-values and T-values for the main effects (IV stands for independent variable) of the WM and VL manipulations, averaged across each ROI (**bold face with a *** denotes significance ($p<0.05$)). The contrasts run are ($H^{wm}L^{vl}$ + $H^{wm}H^{vl}$) – ($L^{wm}L^{vl}$ + $L^{wm}H^{vl}$) for the WM load main effect (first row for each variable: ß or T) and ($L^{wm}H^{vl}$ + $H^{wm}H^{vl}$) – ($L^{wm}L^{vl}$ + $H^{wm}L^{vl}$) for the VL main effect (second row of each variable).

HbO						**HbR**				
ß	WM ROI	VL ROI	MT ROI	AL ROI		**ß**	WM ROI	VL ROI	MT ROI	AL ROI

WM IV	-0.65	2.59	-1.56	-0.29		WM IV	**-3.61***	**-2.60***	**-2.82***	-1.01
VL IV	**-4.26***	**3.70***	**-5.93***	-0.78		VL IV	**-3.52***	**-2.08***	-1.46	0.12
T	WM ROI	VL ROI	MT ROI	AL ROI		**T**	WM ROI	VL ROI	MT ROI	AL ROI
WM IV	-0.36	1.9	-0.67	-0.15		WM IV	**-5.07***	**-4.71***	**-2.48***	-1.23
VL IV	**-2.35***	**2.70***	**-2.53***	-0.4		VL IV	**-4.91***	**-3.76***	-1.28	0.14

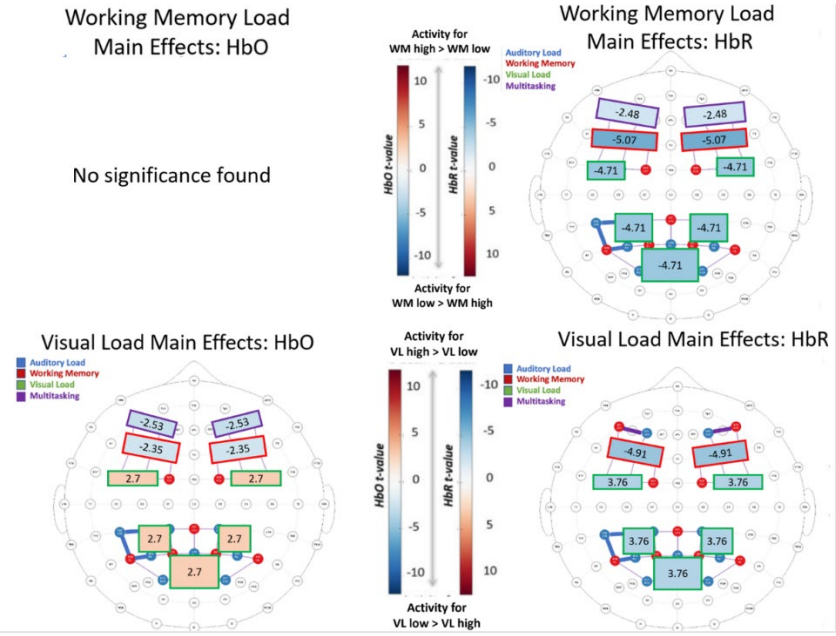


Figure 8: T-values from Table 3, overlayed over the fNIRS ROIs of multitasking, WM, VL, and AL. For HbO (left side) the red spectrum indicates increased activation, with darker red indicating more increased activation. For HbR (right) blue suggests more activation at that region, with darker blue indicating higher levels of activation.

As shown in Table 3, greater oxygenation is indicated, in the left half of the table, by more positive values for HbO and, in the right half, by more negative values for HbR. Within each half, the effect size of the manipulation is depicted in two different, but highly correlated measures, ß (upper half of each half) describes the size of the difference between low and high workload by the corresponding difference in ß, as derived from the toolbox. T (lower half of each half) describes the statistical significance of this effect as assessed by t-tests, also provided by the toolbox. These effects can be considered as equivalent to main effects of each of the two manipulations of workload (WM and VL). Within each of these 4 sub-tables are the critical effects of manipulating the two kinds of mental load (the two rows) on the activation within the four ROIs (four columns).

Viewing Table 3, we find support in the affirmative for *RQ1* and *RQ2*. fNIRS is indeed sensitive to WM levels (RQ1) as well as to VL levels (*RQ2*) in complex task environments. The boldfaced values (signaling significance) are high, and frequently occurring. Specifically, we note:

1. Overall, HbR (right side) appears to be more sensitive than HbO (left side) in that the values are both generally higher and more likely to be significant for HbR and show a consistent sign (negative) for the direction of the effect (increase workload) across all 8 cells, whereas HbO does not show such consistency, addressing *RQ4*.
2. Examining the pattern of effects **within** HbR in particular (right side), the effect of WM is considerably more powerful than is the effect of VL, whether assessed by T- or by ß-values. The difference in power between the two manipulations is consistent with the difference in effect size for the more traditional measures of workload reported above by secondary task performance, subjective workload, and pupil diameter. **It is noteworthy** that this disproportionately higher influence of WM on HbR was particularly evident in the multi-tasking (MT) ROI where both the T-value and the ß-value value are approximately twice as big for the manipulation of WM as for that of VL. This aligns with the findings of (Tomasi et al., 2007; Wallis et al., 2015) who found that WM is closely tied to the frontoparietal attention network, which encompasses regions in both of our WM and MT ROIs.
3. The WM ROI (column 1) was generally more sensitive to workload manipulations than was the VL ROI and the MT ROI. But the VL ROI is nevertheless somewhat sensitive to both manipulations. The same cannot be said for the MT ROI, which, within HbR has a relatively low and non-significant sensitivity to VL.

We see little support for *RQ3*, that fNIRS data viewed through our four ROIs as shown in Table 3 and Figure 8, is not diagnostic to the type of load, but the results are not straightforward. More specifically, viewing Table 3, the following conclusions emerge with respect to **diagnosticity**:

4. Regarding diagnosticity, in HbR we do **not** see the sort of specificity that would have been reflected in an interaction, whereby the VL ROI was more affected by VL than WM, and WM ROI was more affected by WM than VL. Instead, as described above, our manipulation of WM is consistently more powerful than that of VL, and the working

memory ROI is consistently more sensitive than the VL ROI, whether ß-value differences, or their significance in T-values is considered.

In an effort to quantify the differential diagnosticity of fNIRS in response to the two workload manipulations imposed, we asked: if a workload researcher were examining an fNIRS response, with no prior knowledge of the resources imposed by a manipulation, how accurately could she assess that the workload increase was imposed on one resource vs the other? That is, we make a **differential** diagnosis. To do this, in a quasi-Bayesian approach we approximate odds (a probability) by the strength of a signal, in T that represents the magnitude of a workload increase. In particular, we examine the ratios of T-values as follows:

[$WM^{ROI} \div VL^{ROI}$] *given that WM was increased*

to the

[$WM^{ROI} \div VL^{ROI}$] *given that VL was increased*

That is, the ratio of two odds ratios. When this ratio is 1.0, we argue that the hemodynamic values reflected by the two ROIs in question are undiagnostic. In calculating diagnosticity in this manner, we have chosen to use HbR, because examination of Table 3 reveals that it is the more sensitive measure of the two. Using the values in Table 3, we calculate that this diagnosticity ratio as:

$$\frac{5.07}{4.71} \div \frac{4.91}{3.76} = \frac{1.07}{1.30} = 0.82$$

This ratio, being close to 1.0 and certainly not substantially **greater** than 1.0, suggests that the global ROI measures, averaging as they do over several separate channels are not diagnostic of the source of workload. In interpreting this negative result, we note that Figure 8 reveals that several separate channels are involved in each ROI, and hence, a better reading of diagnosticity may come from examining the individual channel response as we do in the following section.

5.2.2 Sensitivity and Diagnosticity via Channel-Specific ROI Analyses

In the last section, we discussed the average-across channel ROI analysis, in which the responses of individual channels in our four pre-defined ROIs were averaged together for a 'big picture' analysis. From these results, we find our ROIs to indeed be quite sensitive to manipulated

workload, but they are *not* diagnostic to the type of load. To identify important differences within the individual channels, in this section we discuss the Channel-Specific ROI analysis in which individual channels were evaluated for statistical differences. Channel-wise statistics can be used to identify significant activation changes in more fine-grained functional ROIs than can be done with averaging across multiple channels. Because the channel-wise statistical comparisons have a larger chance of generating Type II errors, we use q-values rather than p-values to set our threshold of significance at .05. q-values are based on Benjamini-Hochberg false-discovery rate-corrected p-values (Benjamini & Hochberg, 1995). Each contrast in this section includes tables of the results that were significant ($q<0.05$) and were in the direction that corresponds to increased brain activation (positive values for HbO and negative values for HbR). We also include the corresponding LONI region and Brodmann Area that each channel covers. To evaluate the main effects, we ran contrast statements that corresponded to WM main effects (high WM – low WM) and VL main effects (high VL – low VL). Results are shown in Figure 9, with the full statistical results available in Appendix 2.

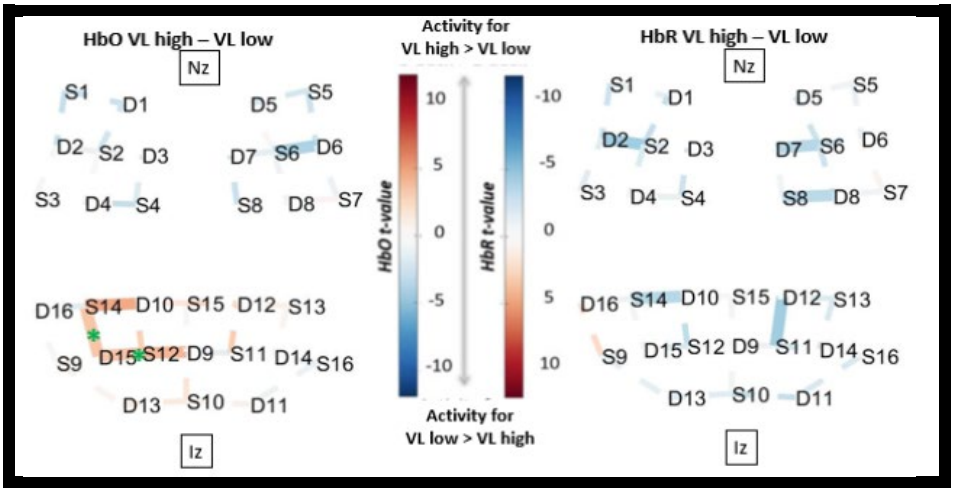


Figure 9: Working Memory Main Effects ($H^{WM} – L^{WM}$) and VL Main Effects $H^{VL} – L^{VL,}$ overlaid over a brain, with nasion (Nz) and inion (Iz) locations added for reference. Only significant channels ($q < 0.05$) are shown. For HbO, positive *t*-values (red) correspond to relatively larger activity for the first term in the contrast, and negative *t*-values (blue) correspond to larger activity for the second term. For HbR contrasts, negative *t*-values (blue) correspond to larger activation in that region. Green * indicates regions that are mutually exclusive (shown by one but not the other) between the WM and VL main effects tests.

In Figure 9, we note that there are six distinct channels (covering five distinct anatomical regions) that are uniquely activated (e.g., increased HbO or decreased HbR) for either the WM main effect (top of Fig. 9), or for the VL main effect (bottom of Fig. 9), but not for both. These regions are denoted with a green * in Figure 9, showing the location on the brain of each unique

region, and they are listed in Table 4. For the full set of statistical results that accompany Figure 9, please see Appendix 2.

Table 4: Summary of results from Table 3, with only unique significant activation shown.

Mutually exclusive regions activated for Working Memory main effects		**Mutually exclusive regions activated for VL main effects**	
Hb	Region	Hb	Region
HbO	L superior occipital gyrus	HbO	L angular gyrus (x2 channels)
HbR	L inferior frontal gyrus		
HbR	R precentral gyrus		
HbR	R angular gyrus		

To further evaluate the effects of load type on individual channels, we also wanted to directly contrast the fNIRS data when participants experienced just high WM, from times when they experienced just high VL. To do this we performed the following contrasts: ($WM^{high}VL^{low}$) - ($VL^{high}WM^{low}$) and its inverse ($VL^{high}WM^{low}$) - ($WM^{high}VL^{low}$). Results are shown in Figure 10, with full statistical output in Appendix 2.

Viewing the results shown in Figures 9, 10 and Table 3, we find strong support in the affirmative for *RQ1* and *RQ2*. fNIRS is indeed sensitive to WM levels (RQ1) as well as to VL levels (*RQ2*). Although these findings were already found in the four-region ROI analysis presented previously, the channel-wise results shown here, provide **critical support in the affirmative for RQ3, that our fNIRS channel-wise results are indeed diagnostic to load levels in our complex task**. More specifically, we note that:

From the data presented in Figure 9 and Table 4, we can conclude that:

1. Single-channel measures were diagnostic: one HbO channel (left superior occipital gyrus) and three HbR channels (left inferior frontal gyrus, right precentral gyrus, right angular gyrus) identified WM but not VL; activation of two HbO channels (both over left angular gyrus) signaled VL but not WM. These channels contribute to diagnosticity by differentiating markers of WM from markers of VL.
2. It is notable that the left IFG was uniquely activated for the WM main effect, but not for VL. The left IFG is part of the multitasking ROI that was used in our average-across

channels ROI analysis above (shown in our fNIRS montage in Figure 7). Working memory and multitasking share resources in the brain (A. Baddeley, 1996; A. Baddeley & Della Sala, 1996; Smith & Jonides, 1999), which is likely why we see this increase in the left IFG as more multitasking is needed to support task control when WM increases.

