Adaptive aiding with an individualized workload model based on psychophysiological measures

Grace Teo¹ · Gerald Matthews¹ · Lauren Reinerman-Jones¹ · Daniel Barber¹

Received: 13 February 2019 / Accepted: 13 November 2019 / Published online: 28 November 2019
© The Author(s) 2019

Abstract

Potential benefits of technology such as automation are oftentimes negated by improper use and application. Adaptive systems provide a means to calibrate the use of technological aids to the operator's state, such as workload state, which can change throughout the course of a task. Such systems require a workload model which detects workload and specifies the level at which aid should be rendered. Workload models that use psychophysiological measures have the advantage of detecting workload continuously and relatively unobtrusively, although the inter-individual variability in psychophysiological responses to workload is a major challenge for many models. This study describes an approach to workload modeling with multiple psychophysiological measures that was generalizable across individuals, and yet accommodated inter-individual variability. Under this approach, several novel algorithms were formulated. Each of these underwent a process of evaluation which included comparisons of the algorithm's performance to an at-chance level, and assessment of algorithm robustness. Further evaluations involved the sensitivity of the shortlisted algorithms at various threshold values for triggering an adaptive aid.

Keywords Adaptive systems · Inter-individual variability · Multiple psychophysiological measures · Workload modeling

1 Introduction

One recurrent goal in human systems engineering is to have the ability to adapt the use of technology to the workload needs of the operator. This issue has often been explored with the use of adaptive systems whose function and behavior can be adjusted according to changes in the operator's mental workload state during task performance.

2 Mitigating the adverse effects of workload with adaptive aiding

Adaptive systems that support human performance have been developed and designed with increasingly sophistication and complexity over the years (Karwowski 2012; Dorneich et al. 2016). In their most basic form, they are closed-loop control systems that auto-regulate their effects within a changing environment to fulfill certain criteria or maintain a “set point.” To accomplish this aim, the environment is continuously monitored and its state is assessed against a target criterion or “set point.” When the state deviates from the target criterion, the system acts to return the state to the desired level. Refinements to the basic closed-loop system include targeting a range of criterion values, as in an autopilot that keeps an airplane within a safety envelope, control of dynamic behaviors such as preventing an overshoot of the target state, and tracking a changing criterion as in adaptive cruise control for vehicles. Adaptive systems are best known in engineering contexts such as vehicle control, but there is increasing interest in building systems that can regulate operator functional states such as workload, stress, and fatigue (e.g., Hockey 2003). That is, performance is regulated indirectly rather than directly,

✉ Grace Teo
gteo@ist.ucf.edu

Gerald Matthews
gmatthew@ist.ucf.edu

Lauren Reinerman-Jones
lreinerm@ist.ucf.edu

Daniel Barber
dbarber@ist.ucf.edu

¹ Institute for Simulation and Training, University of Central Florida, 3100 Technology Parkway, Orlando, FL 32826, USA

through supporting the operator's readiness to deal with a range of performance challenges.

The present study addresses adaptive automation designed to limit mental workload as it fluctuates over the duration of a cognitively demanding task (e.g., Freeman et al. 2000; Prinzel et al. 2000; Bailey et al. 2006). Mental workload has been defined in terms of the attentional resources that are needed to meet task demands ("taskload"), which may be mediated by the operator's functional state, past experience, and external support (Hockey 2003; Young and Stanton 2005; Matthews et al. 2015a). It is the result of the combination of task features, environmental factors, and operator characteristics (Young et al. 2015). Extreme levels of workload can be detrimental to task performance (Young and Stanton 2002; Young et al. 2015), so a system that adjusts its behavior to keep the operator's workload level within an optimum range (Hancock and Warm 1989) would be useful. In order to do this, the adaptive system would have sensors that measure and monitor the level of operator workload such that when workload reaches an excessive level, the system can then adapt appropriately to the operator's state by rendering a suitable aid to relieve task load, which, in turn, influences workload (Hancock and Caird 1993; Matthews and Reinerman-Jones 2017; Hancock and Matthews 2019). In safety-critical domains, a system with such a capability can contribute to a reduction in accidents and errors that may result from fatigue, attention lapses, distractions, or boredom that are typically precipitated by operator overload or underload (Brookhuis and Waard 2001; Young et al. 2015).