Looking at Figure 9, we see further support of diagnosticity between WM and VL. More specifically:

1. Looking at the differences in HbO and HbR in Figure 10 (top), our contrast WM^{high}-VL^{high} yields HbO deactivation in the frontal regions and significant activation in the parietal gyrus, and HbR shows activation of the left frontal gyrus and left occipital gyrus.
2. In Figure 10 (bottom), we see a similar trend between HbO and HbR in the inverse contrast (VL^{high}-WM^{low}) whereas HbO is significantly activated throughout the bilateral frontal gyrus and left supramarginal gyrus while HbR is activated in the right precentral gyrus. The differences across these two measures of hemoglobin show how they are inversely related to one another: when HbO shows activation in a region, HbR often shows deactivation in the same region and vice versa. VL^{high}-WM^{low} shows greater HbO activation in the bilateral frontal regions and HbR activation in the right precentral gyrus.
3. The above two findings suggest that WM tasks induce activation of the parietal and left frontal regions, while VL tasks also activate the frontal and precentral regions. We again conclude that there is overlap between the WM and VL activation regions, especially in the frontal gyrus, where much of both memory and visual processing occur.

4.

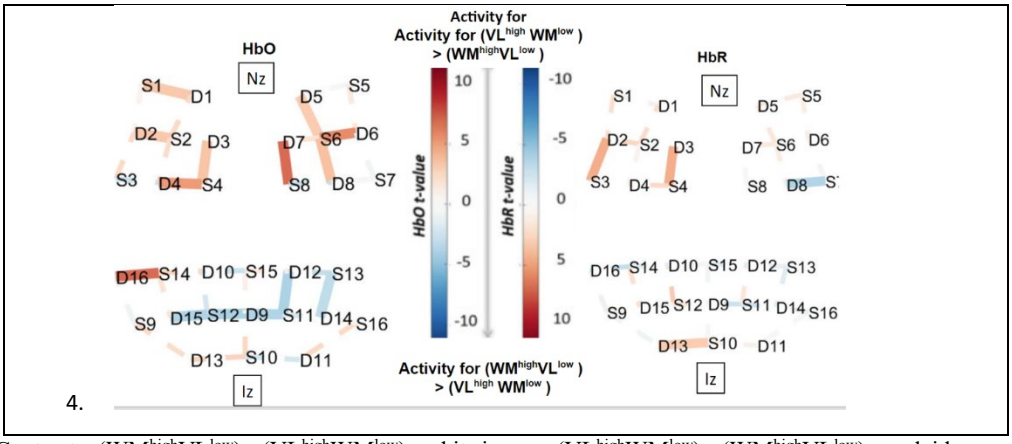


Figure 10: Contrasts: ($WM^{high}VL^{low}$) - ($VL^{high}WM^{low}$) and its inverse ($VL^{high}WM^{low}$) - ($WM^{high}VL^{low}$) overlaid over a brain, with nasion (Nz) and inion (Iz) locations added for reference. Only significant channels ($q < 0.05$) are shown. For HbO, positive *t*-values (red) correspond to relatively larger activity for the first term in the contrast, and negative *t*-values (blue) correspond to larger activity for the second term. For HbR contrasts, negative *t*-values (blue) correspond to larger activation in that region.

6 Discussion

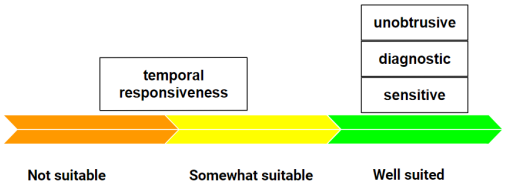
The main goal of this paper was to evaluate the utility of fNIRS for workload-based adaptive automation through the lens of the four principles of unobtrusiveness, temporal responsiveness, sensitivity, and diagnosticity criteria. Table 5 summarizes our findings, which we discuss here.

6.1 Meta-Review Goals and Findings

We explored the four criteria via a **meta-review of related workload-focused fNIRS literature.** With respect to *unobtrusiveness***,** we found a combination of research literature and commercial bio-technology developments that suggest that fNIRS is a device well- suited for being unobtrusive in AA. We expect future fNIRS devices to be designed for specific-use cases, where the number and layout of channels, comfort of the probes, quality of the signal vs. cost, are all considered for a specific use case. We summarize our findings in Table 5.

With respect to *temporal responsiveness,* it is generally agreed upon that the underlying hemodynamic response is quite slow by nature throughout fNIRS literature, as most studies in our meta-review ran analyses on fNIRS data with tasks lasting more than 25 seconds (see Appendix 1). Thus, fNIRS is not ideally suited for AA because of the resulting lag induced in a closed loop adaptive system, and inherent instability when that lag approaches the time constant of workload changes within the task (C. D. Wickens et al., 2022).

Table 5: Summary of findings collated from our experiment and the meta-review. Top: a graphic depicting our findings, with mappings on the suitability of fNIRS for workload-based AA based on the four criteria of temporal responsiveness, unobtrusiveness, diagnosticity, and sensitivity. Bottom: Text summary of our findings regarding the four criteria.



Criteria	Summary of Meta-Review and Experimental Findings
Unobtrusiveness	**Meta-review** findings show a strong trend toward devices continuing to be more wearable, practical, and specialized to specific use cases. **Empirical** results were achieved in this study using a NIRSport 2. The wireless NIRSport2 was equipped with probe tips specially designed for comfort on the scalp.
Temporal responsiveness	**Meta-review** findings suggest that like fMRI, fNIRS on its own, measures a slowly moving hemodynamic response, which makes its temporal responsiveness relatively slow. More work is needed, following the lead of researchers who have focused on exploring short sliding windows of time in ML classification and on hybrid EEG/fNIRS adaptive systems. **Empirical** results were generated using statistical tests on task lengths of 45 seconds in duration, which does not further our understanding of the temporal responsiveness of the fNIRS signal.

Sensitivity	**Meta-review** showcased strong results of the sensitivity to workload manipulations; however with the majority of work focused on extremely simple, highly controlled benchmark tasks (e.g. n-back tasks), rather than those tasks typical of an extra-laboratory working environment. **Empirical** results indicate that fNIRS is sensitive to changes in visual) and working memory load levels. HbR appeared to be more sensitive than HbO for WM, while both HbR and HbO appeared to be sensitive to VL manipulations (as shown in Tables 3 and 4).
Diagnosticity	**Meta-Review** found very little prior work on diagnosticity. Of that handful of work, the vast majority has been done on simple, highly controlled tasks. **Empirical** results indicate that fNIRS is diagnostic to type of load, specifically to visual vs WM, but these findings are not clear cut. In the channel-wise analysis, we found unique regions that are activated in the WM main effects comparison that were not activated by the VL main effects, and vice versa. Such diagnosticity was not revealed by the ROI analysis where we condensed the data into four ROIs. When the data was kept in its channel-wise form, we did see diagnosticity: For WM, we see one HbO channel and three HbR channels that are unique for differentiating WM (HbO: Left superior occipital gyrus, HbR: L inferior frontal gyrus, R precentral gyrus, R angular gyrus). For diagnosticity for VL, we see two HbO channels uniquely differentiating VL, both measure the L angular gyrus.

To combat this issue, much recent research has focused on exploring shorter time windows for machine learning classification, with a sliding window approach being particularly well suited for the fNIRS signal (see R. Liu et al. (2021) for a thorough review of time windows used to date). While the fNIRS signal is bound by the nature of the hemodynamic response to be relatively sluggish, a number of studies have taken a multimodal approach, merging fNIRS data with other behavioral and physiological measurements that have higher temporal responsiveness. Some such approaches utilize hybrid fNIRS/EEG systems for future AA systems that combine the spatial resolution of fNIRS with the high temporal responsiveness of the EEG signal (Kwon et al., 2020; Putze et al., 2014).

In terms of *sensitivity* and *diagnosticity* of the fNIRS signal for workload measurements, the findings in our meta-review were less conclusive. Research conducted to date suggests a complex interplay between task difficulty, practice effects, and human performance, which all have an unpredictable effect on fNIRS measures of HbO and HbR taken from the outer cortex of the brain (Herff et al., 2014; R. McKendrick, Ayaz, Olmstead, & Parasuraman, 2013; Kimberly L Meidenbauer et al., 2021). More work is needed in this area to better understand the relationship between these factors on sensitivity and diagnosticity of fNIRS signals (Herff et al., 2014).

6.2 Experiment Findings and Interpretations

As outlined below, our results lent clear, affirmative support for RQ1, RQ2 and RQ4. With respect to RQ3, and diagnosticity, only our channel-wise analysis lent support for our ability to differentiate changes in WM from changes in VL. Importantly, our findings align with prior neuroscience work on visual attention, WM, executive processing, and multitasking (A. Baddeley, 1996; A. Baddeley & Della Sala, 1996; Smith & Jonides, 1999; Tomasi et al., 2007). Regarding the primary purpose of validating the use of fNIRS in AA, our findings contribute to the studies that have also done this in realistic tasks (e.g., Hasan Ayaz et al. (2012); Gateau et al. (2018)). Our conventional measures of primary and secondary task performance, subjective measures (NASA-TLX) and neuroergonomic measures (pupil diameter), as shown in Figures 2-4, validated that the two workload manipulations were effective and revealed that our manipulation of WM was considerably more powerful than that of VL. Furthermore, our fNIRS results converged on and replicated these two trends.

We ran our analyses using two different techniques for ROI mappings: Average-Across Channels and Channel-Specific. Both techniques yielded **clear and convergent evidence of sensitivity of load level.** Our fNIRS measure of global activation at each of our four ROIs (WM, VL, AL, and MT) were shown in Table 3, where we note that most of the T and ß were in the "expected" direction, signaling greater oxygenation (higher HbO, and, particularly, lower HbR and thus more **sensitivity**) with increased workload. They also showed greater effects for the manipulation of working memory than of visual workload. This differential effect of manipulation power on sensitivity was also reflected in the multitasking ROI. Thus, the fNIRS data provided a **differentially** sensitive measure of workload. It also became clear that these effects were more strongly reflected in HbR, than in the more frequently used HbO.

Although the fNIRS analyses performed by averaging data into four ROIs did not show support for diagnosticity, the fNIRS results completed at the channel-level did yield clear findings to support diagnosticity. These results found with respect to **diagnosticity unique to WM**, both HbO and HbR make contributions toward **diagnosticity**. For the WM and VL main effects comparisons (Fig 9, Table 3), we note that one HbO channel and three HbR channels were unique for differentiating WM (HbO: Left superior occipital gyrus, HbR: L inferior frontal gyrus, R precentral gyrus, R angular gyrus).

These findings dovetail with literature on the WM network in complex tasks, that it engages an interconnected network of brain regions that work together to maintain and update working memory while shifting attention and executing tasks based on the changing task environment. Notably, the Left inferior frontal gyrus (IFG) plays a key role in the executive function behind working memory (Nee et al., 2013). This aligns with the prior work by Gateau et al. (2018) in which they found the left PFC (their ROI #2) experienced greater HbO changes during the real flight condition than in the simulator condition when carrying out high WM tasks. Those authors note that multitasking became critical when carrying out a high WM task rather than a low WM task while also navigating a real aircraft.

With respect to **diagnosticity unique to the VL main effect**, only HbO contributes toward **diagnosticity** (as shown in Table 3, and by the green * in Figure 9). For unique diagnosticity for VL, there were two HbO channels uniquely differentiating VL, both of which cover the L angular gyrus. The angular gyrus has long been recognized as a key region involved in visual attention and visuospatial processing (Göbel, Walsh, & Rushworth, 2001; Studer, Cen, & Walsh, 2014), and indeed seems to be an important region distinguishing VL manipulations (e.g., making our target shapes more/less like the surrounding distractors) from the WM manipulations. Support for diagnosticity was further found by contrasting the two conditions of $WM^{high}VL^{low}$ with $Vl^{high}WM^{low}$. The contrast results (Appendix 2, Table B) suggest that WM tasks induce activation of the parietal and left frontal regions, while VL tasks also activate the frontal and precentral regions. We again conclude that there is overlap between the WM and VL activation regions, especially in the frontal gyrus, where much of memory and visual processing occur.

In summary, the unique brain regions identified as unique to WM and unique to VL **dovetail with research from the fMRI domain** about the brain regions involved in VL versus WM, with WM activating a more diverse interconnected network of brain regions that spans from the prefrontal cortex, back through the parietal region, and into the occipital cortex. VL, on the other hand, only engaged unique brain regions in the left angular gyrus, which has been repeatedly linked to visuospatial attention. These results align with prior literature in the neuroscience domain; suggesting that the brain resources engaged in visual attention and WM are highly overlapping, but not identical (Tomasi et al., 2007). For example, one study that evaluated verbal WM and spatial attention (SA) tasks using fMRI found a common activation network made of

the frontal, temporal, and parietal cortices, suggesting that tasks share a common dynamic shifting of attentional resources in these common areas (LaBar, Gitelman, Parrish, & Mesulam, 1999). More aligned with our experiment, Tomasi et al. (2007) employed a similar paradigm to evaluate the effects of high and low WM and visual attention (VA) tasks. They found that despite the differential attentional requirements of the tasks, they both activated a common network including the prefrontal, parietal, and occipital cortices.

Viewing our findings **through the lens of resource theory,** we expected to see increased load placed on brain regions responsible for multitasking, as a result of coordination of our complex task, with increased WM in particular placing greater demands on MT regions than VL increases. Our observed pattern of effects on the MT ROI are readily interpretable. Multitasking is heavily supported by executive control (C. D. Wickens et al., 2022). So is working memory which, at higher load, involves more mental juggling of the subtasks of maintenance and processing (A. D. Baddeley & Hitch, 1974; Engle, 2002). In contrast, VL increases impose primarily input-output processing, not imposing greater multi-tasking requirements. To our knowledge, none of the studies reviewed above, nor the larger set contained in Appendix 1, examined this specific multi-tasking ROI as we do here.

6.3 Implications for Workload Based Adaptive System Designs

While our work revealed, as other's also have, that fNIRS is quite sensitive to variations in visual and working memory load (particularly the latter), its feasibility for adaptive systems remains constrained by the lag in its measurement, as seen in the current experiment, and replicating many earlier studies. This lag is bound by the nature of the hemodynamic response (Figure 1) with it taking roughly 8 seconds after a stimulus onset for HbO and HbR to peak (Huppert et al., 2006). This naturally occurring 8-second lag clearly places a lower bound that makes it challenging to classify load levels on fNIRS time windows. As noted previously (section 3.2 on temporal responsiveness), most research has found that ~25-second continuous tasks, paired with ~25-second-long window lengths, has yielded most success to date in single trial classification of fNIRS workload levels.

A lag value of this magnitude does not preclude fNIRS use in adaptive automation. However, it will only reliably reflect changes in workload in circumstances when the workload changes that AA is designed to compensate for, are themselves relatively gradual, such as the increased workload imposed by fading illumination at dusk, or that associated with cumulative

mental fatigue. In this regard, statistically significant classification rates between low and high workload conditions in a shorter period of time are not sufficient to justify incorporation in adaptive automation, simply because even a small loss in classification accuracy (associated say with a 90% classification rate) will be likely to undermine user trust in the system. A promising option is to couple fNIRS with other unobtrusive workload assessment techniques, such as EEG alpha/theta ratio, or pupil diameter that may have a much faster response rate, even if those are less sensitive, and less diagnostic; in short, a team approach to on-line workload assessment.

Given the meta-review and empirical findings summarized in Table 5, it is clear that fNIRS developers of workload-based AA systems should consider using multiple measurement modalities (e.g., hybrid EEG/fNIRS) to improve temporal responsiveness, and inclusion of **both** HbO and HbR in the measurement and modeling approaches will provide complementary measurements toward load sensitivity and diagnosticity.

7 Study Limitations

Group-Statistics vs Single-Trial Classification. In this study we have evaluated the utility of fNIRS for diagnosticity, sensitivity, and temporal responsiveness at the group statistical level, we have not run machine learning analyses per individual. Before turning to machine learning and single trial analyses, we opted to take the important step of first using group-level statistics to establish the statistical reliability of the time-varying response and the reliability of the fNIRS signal to distinguish between the different forms of load. Thus, our findings can only be interpreted at the group level; and our ability to extrapolate findings to the individual-level, where adaptive systems would operate, is limited. Yet determining the statistical differences at this group level is essential to extending the technique to adaptive automation, where machine learning can classify the differences in inferred workload based on individual responses. Future work should extend these findings to investigate load diagnosticity, sensitivity, and temporal responsiveness at the individual level, and we hope that our use of cognitive load theory and development of a complex load manipulation testbed provides a pathway to extend this work toward individual level measurement and modeling.

fNIRS Preprocessing. Of major concern within the fNIRS signal is the presence of serially correlated errors due to high sampling rate and heavy-tailed noise distributions (Kimberly L Meidenbauer et al., 2021; H. Santosa et al., 2018) due to noise in the signal. We did not use

short-channel regression in our preprocessing pipeline. This is a limitation of this work as short-channel regression is the optimal technique of distinguishing task-evoked non-neuronal response (systemic noise) from the neuronal signal of interest (Tachtsidis & Scholkmann, 2016; M. Yücel et al., 2021). Our pipeline utilized both accelerometer regression and AR-IRLS pre-whitening, which have been applied widely throughout fNIRS literature to reduce motion and physiological-related noise. The AR-IRLS model uses an auto-regressive filter to minimize these errors and iteratively down-weighs outliers in a weighted regression, and has been adapted for real-time filtering (J. Barker, Rosso, Sparto, & Huppert, 2016; J. W. Barker et al., 2013). However, it is best practice to utilize short-separation channel regression to obtain the true hemodynamic response signal. We encourage future work to use this technique, especially within the context of AA.

8 Conclusions and Future Work

Although the concept of workload-based adaptive automation has been discussed frequently in the fields of HCI and human factors, these intelligent systems have proven very difficult to achieve. In this paper we focused on the utility of fNIRS for addressing four measurement criteria that are essential to consider if we are to realize the vision of workload-based AA with fNIRS and described a meta-review and empirical study to explore these criteria. We found that fNIRS has relatively poor **temporal responsiveness**, but it rates highly with respect to **unobtrusiveness**. Further, the fNIRS signal is adequately sensitive to gradations of load level changes (**sensitivity**), and when data are viewed in channel-wise format, the fNIRS device does appear to offer **diagnosticity**; whereby the type of load being modulated (in our case WM and VL) can be uniquely identified. Although our findings showed support for sensitivity and diagnosticity of the fNIRS signal, we note the strong need for more research to be conducted by fNIRS researchers in the HF and HCI domains, if we are to build workload-based AA using the fNIRS signal. Future research should focus on diagnosticity and sensitivity of fNIRS for measuring workload changes in studies that utilize complex, ecologically valid tasks, with a suitable number of channels for differentiation of different types of workload. Also, adaptive systems are composed of data-hungry algorithms, which cannot be adequately trained on small datasets, especially given the high dimensional features space of brain data. Even if the fNIRS signal is suitable for differentiating between different types of load and different levels of load,

there is a great need for the research community to make dataset and testbed sharing a priority. Shared fNIRS data should contain information about the anatomical brain region measured by each channel, as well as access to raw data streams, so that researchers can train models on datasets aggregating different experiments, different devices, and different labs. These large-scale efforts will be instrumental in order to fully realize the goals of using fNIRS as a basis for workload-based AA.

ACKNOWLEDGEMENTS

We thank the National Science Foundation for support of this research (Award # 1909864). We'd also like to thank Michal Bodzianowski for his phenomenal support creating and maintaining the search and sort testbed used in our studies.

REFERENCES

Abdullah, A., & Khan, I. H. (2018). *Application of Near-Infra-Red Spectroscopy for the Analysis of the Impact of Black Color on Neural Response.* Paper presented at the Proceedings of the 2nd International Conference on Information System and Data Mining.

Afergan, D., Peck, E. M., Solovey, E. T., Jenkins, A., Hincks, S. W., Brown, E. T., . . . Jacob, R. J. K. (2014). *Dynamic difficulty using brain metrics of workload*. Paper presented at the Proceedings of the 32nd annual ACM conference on Human factors in computing systems, Toronto, Ontario, Canada.

Aghajani, H., Garbey, M., & Omurtag, A. (2017). Measuring mental workload with EEG+ fNIRS. *Frontiers in human neuroscience, 11*, 359.

Aihara, T., Shimokawa, T., Ogawa, T., Okada, Y., Ishikawa, A., Inoue, Y., & Yamashita, O. (2020). Resting-state functional connectivity estimated with hierarchical bayesian diffuse optical tomography. *Frontiers in neuroscience, 14*, 32.

Aksoy, E., Izzetoglu, K., Baysoy, E., Agrali, A., Kitapcioglu, D., & Onaral, B. (2019). Performance monitoring via functional near infrared spectroscopy for virtual reality based basic life support training. *Frontiers in Neuroscience, 13*, 1336.

Al-Hudhud, G., Alqahtani, L., Albaity, H., Alsaeed, D., & Al-Turaiki, I. (2019). Analyzing Passive BCI Signals to Control Adaptive Automation Devices. *Sensors (Basel, Switzerland), 19*(14), 3042. doi:10.3390/s19143042

Anderson, A. A., Smith, E., Chowdhry, F. A., Thurm, A., Condy, E., Swineford, L., . . . Gandjbakhche, A. H. (2017). Prefrontal hemodynamics in toddlers at rest: A pilot study of developmental variability. *Frontiers in neuroscience, 11*, 300.

Aranyi, G., Charles, F., & Cavazza, M. (2015). *Anger-based BCI using fNIRS neurofeedback.* Paper presented at the Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology.

Asgher, U., Ahmad, R., Naseer, N., Ayaz, Y., Khan, M. J., & Amjad, M. K. (2019). Assessment and classification of mental workload in the prefrontal cortex (PFC) using fixed-value modified beer-lambert law. *IEEE Access, 7*, 143250-143262.

Asgher, U., Khalil, K., Khan, M. J., Ahmad, R., Butt, S. I., Ayaz, Y., . . . Nazir, S. (2020). Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface. *Frontiers in neuroscience, 14*, 584.

Ayaz, H., Baker, W. B., Blaney, G., Boas, D. A., Bortfeld, H., Brady, K., . . . Carp, S. A. (2022). Optical imaging and spectroscopy for the study of the human brain: status report. *Neurophotonics, 9*(S2), S24001.

Ayaz, H., & Dehais, F. (2018). *Neuroergonomics The Brain at Work and in Everyday Life*: Academic Press.

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

Backs, R. W., Lenneman, J. K., Wetzel, J. M., & Green, P. (2003). Cardiac measures of driver workload during simulated driving with and without visual occlusion. *Human Factors, 45*(4), 525-538.

Baddeley, A. (1996). Exploring the Central Executive. *The Quarterly Journal of Experimental Psychology, 49*.

Baddeley, A., & Della Sala, S. (1996). Working memory and executive control. *Philosophical Transactions of the Royal Society of London, 351*.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47-89): Elsevier.

Baker, J. M., Bruno, J. L., Gundran, A., Hosseini, S. H., & Reiss, A. L. (2018). fNIRS measurement of cortical activation and functional connectivity during a visuospatial working memory task. *PloS one, 13*(8), e0201486.

Barker, J., Rosso, A., Sparto, P., & Huppert, T. (2016). Correction of motion artifacts and serial correlations for real-time functional near-infrared spectroscopy. *Neurophotonics, 3*(3), 031410. doi:10.1117/1.NPh.3.3.031410

Barker, J. W., Aarabi, A., & Huppert, T. J. (2013). Autoregressive model based algorithm for correcting motion and serially correlated errors in fNIRS. *Biomedical optics express, 4*(8), 1366-1379.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300. doi:10.1016/S0166-4328(01)00297-2

Brouwer, A.-M., Zander, T. O., van Erp, J. B. F., Korteling, J. E., & Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Frontiers in Neuroscience, 9*(136). doi:10.3389/fnins.2015.00136

Bunce, S. C., Izzetoglu, K., Ayaz, H., Shewokis, P., Izzetoglu, M., Pourrezaei, K., & Onaral, B. (2011, 2011//). *Implementation of fNIRS for Monitoring Levels of Expertise and Mental Workload.* Paper presented at the Foundations of Augmented Cognition. Directing the Future of Adaptive Systems, Berlin, Heidelberg.

Cakir, M. P., Çakir, N. A., Ayaz, H., & Lee, F. J. (2015). *An optical brain imaging study on the improvements in mathematical fluency from game-based learning.* Paper presented at the Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play.

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., & Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific reports, 7*(1), 1-15.

Chance, B., Zhuang, Z., Chu, U., Alter, C., & Lipton, L. (1993). Cognition activated low frequency modulation of light absorption in human brain. *90*, 2660-2774

Charles, F., De Castro Martins, C., & Cavazza, M. (2020). Prefrontal asymmetry BCI neurofeedback datasets. *Frontiers in Neuroscience, 14*, 601402.

Chen, L.-C., Sandmann, P., Thorne, J., Herrmann, C., & Debener, S. (2015). *Association of Concurrent fNIRS and EEG Signatures in Response to Auditory and Visual Stimuli* (Vol. 28).

Chen, Y., Tang, J., Chen, Y., Farrand, J., Craft, M. A., Carlson, B. W., & Yuan, H. (2020). Amplitude of fNIRS resting-state global signal is related to EEG vigilance measures: A simultaneous fNIRS and EEG study. *Frontiers in Neuroscience*, 1265.

Chu, H., Cao, Y., Jiang, J., Yang, J., Huang, M., Li, Q., . . . Jiao, X. (2022). Optimized electroencephalogram and functional near-infrared spectroscopy-based mental workload detection method for practical applications. *BioMedical Engineering OnLine, 21*(1), 1-17.

Crottaz-Herbette, S., Anagnoson, R. T., & Menon, V. (2004). Modality effects in verbal working memory: differential prefrontal and parietal responses to auditory and visual stimuli. *NeuroImage, 21*(1), 340-351.

Cui, X., Bray, S., Bryant, D. M., Glover, G. H., & Reiss, A. L. (2011). A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. *Neuroimage, 54*(4), 2808-2821.

Dorneich, M. C., Dudley, R., Rogers, W., Letsu-Dake, E., Whitlow, S. D., Dillard, M., & Nelson, E. (2015). Evaluation of information quality and automation visibility in Information automation on the flight deck. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59*(1), 284-288. doi:10.1177/1541931215591058

Dravida, S., Noah, J. A., Zhang, X., & Hirsch, J. (2017). Comparison of oxyhemoglobin and deoxyhemoglobin signal reliability with and without global mean removal for digit manipulation motor tasks. *Neurophotonics, 5*(1), 011006.

Duan, Y., Liu, X., & Lian, Y. (2021). *Progress on Hybrid EEG-fNIRs system and its application*. Paper presented at the Proceedings of the 2nd International Symposium on Artificial Intelligence for Medicine Sciences, Beijing, China.

Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., . . . Giannopoulos, I. (2018). *The Index of Pupillary Activity: Measuring Cognitive Load vis-à-vis Task Difficulty with Pupil Oscillation*. Paper presented at the Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal QC, Canada.

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

Engle, R. W. (2002). Working memory capacity as executive attention. *Current directions in psychological science, 11*(1), 19-23.

Fairclough, S., Ewing, K., Burns, C., & Kreplin, U. (2019). Chapter 12 - Neural Efficiency and Mental Workload: Locating the Red Line. In H. Ayaz & F. Dehais (Eds.), *Neuroergonomics* (pp. 73-77): Academic Press.

Fishburn, F. A., Norr, M. E., Medvedev, A. V., & Vaidya, C. J. (2014). Sensitivity of fNIRS to cognitive state and load. *Frontiers in human neuroscience, 8*, 76.

Foy, H. J., Runham, P., & Chapman, P. (2016). Prefrontal cortex activation and young driver behaviour: a fNIRS study. *PLoS one, 11*(5), e0156512.

Friedman, L., Walker, E., & Solovey, E. (2018). *Integrating non-invasive neuroimaging and computer log data to improve understanding of cognitive processes*. Paper presented at the Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data, Boulder, Colorado.

Gateau, T., Ayaz, H., & Dehais, F. (2018). In silico vs. over the clouds: on-the-fly mental state estimation of aircraft pilots, using a functional near infrared spectroscopy based passive-BCI. *Frontiers in human neuroscience, 12*, 187.

Geng, S., Liu, X., Biswal, B. B., & Niu, H. (2017). Effect of resting-state fNIRS scanning duration on functional brain connectivity and graph theory metrics of brain network. *Frontiers in neuroscience, 11*, 392.

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

Girouard, A., Solovey, E., Hirshfield, L. M., Chauncey, K., Sassaroli, A., Fantini, S., & Jacob, R. (2009). *Distinguishing Difficulty Levels with Non-invasive Brain Activity Measurements.* Paper presented at the Proc. INTERACT Conference.

Göbel, S., Walsh, V., & Rushworth, M. F. (2001). The mental number line and the human angular gyrus. *Neuroimage, 14*(6), 1278-1289.

Goshvarpour, A., & Goshvarpour, A. (2023). Matching pursuit-based analysis of fNIRS in combination with cascade PCA and reliefF for mental task recognition. *Expert Systems with Applications, 213*, 119283.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theorical research. In P. Hancock, Meshkati, N. (Ed.), *Human Mental Workload* (pp. pp 139 - 183). Amsterdam.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. *Frontiers in human neuroscience, 7*, 935.

Hirshfield, L., Gulotta, R., Hirshfield, S., Hincks, S., Russell, M., Williams, T., & Jacob, R. (2011). *This is your brain on interfaces: enhancing usability testing with functional near infrared spectroscopy.* Paper presented at the SIGCHI.

Hirshfield, L. M., Chauncey, K., Gulotta, R., Girouard, A., Solovey, E. T., Jacob, R. J. K., . . . Fantini, S. (2009). *Combining Electroencephalograph and Near Infrared Spectroscopy to Explore Users' Mental Workload States* Paper presented at the HCI International.

Holmes, C., Hoge, R., Collins, L., Woods, R., Toga, A., & Evans, A. (1998). Enhancement of MR images using registration for signal averaging. *Journal of Computer Assist Tomography, 22*(2), 324–333.

Holmes, E., Barrett, D. W., Saucedo, C. L., O'Connor, P., Liu, H., & Gonzalez-Lima, F. (2019). Cognitive enhancement by transcranial photobiomodulation is associated with cerebrovascular oxygenation of the prefrontal cortex. *Frontiers in Neuroscience, 13*, 1129.

Huppert, T. J., Franceschini, M. A., & Boas, D. A. (2009). Noninvasive imaging of cerebral activation with diffuse optical tomography. *In Vivo Optical Imaging of Brain Function. 2nd edition*.

Huppert, T. J., Hoge, R. D., Diamond, S. G., Franceschini, M. A., & Boas, D. A. (2006). A temporal comparison of BOLD, ASL, and NIRS hemodynamic responses to motor stimuli in adult humans. *Neuroimage, 29*(2), 368-382.

İşbilir, E., Çakır, M. P., Acartürk, C., & Tekerek, A. Ş. (2019). Towards a multimodal model of cognitive workload through synchronous optical brain imaging and eye tracking measures. *Frontiers in human neuroscience, 13*, 375.

Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., & Chance, B. (2004). Functional Optical Brain Imaging Using Near-Infrared During Cognitive Tasks. *International Journal of Human-Computer Interaction, 17*(2), 211-231.

Jacobs, K. M. (2011). Brodmann's areas of the cortex. *Encyclopedia of clinical neuropsychology*, 459-459.

Jacques, S. L. (2013). Optical properties of biological tissues: a review. *Physics in Medicine & Biology, 58*(11), R37.

Jin, L., Jia, H., & Yu, D. (2018). *A Novel Indicator for Assessing Conceptual Change: A Study Based on Functional Connectivity Measured with Near-infrared Spectroscopy.* Paper presented at the Proceedings of the 2018 International Conference on Education Technology Management.

Kaber, D. B., & Kim, S.-H. (2011). Understanding Cognitive Strategy With Adaptive Automation in Dual-Task Performance Using Computational Cognitive Models. *Journal of Cognitive Engineering and Decision Making, 5*(3), 309-331. doi:10.1177/1555343411416442

Kerr, J., Reddy, P., Shewokis, P. A., & Izzetoglu, K. (2022). Cognitive Workload Impacts of Simulated Visibility Changes During Search and Surveillance Tasks Quantified by Functional Near Infrared Spectroscopy. *IEEE Transactions on Human-Machine Systems*.

Keshmiri, S., Sumioka, H., Okubo, M., & Ishiguro, H. (2019). An information-theoretic approach to quantitative analysis of the correspondence between skin blood flow and functional near-infrared spectroscopy measurement in prefrontal cortex activity. *Frontiers in neuroscience, 13*, 79.

Khan, M. J., Liu, X., Bhutta, M. R., & Hong, K.-S. (2016). *Drowsiness detection using fNIRS in different time windows for a passive BCI.* Paper presented at the 2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob).

Kuruvilla, M. S., Green, J. R., Ayaz, H., & Murman, D. L. (2013). Neural correlates of cognitive decline in ALS: An fNIRS study of the prefrontal cortex. *Cognitive neuroscience, 4*(2), 115-121.

Kwon, J., Shin, S., & Im, C. (2020). Toward a compact hybrid brain-computer interface (BCI): Performance evaluation of multi-class hybrid EEG-fNIRS BCIs with limited number of channels. . *PLoS ONE, 15*(3).

LaBar, K. S., Gitelman, D. R., Parrish, T. B., & Mesulam, M.-M. (1999). Neuroanatomic overlap of working memory and spatial attention networks: a functional MRI comparison within subjects. *Neuroimage, 10*(6), 695-704.

Lai, C. Y., Ho, C. S., Lim, C. R., & Ho, R. C. (2017). Functional near-infrared spectroscopy in psychiatry. *BJPsych Advances, 23*(5), 324-330.

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychological Human Perception and Performance 21*, 451-468.

Lavie, N., Hirst, A., de Fockert, J. W., & Viding, E. (2004). Load theory of selective attention and cognitive control. *Journal of Expirmental Psychology, 133*, 339-354.

Le, A. S., Xuan, N. H., & Aoki, H. (2022). Assessment of senior drivers' internal state in the event of simulated unexpected vehicle motion based on near-infrared spectroscopy. *Traffic injury prevention*, 1-5.

Lee, G., Jung, Y.-J., Park, J.-s., & Hong, J.-Y. (2019). *ActiView: A MATLAB-based Toolbox for Realtime Cortical Activation Analysis Using Functional Near-infrared Spectroscopy.* Paper presented at the Proceedings of the 2019 2nd International Conference on Electronics and Electrical Engineering Technology.

Lei, M., Miyoshi, T., Dan, I., & Sato, H. (2020). Using a Data-Driven Approach to Estimate Second-Language Proficiency From Brain Activation: A Functional Near-Infrared Spectroscopy Study. *Frontiers in Neuroscience, 14*. doi:10.3389/fnins.2020.00694

Li, W., Li, R., Xie, X., & Chang, Y. (2022). Evaluating mental workload during multitasking in simulated flight. *Brain and Behavior, 12*(4), e2489.

Liu, R., Walker, E., Friedman, L., Arrington, C. M., & Solovey, E. T. (2021). fNIRS-based classification of mind-wandering with personalized window selection for multimodal learning interfaces. *Journal on Multimodal User Interfaces, 15*(3), 257-272.

Liu, Y., & Ayaz, H. (2018). Speech recognition via fNIRS based brain signals. *Frontiers in neuroscience, 12*, 695.

Logothetis, N. K., Pauls, J., Augath, M., Trinath, T., & Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. *nature, 412*(6843), 150-157.

Lohani, M., Payne, B. R., & Strayer, D. L. (2019). A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving. *Frontiers in human neuroscience, 13*, 57-57. doi:10.3389/fnhum.2019.00057

MacIntosh, B. J., Klassen, L. M., & Menon, R. S. (2003). Transient hemodynamics during a breath hold challenge in a two part functional imaging study with simultaneous near-infrared spectroscopy in adult humans. *NeuroImage, 20*(2), 1246-1252.

MacNeil, E., Bishop, A., & Izzetoglu, K. (2022). *Study of Different Classifiers and Multi-modal Sensors in Assessment of Workload.* Paper presented at the International Conference on Human-Computer Interaction.

Maior, H. A., Wilson, M. L., & Sharples, S. (2018). Workload alerts—using physiological measures of mental workload to provide feedback during tasks. *ACM Transactions on Computer-Human Interaction, 25*(2).

Malonek, D., & Grinvald, A. (1996). Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: implications for functional brain mapping. *Science, 272*(5261), 551-554.

Mandrick, K., Chua, Z., Causse, M., Perrey, S., & Dehais, F. (2016). Why a Comprehensive Understanding of Mental Workload through the Measurement of Neurovascular Coupling Is a Key Issue for Neuroergonomics? *Frontiers in Human Neuroscience, 10*(250). doi:10.3389/fnhum.2016.00250

Mandrick, K., Derosiere, G., Dray, G., Coulon, D., Micallef, J.-P., & Perrey, S. (2013). Prefrontal cortex activity during motor tasks with additional mental load requiring attentional demand: a near-infrared spectroscopy study. *Neuroscience research, 76*(3), 156-162.

Mansouri, F. A., Koechlin, E., Rosa, M. G. P., & Buckley, M. J. (2017). Managing competing goals — a key role for the frontopolar cortex. *Nature Reviews Neuroscience, 18*(11), 645-657. doi:10.1038/nrn.2017.111

McKendrick, R., Ayaz, H., Olmstead, R., & Parasuraman, R. (2013). Enhancing Dual-Task Performance with Verbal and Spatial Working Memory Training: Continuous Monitoring of Cerebral Hemodynamics with NIRS. . *Neuroimage*.

McKendrick, R., Parasuraman, R., Murtza, R., Formwalt, A., Baccus, W., Paczynski, M., & Ayaz, H. (2016). Into the Wild: Neuroergonomic Differentiation of Hand-Held and Augmented Reality Wearable Displays during Outdoor Navigation with Functional Near Infrared Spectroscopy. *Frontiers in Human Neuroscience, 10*(216). doi:10.3389/fnhum.2016.00216

Meidenbauer, K. L., Choe, K. W., Cardenas-Iniguez, C., Huppert, T. J., & Berman, M. G. (2020). Load-Dependent Relationships between Frontal fNIRS Activity and Performance: A Data-Driven PLS Approach. *bioRxiv*, 2020.2008.2021.261438. doi:10.1101/2020.08.21.261438

Meidenbauer, K. L., Choe, K. W., Cardenas-Iniguez, C., Huppert, T. J., & Berman, M. G. (2021). Load-dependent relationships between frontal fNIRS activity and performance: A data-driven PLS approach. *NeuroImage, 230*, 117795.

Moray, N. (1979). Models and measures of mental workload. *Mental workload: Its theory and measurement*, 13-21.

Muller-Plath, G. (2008). Localizing subprocesses of visual search by correlating local brain activation in fMRI with response time model parameters. *Journal of Neuroscience Methods, 171*(2), 316-330.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological review, 86*(3), 214.

Nazeer, H., Naseer, N., Mehboob, A., Khan, M. J., Khan, R. A., Khan, U. S., & Ayaz, Y. (2020). Enhancing classification performance of fNIRS-BCI by identifying cortically active channels using the z-score method. *Sensors, 20*(23), 6995.

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2013). A meta-analysis of executive components of working memory. *Cerebral cortex, 23*(2), 264-282.

Neupane, A., Saxena, N., & Hirshfield, L. (2017). *Neural underpinnings of website legitimacy and familiarity detection: An fnirs study.* Paper presented at the Proceedings of the 26th International Conference on World Wide Web.

Novi, S. L., Forero, E. J., Rubianes Silva, J. A. I., De Souza, N. G. S., Martins, G. G., Quiroga, A., . . . Mesquita, R. C. (2020). Integration of spatial information increases reproducibility in functional near-infrared spectroscopy. *Frontiers in Neuroscience, 14*, 746.

Obrig, H., Neufang, M., Wenzel, R., Kohl, M., Steinbrink, J., Einhäupl, K., & Villringer, A. (2000). Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults. *Neuroimage, 12*(6), 623-639.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human factors, 56*(3), 476-488.

Peck, E. M., Afergan, D., & Jacob, R. J. (2013). *Investigation of fNIRS brain sensing as input to information filtering systems.* Paper presented at the Proceedings of the 4th Augmented Human International Conference, Stuttgart, Germany.

Peck, E. M., Yuksel, B. F., Ottley, A., Jacob, R. J., & Chang, R. (2013). *Using fNIRS brain sensing to evaluate information visualization interfaces.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France.

Pike, M. F., Maior, H. A., Porcheron, M., Sharples, S. C., & Wilson, M. L. (2014). *Measuring the effect of think aloud protocols on workload using fNIRS.* Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.

Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., & Burgess, P. W. (2020). The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences, 1464*(1), 5-29.

Putze, F., Hesslinger, S., Tse, C.-Y., Huang, Y., Herff, C., Guan, C., & Schultz, T. (2014). Hybrid fNIRS-EEG based classification of auditory and visual perception processes. *Frontiers in Neuroscience, 8*. doi:10.3389/fnins.2014.00373

Recarte, M. A., & Nunes, L. M. (2003). Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied, 9*(2), 119.

Rouse, W. B. (1988). Adaptive Aiding for Human/Computer Control. *Human Factors, 30*(4), 431-443. doi:10.1177/001872088803000405

Saikia, M. J., Kuanar, S., Borthakur, D., Vinti, M., & Tendhar, T. (2021). *A machine learning approach to classify working memory load from optical neuroimaging data.* Paper presented at the Optical Techniques in Neurosurgery, Neurophotonics, and Optogenetics.

Santosa, H., Fishburn, F., Zhai, X., & Huppert, T. J. (2019). Investigation of the sensitivity-specificity of canonical-and deconvolution-based linear models in evoked functional near-infrared spectroscopy. *Neurophotonics, 6*(2), 025009-025009.

Santosa, H., Zhai, X., Fishburn, F., & Huppert, T. (2018). The NIRS Brain AnalyzIR Toolbox. *Algorithms 11*(73).

Sauer, J., Kao, C.-S., & Wastell, D. (2012). A comparison of adaptive and adaptable automation under different levels of environmental stress. *Ergonomics, 55*(8), 840-853.

Schroeter, M. L., Kupka, T., Mildner, T., Uludağ, K., & von Cramon, D. Y. (2006). Investigating the post-stimulus undershoot of the BOLD signal—a simultaneous fMRI and fNIRS study. *Neuroimage, 30*(2), 349-358.

Shattuck, D., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K., . . . Toga, A. (2008). Construction of a 3D probabilistic atlas of human cortical structures. *Neuroimage, 39*(3), 1064-1080.

Shirzadi, S., Einalou, Z., & Dadgostar, M. (2020). Investigation of functional connectivity during working memory task and hemispheric lateralization in left-and right-handers measured by fNIRS. *Optik, 221*, 165347.

Smith, E., & Jonides, J. (1999). Storage and Executive Processes n the Frontal Lobes. *Science, 283*.

Solovey, E., Girouard, A., Chauncey, K., Hirshfield, L., Sassaroli, A., Zheng, F., . . . Jacob, R. (2009). *Using fNIRS Brain Sensing in Realistic HCI Settings: Experiments and Guidelines* Paper presented at the ACM UIST Symposium on User Interface Software and Technology.

Solovey, E., Schoorman, P., Scheutz, M., Sassaroli, A., Fantini, S., & Jacob, R. (2012). *Brainput: Enhancing Interactive Systems with Streaming fNIRS Brain Input* Paper presented at the Proc. ACM Conference on Human Factors in Computing Systems

Solovey, E. T., Afergan, D., Peck, E. M., Hincks, S. W., & Jacob, R. J. K. (2015). Designing Implicit Interfaces for Physiological Computing: Guidelines and Lessons Learned Using fNIRS. *ACM Trans. Comput.-Hum. Interact., 21*(6), 1-27. doi:10.1145/2687926

Solovey, E. T., Lalooses, F., Chauncey, K., Weaver, D., Parasi, M., Scheutz, M., . . . Girouard, A. (2011). *Sensing cognitive multitasking for a brain-based adaptive user interface.* Paper presented at the Proceedings of the SIGCHI conference on Human Factors in Computing Systems.

Solovey, E. T., Okerlund, J., Hoef, C., Davis, J., & Shaer, O. (2015). *Augmenting spatial skills with semi-immersive interactive desktop displays: do immersion cues matter?* Paper presented at the Proceedings of the 6th Augmented Human International Conference.

Strangman, G., Franceschini, M. A., & Boas, D. A. (2003). Factors affecting the accuracy of near-infrared spectroscopy concentration calculations for focal changes in oxygenation parameters. *NeuroImage, 18*(4), 865–879.

Studer, B., Cen, D., & Walsh, V. (2014). The angular gyrus and visuospatial attention in decision-making under risk. *Neuroimage, 103*, 75-80.

Suh, B., Song, I., Jeon, W., Cha, Y., Che, K., Lee, S. H., . . . An, J. (2019). *Cortical regions associated with visual-auditory integration: an fNIRS study.* Paper presented at the 2019 7th International Winter Conference on Brain-Computer Interface (BCI).

Tachtsidis, I., & Scholkmann, F. (2016). False positives and false negatives in functional near-infrared spectroscopy: issues, challenges, and the way forward. *Neurophotonics, 3*(3), 031405. doi:10.1117/1.NPh.3.3.031405

Tomasi, D., Chang, L., Caparelli, E., & Ernst, T. (2007). Different activation patterns for working memory load and visual attention load. *Brain research, 1132*, 158-165.

Tran, N., Grant, T., Phung, T., Hirshfield, L., Wickens, C., & Williams, T. (2021). *Get This!? Mixed Reality Improves Robot Communication Regardless of Mental Workload*. Paper presented at the Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction, Boulder, CO, USA.

Tran, N., Grant, T., Phung, T., Hirshfield, L., Wickens, C., & Williams, T. (2021). *Robot-Generated Mixed Reality Gestures Improve Human-Robot Interaction.* Paper presented at the Social Robotics. ICSR 2021. Lecture Notes in Computer Science.

Unni, A., Ihme, K., Jipp, M., & Rieger, J. W. (2017). Assessing the driver's current level of working memory load with high density functional near-infrared spectroscopy: a realistic driving simulator study. *Frontiers in human neuroscience, 11*, 167.

Volkening, N., Unni, A., Becker, S., Rieger, J. W., Fudickar, S., & Hein, A. (2018). *Development of a Mobile Functional Near-infrared Spectroscopy Prototype and its Initial Evaluation: Lessons Learned.* Paper presented at the Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference.

von Lühmann, A., Zheng, Y., Ortega-Martinez, A., Kiran, S., Somers, D. C., Cronin-Golomb, A., . . . Yücel, M. A. (2021). Toward Neuroscience of the Everyday World (NEW) using functional near-infrared spectroscopy. *Current opinion in biomedical engineering, 18*, 100272.

Wallis, G., Stokes, M., Cousijn, H., Woolrich, M., & Nobre, A. C. (2015). Frontoparietal and cingulo-opercular networks play dissociable roles in control of working memory. *Journal of Cognitive Neuroscience, 27*(10), 2019-2034.

Wickens, C. (1984). Processing resources in attention. In R. Parasuraman & D. Davies (Eds.), *Varieties of attention* (pp. 63–102). New York: Academic Press.

Wickens, C., & Tsang, P. (2014). Handbook of human-systems integration. *Handbook of human-systems integration. American Psychological Association, Washington, DC*.

Wickens, C. D. (1980). The structure of attentional resources. *Attention and performance VIII, 8*, 239-257.

Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2022). *Engineering psychology and human performance*: Routledge.

Wyser, D., Mattille, M., Wolf, M., Lambercy, O., Scholkmann, F., & Gassert, R. (2020). Short-channel regression in functional near-infrared spectroscopy is more effective when considering heterogeneous scalp hemodynamics. *Neurophotonics, 7*(3), 035011.

Xu, J., Slagle, J. M., Banerjee, A., Bracken, B., & Weinger, M. B. (2019). Use of a portable functional near-infrared spectroscopy (fnirs) system to examine team experience during crisis event management in clinical simulations. *Frontiers in human neuroscience, 13*, 85.

Yamamura, H., Baldauf, H., & Kunze, K. (2021). *Hemodynamicvr-adapting the user's field of view during virtual reality locomotion tasks to reduce cybersickness using wearable functional near-infrared spectroscopy.* Paper presented at the Augmented Humans Conference 2021.

Yamazaki, H., Kanazawa, Y., & Omori, K. (2020). Advantages of double density alignment of fNIRS optodes to evaluate cortical activities related to phonological short-term memory using NIRS-SPM. *Hearing Research, 395*, 108024.

Yeh, Y.-y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30*, 111-120.

Yücel, M., Lühmann, A., Scholkmann, F., Gervain, J., Dan, I., Ayaz, H., . . . Wolf, M. (2021). Best practices for fNIRS publications. *Neurophotonics, 8*(1), 012101. doi:10.3390/a11050073

Yücel, M. A., Selb, J. J., Huppert, T. J., Franceschini, M. A., & Boas, D. A. (2017). Functional near infrared spectroscopy: enabling routine functional brain imaging. *Current opinion in biomedical engineering, 4*, 78-86.

Zhai, X., Santosa, H., & Huppert, T. J. (2020). Using anatomically defined regions-of-interest to adjust for head-size and probe alignment in functional near-infrared spectroscopy. *Neurophotonics, 7*(3), 035008.

Zhang, Q., Strangman, G. E., & Ganis, G. (2009). Adaptive filtering to reduce global interference in non-invasive NIRS measures of brain activation: how well and when does it work? *Neuroimage, 45*(3), 788-794.

Zhang, Y., Brooks, D. H., Franceschini, M. A., & Boas, D. A. (2005). Eigenvector-based spatial filtering for reduction of physiological interference in diffuse optical imaging. *Journal of biomedical optics, 10*(1), 011014.

Zhang, Y., & Zhu, C. (2020). Assessing brain networks by resting-state dynamic functional connectivity: An fNIRS-EEG study. *Frontiers in Neuroscience*, 1430.

Zhao, H., & Cooper, R. J. (2018). Review of recent progress toward a fiberless, whole-scalp diffuse optical tomography system. *Neurophotonics, 5*(1), 011012-011012.

Zhao, J., Liu, J., Jiang, X., Zhou, G., Chen, G., Ding, X. P., . . . Lee, K. (2016). Linking resting-state networks in the prefrontal cortex to executive function: a functional near infrared spectroscopy study. *Frontiers in neuroscience, 10*, 452.

Appendix 1: Meta-Review Table

Citation	N	fNIRS Set Up (# of channels, regions)	Device Type	Time Window	Manip of WL (Y/N)	Convergent Measure (NASA-TLX, pupil, secondary task) (Y/N)	Task
(E. T. Solovey, Okerlund, et al., 2015)	48	10, PFC	Imagent (ISS Inc.)	not specified	Y	Y, subjective measures	3D spatial reasoning puzzles
(Maior, Wilson, & Sharples, 2018)	32	16, PFC	fNIR300 (Biopac®)	30s	Y	Y, performance data	Air Traffic control game
(Asgher et al., 2020)	15	12, PFC	P-fNIRS System	20s	Y	Y, NASA-TLX	supervised mental workload experimentation with 4 varying MWL levels
(Hasan Ayaz et al., 2012)	24	16, PFC	fNIR Device model 1000 (NIRDevices LLC.)	not specified	Y	Y, self-reported rating, behavioral measures, NASA-TLX	Study 1: n-back and ATC tasks, Study 2: n-back and UAV tasks
(Peck, Yuksel, et al., 2013)	16	8, right & left PFC	OxiplexTS (ISS Inc.)	40.7s	Y	Y. NASA-TLX & performance data	n-back task and visualizing bar graphs and pie charts
(Afergan et al., 2014)	12	8, right & left PFC	Imagent (ISS Inc.)	25s	Y	Y, dependent measures	UAV Navigation Task and n-back task
(E. T. Solovey, Afergan, Peck, Hincks, & Jacob, 2015)	65	16, PFC	Imagent (ISS Inc.)	calibration phase: 25s	Y	N	n-back and multitasking paradigms
(Durantin et al., 2014)	12	16, PFC	fNIR100 (Biopac®)	6 min	Y	Y, NASA-TLX, HRV, and performance data	simulation of a ROV where they followed a dynamic target with their aircraft under different levels of control difficulty and processing load
(Gateau et al., 2018)	28	16, PFC	fNIR100 (Biopac®)	30s	Y	Y, performance data	flight simulator task with manipulated workload using ATC sending auditory messages and subject repeating them back
(Bunce et al., 2011)	8	8 optodes (# of channels not specified), DLPFC	not specified	75s	Y	Y, secondary task	ship-based navy command and control environment task and a secondary verbal task
(Suh et al., 2019)	9	36, Superior Temporal Gyrus (STG) and Middle Temporal Gyrus (MTG)	FORIE-3000 (Shimadzu Corp.,)	55s	Y	Y, performance data	rhythm game which offered visual-auditory stimulation with synchronous and asynchronous conditions

(Baker, Bruno, Gundran, Hosseini, & Reiss, 2018)	15	40, DLPFC, the left intraparietal sulcus (lPAR), and the right intraparietal sulcus (rPAR)	NIRsport (NIRx)	mixed durations	Y	Y. behavioral data	Visuospatial working memory task and Just Noticeable Difference task
(Herff et al., 2014)	10	8, PFC	Oxymon Mark III (Artinis Medical Systems)	44s	Y	Y, post-experiment questionnaire	n-back task
(L. Hirshfield et al., 2011)	10	8, PFC	OxiplexTS (ISS Inc.)	50s	Y	Y, post-questionnaire survey	Finding A's, Stroop, n-back, driving simulator, conducting web searches
(Putze et al., 2014)	12	170, visual cortex and temporal cortex	Imagent (ISS Inc.)	10-15s*	N	Y, mixed task & EEG	visual and auditory stimuli
(Pike et al., 2014)	20	16, PFC, Brodmann area 10	fNIR300 (Biopac®)	mixed durations	N	Y, EEG and subjective measurement	mathematical computations while speaking aloud
(E. Solovey et al., 2009)	10	8, right & left anterior PFC	OxiplexTS (ISS Inc.)	15s*	N	Y, performance data	Memory task
(Cakir, Çakir, Ayaz, & Lee, 2015)	27	16, PFC	fNIR Devices LLC	15s*	N	Y, accuracy and response time	math tasks
(Friedman, Walker, & Solovey, 2018)	12	8, PFC	8-channel fNIRS device (ISS, Inc.)	40s window	N	Y, behavioral and subjective data	SART Task
(Aranyi, Charles, & Cavazza, 2015)	12	16, PFC	fNIR400 (Biopac®)	20-22s	N	Y, subjective measurement	Anger Task
(E. T. Solovey et al., 2011)	12	8, PFC	OxiplexTS (ISS Inc.)	40s	N	Y, behavioral data	sort rocks and watch location of rocks
(Jin, Jia, & Yu, 2018)	20	38, PFC	LabNIRS, (Shimadzu Corp.)	30s	N	Y, performance data	Solving a science problem
(L.-C. Chen et al., 2015)	24	20, temporal and occipital	NIRScout (NIRx)	10s*	N	Y, EEG & subjective measures	auditory and visual tasks
(Aihara et al., 2020)	20	152, bilateral frontal and parietal	SMARTNIRS (Shimadzu Corp.)	not specified	N	Y, fMRI	10 min resting state condition and two-back working memory task
(Yujin Zhang & Zhu, 2020)	20	40, frontal, sensimotor, occipital, temporal, pariental	LabNIRS (Shimadzu Corp.)	20, 30, and 60 s	N	Y, EEG	Resting state with eyes opened and then eyes closed
(Y. Chen et al., 2020)	19	105, covering the areas from the forehead to the occipital lobe	NirScout (NIRx)	30s	N	Y, EEG	Exp 1: two separate sessions, eyes-open (EO) and eyes-closed (EC), while standing, sitting, and supine. Exp 2: rest still and allowed to fall asleep during a 45-min recording, while subjects laid supine in an adjustable recliner
(E. Solovey et al., 2012)	11	2, anterior PFC	OxiplexTS (ISS Inc.)	40s	N	Y, NASA-TLX	Robot navigation task
(Anderson et al., 2017)	29	16, PFC	fNIRS Devices LLC.	50s	N	Y, Non-verbal and Verbal	Toddlers underwent a vanilla baseline

						Developmental Quotients	recording, during which they watched two 50-s clips from children's shows, presented in audiovisual format
(Neupane, Saxena, & Hirshfield, 2017)	20	46, Whole head (frontal, temporal, parietal, and occipital lobes)	ETG 4000 (Hitachi)	20s	N	Y, fMRI	trust task for websites
(Aksoy et al., 2019)	22	16, PFC	fNIR Devices, LLC.	15s*-120s	N	Y, BLS scores	VR learning task
(Yamamura, Baldauf, & Kunze, 2021)	10	not specified, PFC	HOT-1000 (Hitachi Medical Systems)	5 min	N	Y, simulator sickness questionnaire (SSQ)	VR task
(Yamazaki, Kanazawa, & Omori, 2020)	14	53, Left hemisphere (frontal, temporal, and parietal lobes)	LABNIRS (Shimadzu Corp.)	3-25s*	N	N	Pseudoword audio and visual STM tasks
(E. Holmes et al., 2019)	34	20, PFC	NirScout (NIRx)	2-6 min	N	N	Psychomotor vigilance task and delayed match-to-sample task
(Geng et al., 2017)	21	46, Whole head (frontal, temporal, parietal, and occipital lobes)	CW6, (TechEn Inc.)	11 min	N	N	resting state
(J. Zhao et al., 2016)	90	24, bilateral PFC	ETG-4000 (Hitachi Medical Systems)	3s*	N	N	CANTAB - Stockings of Cambridge (SOC), Spatial Working Memory (SWM), Spatial Span (SSP), and Intra-dimensional/Extra-dimensional Shifts (IED)
(Lei, Miyoshi, Dan, & Sato, 2020)	131	22, bilateral frontal and temporal areas	ETG-4000 (Hitachi Medical Systems)	18-21s	N	N	Listening task in different languages
(Charles, De Castro Martins, & Cavazza, 2020)	S1:11 S2: 17 S3: 11	16, Right & left DLPFC	fNIR400 (Biopac®)	S1: 22s S2: 40s S3: 40s	N	N	Expression anger, engaging with virtual character, and expressing motivation
(Keshmiri, Sumioka, Okubo, & Ishiguro, 2019)	S1: 34 S2: 36 S3: 26	4, PFC	HOT-1000 (Hitachi)	S1: 6s* S2: 20 min S3: 50s	N	N	verbal fluency task, conversation task, logical memory test
(Novi et al., 2020)	10	64, primary and secondary motor cortices, frontal, and parietal	NirScout (NIRx)	2s*	N	N	Right-hand finger-tapping
(Y. Liu & Ayaz, 2018)	19	40, anterior PFC and parietal cortex	fNIR Imager Model 1100; (fNIR Devices), LLC and ETG 4000 (Hitachi Medical Systems)	100s task, 25-50s segments for classification	N	N	Listening to English stories
(Abdullah & Khan, 2018)	9	17, temporal, parietal, and occipital lobes	forty four-channel NIRS device	3-5s*	N	N	Color association task
(Lee, Jung, Park, & Hong, 2019)	7	20, M1, M2 and SMA, approx. 3 cm apart	NirSport 8x8 (NIRx)	20s	N	N	Finger tapping task

(Peck, Afergan, & Jacob, 2013)	14	8, PFC	OxiplexTX (ISS Inc.)	25s	N	N	Viewing favorite and least favorite movies
(Volkening et al., 2018)	S1:5 S2: 9	S1: 16 channels S2: 10-5 layout, Brodmann areas 1-4 and 6	mofNIRS & NirScout (NIRx)	S1: 2 min S2: 15s*	N	N	Grip movements of the hands using hand-held strength trainers
(Le, Xuan, & Aoki, 2022)	17	4, PFC	Brain Activity Monitor (Astem Co., Ltd.)	1s	Y	N	Driving simulator
(Chu et al., 2022)	20	2, PFC	PORTALITE (Artinis)	3s	Y	Y	MATB - monitoring task, tracking task, and oil management task
(MacNeil, Bishop, & Izzetoglu, 2022)	9	18, PFC	fNIR Devices 2000M (fNIR Devices LLC.)	8-10s	Y	Y	Simulated use of force training for federal law enforcement
(İşbilir, Çakır, Acartürk, & Tekerek, 2019)	14	16, PFC	Imager 1002 (fNIR Devices LLC.)	100-600s	N	Y	Military landing platform exercise
(R McKendrick et al., 2016)	20	All studies: 4, PFC	fNIR Devices Model 1100W	60s	N	Y	Route navigation with n-back secondary task
(Kerr, Reddy, Shewokis, & Izzetoglu, 2022)	7	18, PFC	fNIR Devices	12 min	Y	Y	Scanning UAV sensor operator images (Simlat C-STAR)
(Li, Li, Xie, & Chang, 2022)	26	8, PFC	OctaMon (Artinis Medical Systems)	180s	Y	Y	Flight task with secondary subtasks: flight target tracking, meter monitoring, emergency handling, residual capacity
(Izzetoglu, Bunce, Onaral, Pourrezaei, & Chance, 2004)	8	16, PFC	Drexel BME Device	75s	Y	Y	Warship Commander Task (WCT)

Appendix 2. fNIRS Statistical Results

Table A: Channel-wise results for ($WM^{high}VL^{low}$)>($VL^{high}WM^{low}$) contrast. With type of hemoglobin (Hb), source-detector position (S D), functional brain regions, and Brodmann Area (BA) listed, as well as T, p, q, power values.

Contrast: ($WM^{high}VL^{low}$)>($VL^{high}WM^{low}$)							
Hb	S D;	Region	BA	T	P	q	power
HbO	1 1;	L anterior frontal gyrus	9, 11, 46	-2.77	0.006	0.033	0.61
HbO	2 2;	L middle frontal gyrus	45, 46	-2.82	0.005	0.029	0.63
HbO	4 3;	L superior frontal gyrus	8, 9	-2.87	0.004	0.028	0.64
HbO	4 4;	L middle frontal gyrus	6, 9, 8	-4.26	<0.001	<0.001	0.96
HbO	6 5;	R anterior frontal gyrus	8, 9	-2.66	0.008	0.043	0.56
HbO	6 6;	R middle frontal gyrus	45, 46	-4.76	<0.001	<0.001	0.99
HbO	6 8;	R middle frontal gyrus	9, 44, 46, 45	-3.05	0.002	0.018	0.71
HbO	8 7;	R superior frontal gyrus	8, 9	-5.76	<0.001	<0.001	0.99
HbO	11 9;	R superior parietal gyrus	7	3.00	0.003	0.020	0.70
HbO	11 12;	R superior parietal gyrus	7	3.29	0.001	0.011	0.79
HbO	12 9;	L superior parietal gyrus	7	3.28	0.001	0.011	0.78

HbO	12 15;	L angular gyrus	7, 19, 39	3.10	0.002	0.016	0.73
HbO	13 14;	R angular gyrus	39, 40, 7	2.59	0.010	0.049	0.54
HbO	14 16;	L supramarginal gyrus	40, 39, 48	-5.79	<0.001	<0.001	0.99
HbR	3 2;	L inferior frontal gyrus	9, 44, 45, 46	-3.73	<0.001	0.004	0.89
HbR	4 3;	L superior frontal gyrus	8, 9	-3.65	<0.001	0.004	0.88
HbR	7 8;	R precentral gyrus	44, 6, 9	3.01	0.002	0.020	0.70
HbR	10 13;	L superior occipital gyrus	18, 19, 17	-2.62	0.009	0.047	0.55

Table B: Channel wise results for the Working Memory Main Effects ($H^{WM} – L^{WM}$) and VL Main Effects $H^{VL} – L^{VL}$ contrasts. With type of hemoglobin (Hb), source-detector position (S D), functional brain regions, and Brodmann Area (BA) listed, as well as T, p, q, power values, and uniqueness. The 'unique?' column in Table 3 has a 'Y' for all channels that are unique to the VL and WM main effects comparisons (present in one of VL or WM, but not present in the other).

Contrast: Working Memory Main Effect ($H^{WM} – L^{WM}$)								
Hb	**S D;**	**Region**	**BA**	**T**	**p**	**q**	**power**	**unique?**
HbO	10 13;	L superior occipital gyrus	18, 19, 17	3.46	0.001	0.007	0.83	Y
HbO	14 10;	L superior parietal gyrus	40, 7, 2	3.26	0.001	0.011	0.78	N
HbR	14 10;	L superior parietal gyrus	40, 7, 2	-4.75	<0.001	<0.001	0.99	N
HbR	2 2;	L middle frontal gyrus	45, 46	-3.99	<0.001	0.002	0.93	N
HbR	1 2;	L inferior frontal gyrus	45, 46	-2.57	0.011	0.049	0.53	Y
HbR	7 8;	R precentral gyrus	44, 6, 9	-3.83	<0.001	0.003	0.91	Y
HbR	11 14;	R angular gyrus	7, 19, 39	-3.13	0.002	0.015	0.74	Y
HbR	8 8;	R middle frontal gyrus	6, 9, 8	-2.94	0.004	0.023	0.67	N
HbR	6 6;	R middle frontal gyrus	45, 46	-3.5	0.001	0.006	0.84	N
Contrast: VL Main Effect ($H^{VL} – L^{VL}$)								
Hb	**S D;**	**Region**	**BA**	**T**	**P**	**q**	**power**	**unique?**
HbO	14 10;	L superior parietal gyrus	40, 7, 2	3.78	<0.001	0.003	0.9	N
HbO	12 9;	L superior parietal gyrus	7	3.35	0.001	0.009	0.81	N
HbO	14 15;	L angular gyrus	39, 40, 7	2.93	0.004	0.024	0.67	Y
HbO	12 15;	L angular gyrus	7, 19, 39	2.85	0.005	0.027	0.64	Y
HbR	14 10;	L superior parietal gyrus	40, 7, 2	-2.9	0.004	0.025	0.66	N
HbR	2 2;	L middle frontal gyrus	45, 46	-3.54	<0.001	0.006	0.85	N
HbR	8 8;	R middle frontal gyrus	6, 9, 8	-2.8	0.006	0.031	0.62	N
HbR	6 7;	R middle frontal gyrus	9, 46	-2.98	0.003	0.021	0.69	N

Table 1: Seven standards for experiments needed to evaluate fNIRS for workload-based AA.

(1) Participants should perform a complex task typical of real-world human-computer interactions.
(2) Workload should be experimentally manipulated in a controlled manner to impose greater or lesser cognitive demands (going beyond just load on/off), in order to evaluate sensitivity of different load levels on a specific resource.
(3) Studies should focus on different specific resources within a multiple resource structure, hence examining diagnosticity
(4) The validity of experiment task manipulations should be assured by including additional workload measures, such as self-report workload, response time, performance, and pupil diameter.
(5) To further examine the diagnosticity of the measures, researchers should measure multiple functional brain regions of interest (ROIs), ideally mapped onto the multiple resources identified in the experimental design, in order to determine if specific ROIs are differentially sensitive to the workload manipulation assumed to be reflected by increased activation there.
(6) Increased activation should be explored via the two different fNIRS measures of HbR and HbO.
(7) Finally, studies should have adequate statistical power, with a suitable N.

Table 2: Meta-review studies reviewed that adhered to atleast three of the seven standards, ordered by # of standards adhered to.

Author	# standards	complexity	workload manipulation	diff resources	convergent measures	multiple ROI	HBO-HBR	N
Isbilir	3		y		y		y	14
Chu	3	y	y				y	20
Lei	3				y	y		131
Hamann	3	y	y	y				35
Izzetoglu	3	y	y		y			8
Peck	3	y		y	y			16
Solovey	4	y	y		y		y	48
McKendrick	4	y			y	y	y	20
Ayaz	4	y	y		y		y	16
Durantin	4	y	y		y			12
Kerr	5	y	y		y	y	y	7
Putze	5		y	y	y	y	y	12
Gateau	6	y	y	y	y	y	y	28

Table 3: HbO and HbR ß-values and T-values for the main effects (IV stands for independent variable) of the WM and VL manipulations, averaged across each ROI (**bold face with a *** denotes significance ($p<0.05$)). The contrasts run are $(H^{wm}L^{vl} + H^{wm}H^{vl}) - (L^{wm}L^{vl} + L^{wm}H^{vl})$ for the WM load main effect (first row for each variable: ß or T) and $(L^{wm}H^{vl} + H^{wm}H^{vl}) - (L^{wm}L^{vl} + H^{wm}L^{vl})$ for the VL main effect (second row of each variable).

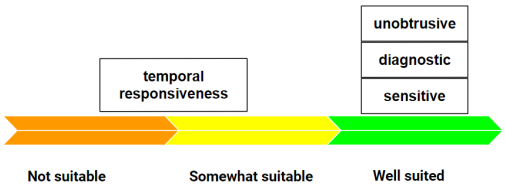
HbO						**HbR**				
ß	WM ROI	VL ROI	MT ROI	AL ROI		**ß**	WM ROI	VL ROI	MT ROI	AL ROI

WM IV	-0.65	2.59	-1.56	-0.29	WM IV	**-3.61***	**-2.60***	**-2.82***	-1.01
VL IV	**-4.26***	**3.70***	**-5.93***	-0.78	VL IV	**-3.52***	**-2.08***	-1.46	0.12
T	WM ROI	VL ROI	MT ROI	AL ROI	**T**	WM ROI	VL ROI	MT ROI	AL ROI
WM IV	-0.36	1.9	-0.67	-0.15	WM IV	**-5.07***	**-4.71***	**-2.48***	-1.23
VL IV	**-2.35***	**2.70***	**-2.53***	-0.4	VL IV	**-4.91***	**-3.76***	-1.28	0.14

Table 4: Summary of results from Table 3, with only unique significant activation shown.

Mutually exclusive regions activated for Working Memory main effects		**Mutually exclusive regions activated for VL main effects**	
Hb	Region	Hb	Region
HbO	L superior occipital gyrus	HbO	L angular gyrus (x2 channels)
HbR	L inferior frontal gyrus		
HbR	R precentral gyrus		
HbR	R angular gyrus		

Table 5: Summary of findings collated from our experiment and the meta-review. Top: a graphic depicting our findings, with mappings on the suitability of fNIRS for workload-based AA based on the four criteria of temporal responsiveness, unobtrusiveness, diagnosticity, and sensitivity. Bottom: Text summary of our findings regarding the four criteria.



Criteria	**Summary of Meta-Review and Experimental Findings**
Unobtrusiveness	**Meta-review** findings show a strong trend toward devices continuing to be more wearable, practical, and specialized to specific use cases. **Empirical** results were achieved in this study using a NIRSport 2. The wireless NIRSport2 was equipped with probe tips specially designed for comfort on the scalp.
Temporal responsiveness	**Meta-review** findings suggest that like fMRI, fNIRS on its own, measures a slowly moving hemodynamic response, which makes its temporal responsiveness relatively slow. More work is needed, following the lead of researchers who have focused on exploring short sliding windows of time in ML classification and on hybrid EEG/fNIRS adaptive systems. **Empirical** results were generated using statistical tests on task lengths of 45 seconds in duration, which does not further our understanding of the temporal responsiveness of the fNIRS signal.

Sensitivity	**Meta-review** showcased strong results of the sensitivity to workload manipulations; however with the majority of work focused on extremely simple, highly controlled benchmark tasks (e.g. n-back tasks), rather than those tasks typical of an extra-laboratory working environment. **Empirical** results indicate that fNIRS is sensitive to changes in visual) and working memory load levels. HbR appeared to be more sensitive than HbO for WM, while both HbR and HbO appeared to be sensitive to VL manipulations (as shown in Tables 3 and 4).
Diagnosticity	**Meta-Review** found very little prior work on diagnosticity. Of that handful of work, the vast majority has been done on simple, highly controlled tasks. **Empirical** results indicate that fNIRS is diagnostic to type of load, specifically to visual vs WM, but these findings are not clear cut. In the channel-wise analysis, we found unique regions that are activated in the WM main effects comparison that were not activated by the VL main effects, and vice versa. Such diagnosticity was not revealed by the ROI analysis where we condensed the data into four ROIs. When the data was kept in its channel-wise form, we did see diagnosticity: For WM, we see one HbO channel and three HbR channels that are unique for differentiating WM (HbO: Left superior occipital gyrus, HbR: L inferior frontal gyrus, R precentral gyrus, R angular gyrus). For diagnosticity for VL, we see two HbO channels uniquely differentiating VL, both measure the L angular gyrus.

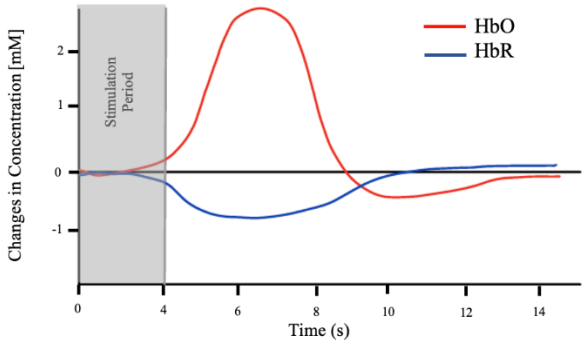


Figure 1: Typical time response of HbO and HbR after stimulus (such as completing a n-back task). HbO peaks between 6-8s following the stimuli and HbR dips at the same time.

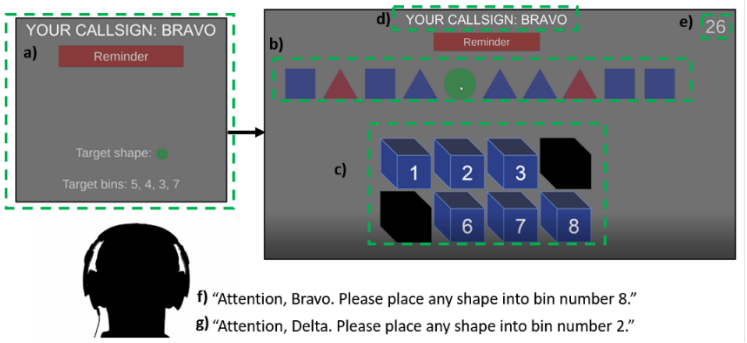


Figure 2: The shape sorting testbed. The a) instruction screen directs the participant on the primary task target shape and target bins. Participants then sort the correct target shape out of a list of possible shapes (b) into the correct numbered bins (c). Participants are assigned a callsign (d) and a secondary auditory task is presented through the right side of the headphones, where the information can either be ignored (g) or where it must be attended to (f). Each task session lasts 45 seconds with time being counted down (e), before filling out surveys and beginning a new task, with new updated instructions.

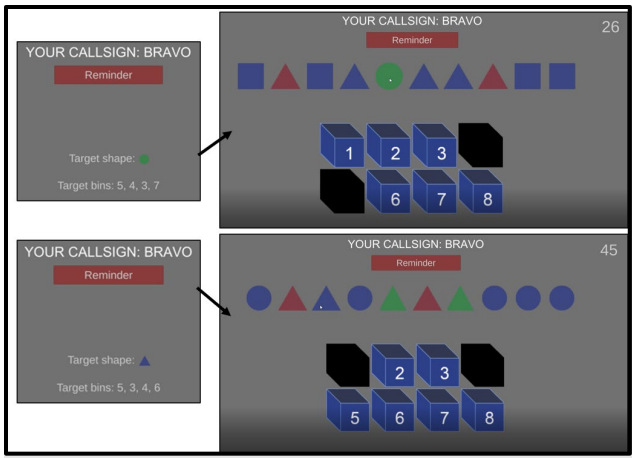


Figure 3: Left (instructions before a task begins). Top right: An example of a search task where the VL is low because target and distractors share no features in common. Bottom right: an example of a search task where the VL is high: 1 feature is shared. See (Nhan Tran et al., 2021) for an example of this task implemented in a mixed reality context.

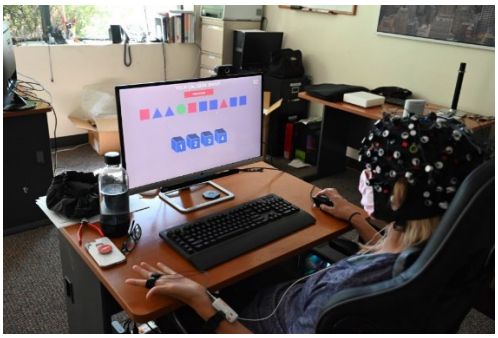


Figure 4: Sensor set-up included a Tobii 4c eye tracker and a NIRx Sport2 fNIRS device.

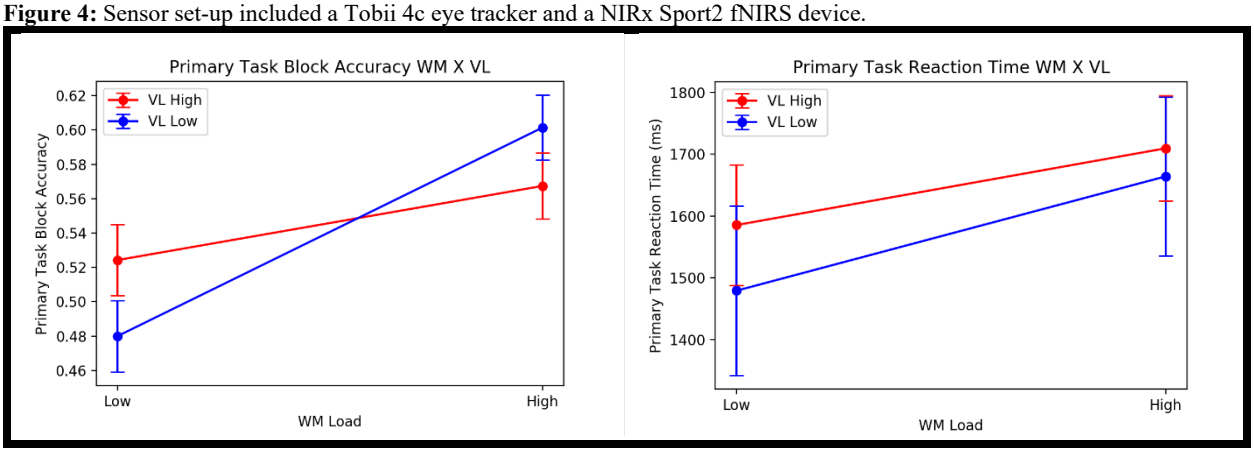


Figure 5: Left: Effect of WM and VL on primary task accuracy. Right: The effects of WM and VL on primary task RT. The error bars represent the unbiased one standard error as implemented by the Pandas sem function.

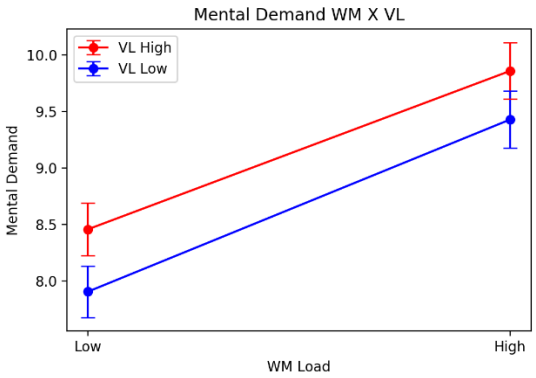
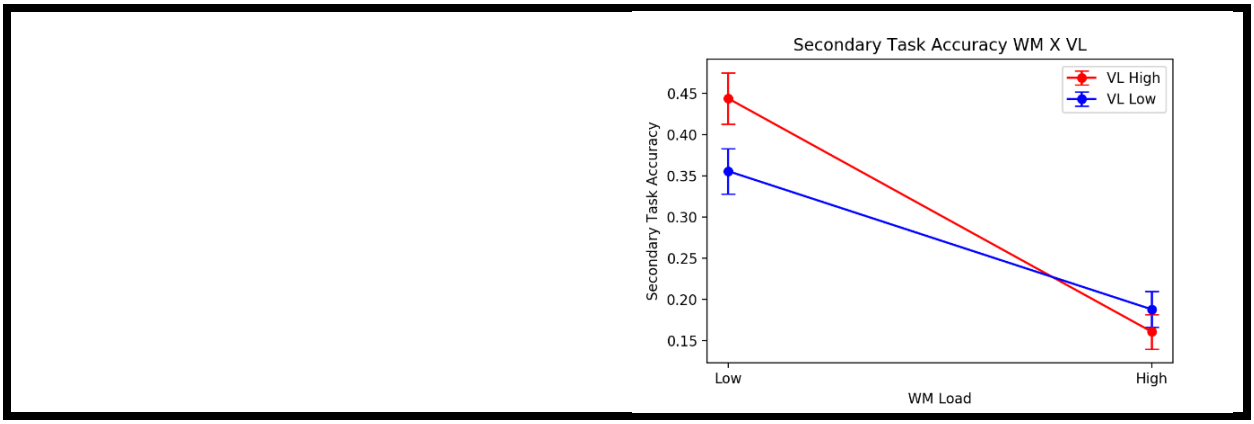


Figure 6. Effects of WM and VL on self-report mental demand. Bottom Right: Effects of WM and VL on secondary task accuracy. The error bars represent the unbiased one standard error as implemented by Pandas 'sem' function.



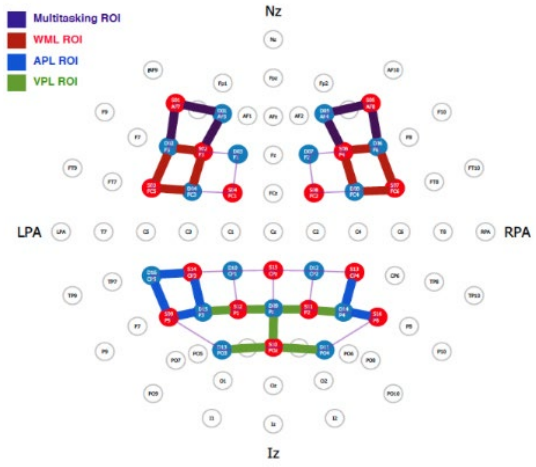


Figure 7: Regions of interest overlaid over the 42-channel fNIRS montage. Red circles represent light sources, blue circles represent light detectors. Red, green, blue, and purple lines represent a channel of measured data that falls into the WM, VL, AL, and MT ROIs, respectively. In the schematic picture, Nz represents the nasion, Iz represents the inion, LPA and RPA represent the left and right pre-auricular regions, respectively (used in standard EEG 10-20 landmarking).

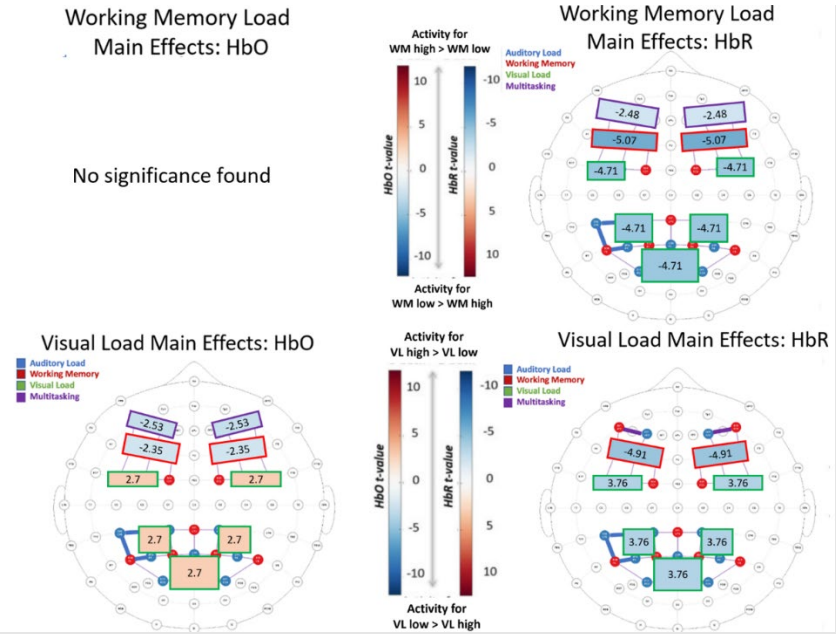
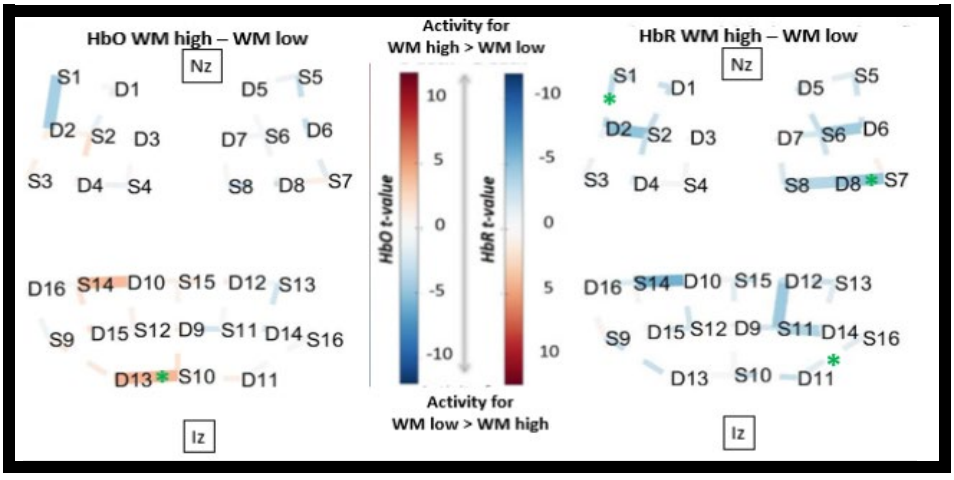


Figure 8: T-values from Table 3, overlayed over the fNIRS ROIs of multitasking, WM, VL, and AL. For HbO (left side) the red spectrum indicates increased activation, with darker red indicating more increased activation. For HbR (right) blue suggests more activation at that region, with darker blue indicating higher levels of activation.



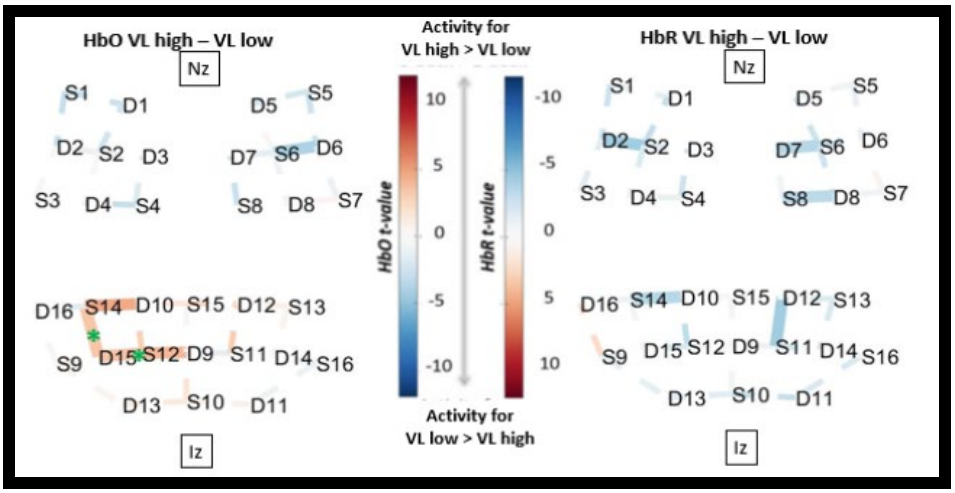
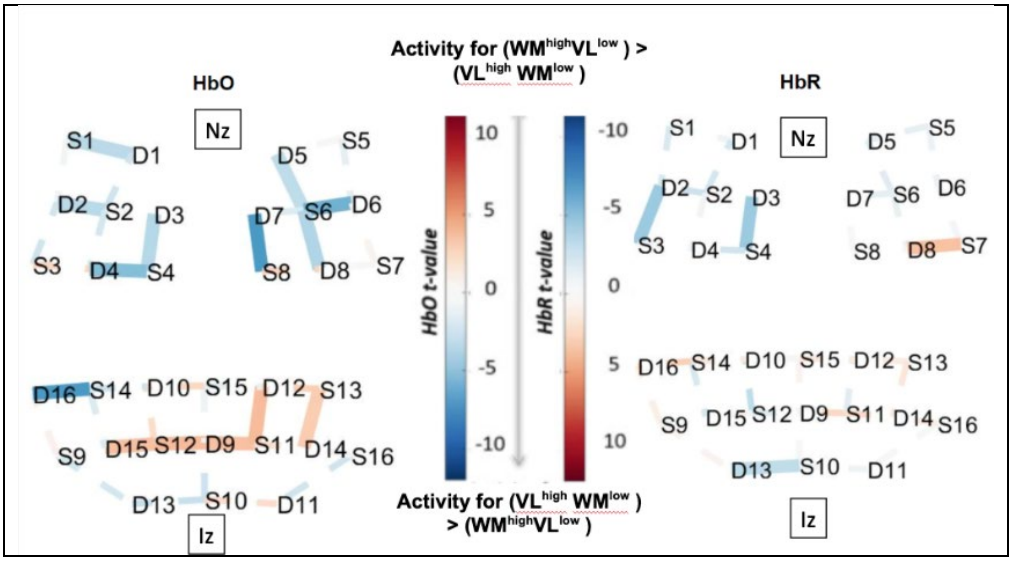


Figure 9: Working Memory Main Effects ($H^{WM} – L^{WM}$) and VL Main Effects $H^{VL} – L^{VL,}$ overlaid over a brain, with nasion (Nz) and inion (Iz) locations added for reference. Only significant channels ($q < 0.05$) are shown. For HbO, positive *t*-values (red) correspond to relatively larger activity for the first term in the contrast, and negative *t*-values (blue) correspond to larger activity for the second term. For HbR contrasts, negative *t*-values (blue) correspond to larger activation in that region. Green * indicates regions that are mutually exclusive (shown by one but not the other) between the WM and VL main effects tests.



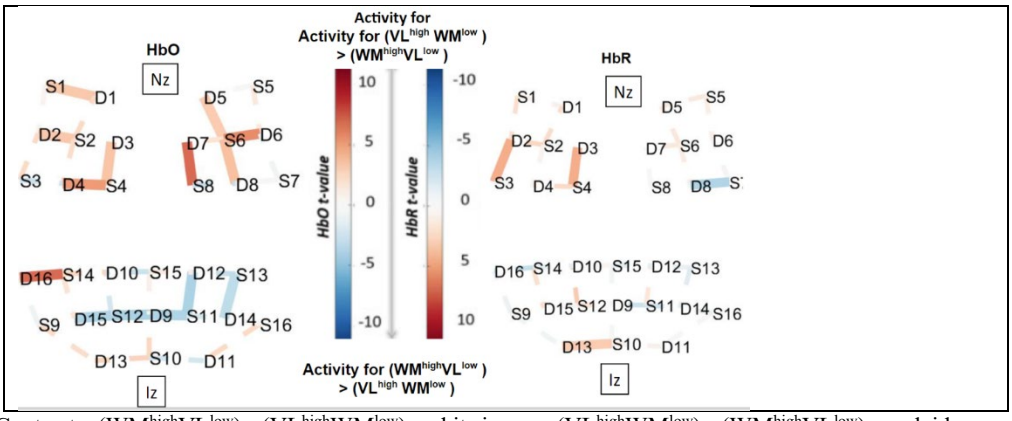


Figure 10: Contrasts: ($WM^{high}VL^{low}$) - ($VL^{high}WM^{low}$) and its inverse ($VL^{high}WM^{low}$) - ($WM^{high}VL^{low}$) overlaid over a brain, with nasion (Nz) and inion (Iz) locations added for reference. Only significant channels ($q < 0.05$) are shown. For HbO, positive *t*-values (red) correspond to relatively larger activity for the first term in the contrast, and negative *t*-values (blue) correspond to larger activity for the second term. For HbR contrasts, negative *t*-values (blue) correspond to larger activation in that region.

Figure Captions

Figure 1: Typical time response of HbO and HbR after stimulus (such as completing a n-back task). HbO peaks between 4-6s following the stimuli and HbR dips at the same time.

Figure 2: The shape sorting testbed. The a) instruction screen directs the participant on the primary task target shape and target bins. Participants then sort the correct target shape out of a list of possible shapes (b) into the correct numbered bins (c). Participants are assigned a callsign (d) and a secondary auditory task is presented through the right side of the headphones, where the information can either be ignored (g) or where it must be attended to (f). Each task session lasts 45 seconds with time being counted down (e), before filling out surveys and beginning a new task, with new updated instructions.

Figure 3: Left (instructions before a task begins). Top right: An example of a search task where the VL is low because target and distractors share no features in common. Bottom right: an example of a search task where the VL is high: 1 feature is shared. See (Nhan Tran et al., 2021) for an example of this task implemented in a mixed reality context.

Figure 4: Sensor set-up included a Tobii 4c eye tracker and a NIRx Sport2 fNIRS device.

Figure 5: Left: Effect of WM and VL on primary task accuracy. Right: The effects of WM and VL on primary task RT. The error bars represent the unbiased one standard error as implemented by the Pandas sem function.

Figure 6. Effects of WM and VL on self-report mental demand. Bottom Right: Effects of WM and VL on secondary task accuracy. The error bars represent the unbiased one standard error as implemented by Pandas 'sem' function.

Figure 7: Regions of interest overlaid over the 42-channel fNIRS montage. Red circles represent light sources, blue circles represent light detectors. Red, green, blue, and purple lines represent a channel of measured data that falls into the WM, VL, AL, and MT ROIs, respectively. In the schematic picture, Nz represents the nasion, Iz represents the inion, LPA and RPA represent the left and right pre-auricular regions, respectively (used in standard EEG 10-20 landmarking).

Figure 8: T-values from Table 3, overlayed over the fNIRS ROIs of multitasking, WM, VL, and AL. For HbO (left side) the red spectrum indicates increased activation, with darker red indicating more increased activation. For HbR (right) blue suggests more activation at that region, with darker blue indicating higher levels of activation.

Figure 9: Working Memory Main Effects ($H^{WM} - L^{WM}$) and VL Main Effects $H^{VL} - L^{VL,}$ overlaid over a brain, with nasion (Nz) and inion (Iz) locations added for reference. Only significant channels ($q < 0.05$) are shown. For HbO, positive *t*-values (red) correspond to relatively larger activity for the first term in the contrast, and negative *t*-values (blue) correspond to larger activity for the second term. For HbR contrasts, negative *t*-values (blue) correspond to larger activation in that region. Green * indicates regions that are mutually exclusive (shown by one but not the other) between the WM and VL main effects tests.

Figure 10: Contrasts: ($WM^{high}VL^{low}$) - ($VL^{high}WM^{low}$) and its inverse ($VL^{high}WM^{low}$) - ($WM^{high}VL^{low}$) overlaid over a brain, with nasion (Nz) and inion (Iz) locations added for reference. Only significant channels ($q < 0.05$) are shown. For HbO, positive *t*-values (red) correspond to relatively larger activity for the first term in the contrast, and negative *t*-values (blue) correspond to larger activity for the second term. For HbR contrasts, negative *t*-values (blue) correspond to larger activation in that region.

Author Bios

Leanne Hirshfield is an Associate Research Professor in the Institute of Cognitive Science at University of Colorado, Boulder. She directs the System Human Interaction with NIRS and EEG (SHINE) Lab. Her research explores the use of non-invasive brain measurement to classify users' social, cognitive, and affective states to support adaptive systems.

Chris Wickens is a Professor at Colorado State University. He has been a productive researcher in human factors for over 30 years, publishing over 250 publications in refereed journals and book chapters and has authored or co-authored 8 books including Introduction to Human Factors; Designing for People, and Engineering Psychology & Human Performance.

Emily Doherty is a PhD student in the Department of Computer Science and Institute of Cognitive Science at the University of Colorado Boulder advised by Leanne Hirshfield. Her research is focused on how physiological signals, specifically brain patterns, can be used to improve team collaboration and human-AI interactions.

Cara Spencer is a computer science PhD student in the computer science department at University of Colorado Boulder. Her research interests are in measurement and modeling of human and team performance through non-invasive, unobtrusive measurement and machine learning, focusing on high performing populations such as pilots and astronauts in extreme conditions.

Tom Williams is an Associate Professor of Computer Science at the Colorado School of Mines. Tom earned a joint PhD in Computer Science and Cognitive Science from Tufts University in 2017. Tom's research focuses on human-robot interaction that is sensitive to environmental, cognitive, social, and moral context.

Lucas Hayne is a computer science PhD student at the University of Colorado Boulder. His research develops methods for analyzing neural data from both artificial and biological neural networks. These methods improve neural network training and interpretability by elucidating the connection between neural network representations and performance.