While there are various measures of workload, several characteristics of psychophysiological workload measures make them suitable for use in adaptive systems. Unlike subjective and self-report measures, they are objective and do not disrupt the task since they do not require any overt response from the operator, who may also not have accurate insight into his or her own level of workload, especially when deeply engaged in the task (Kantowitz and Casper 2017). Psychophysiological workload measures allow continuous monitoring, providing high temporal resolution of operator state. Unlike performance-based workload measures, psychophysiological workload measures can be used to preempt performance declines before operational effectiveness is compromised.

3 Inter-individual variability in psychophysiology

The basis for using psychophysiological measures is that with activation of certain mental processes required for the task, there is a corresponding physiological response that reflects this mental activity. Although large inter-individual variability in mental workload is observed with all workload measures

for the same task performed in the same environment, the issue is particularly troublesome with psychophysiological workload measures (Hancock et al. 1985; Roscoe 1993; Johannes and Gaillard 2014). First, there is wide variability in individual physiology. Psychophysiological workload measures reflect a variety of distinct responses. These include central brain activity assessed using the electroencephalogram (EEG), and peripheral systems such as pupil diameter and cardiac activity measured with the electrocardiogram (ECG). Workload is also indexed by slower hemodynamic responses reflecting metabolic activity, i.e., cerebral blood flow velocity (CBFV) and regional oxygen saturation (rSO₂). Workload responses differ across individuals of various ages, gender, levels of cardiovascular fitness, and physical health. For instance, hypertension impacts ECG signals and cerebral blood flow, and there are age differences present in EEG and pupil diameter (Birren et al. 1950; Bill and Linder 1976; Winn et al. 1994; Pierce et al. 2003; Ang and Lang 2008).

Individuals' psychological responses to the same task demands also vary widely, which can require multiple measures for assessment. Even a well-defined task may not implicate the same set of mental processes in different individuals since measures that index autonomic and central nervous system function can dissociate. For instance, in performing the same task, one individual may show greater changes in brain activity while another may show more changes in cardiac activity (Matthews et al. 2015b). There are also other reasons to use multiple psychophysiological measures in adaptive systems. Different measures are sensitive to different task demands such that one measure may capture the levels of certain workload manipulations, while others may not (Wilson and O'Donnell 1988; Matthews et al. 2015b). For example, an EEG-based workload index and eye fixation durations were sensitive to the single/dual task workload manipulation, but rSO₂ and heart rate variability (HRV) discriminated between different levels of certain single tasks instead (Matthews et al. 2015b). This finding suggests that especially for multitasking environments, the determination of workload levels should not be based only on one measure.

4 Workload models for adaptive aiding

In neuro-ergonomic applications, adaptive systems can alter the extent and schedule of the aid in response to the operator's changing needs throughout the course of a task. In doing so, they can minimize many unintended consequences of indiscriminate and persistent aiding (Carmody and Gluckman 1993; Parasuraman et al. 1993; Endsley and Kiris 1995). Adaptive systems invoked by psychophysiological measures of workload can preempt performance declines and do not depend on the operators to be aware of their need for aid or make explicit requests for aid. Such systems rely on a

workload model based on psychophysiological measures to drive the schedule of adaptive aid. The model encodes when an excessive workload level is reached so that adaptive aid can be provided. To do so, it must be sensitive to and able to differentiate meaningful workload levels (e.g., levels that relate to different levels of performance). It should also provide diagnostic information about the measures associated with the need for aid to help design appropriate aiding behaviors. For example, in knowing that an adaptive aid had been triggered by an unusually high ocular activity, the system can work to relieve the visual demand.

This requirement for transparency is often cited among the limitations of using artificial neural networks (ANN), support vector machines (SVM), and other machine learning algorithms that include non-linear techniques for workload modeling (Mittelstadt et al. 2016). While some of these can result in models with very high accuracy rates (e.g., Wilson and Russell 2003a, b; Yeo et al. 2009; Baldwin and Penaranda 2012), not all are suitable for use in all adaptive systems or real-time applications. Some machine learning algorithms have limited diagnosticity because their features, rules, and criteria are less “transparent” and are often inscrutable. It is typically unclear how the multiple inputs are selected and combined to predict the target outcome. This inscrutability can have serious implications on the user’s trust in adaptive systems (Knight 2017; Ribeiro et al. 2016).

In addition, the model should include a variety of measures to detect a range of workload responses capturing the operator’s workload, as well as accommodate large inter-individual variability in psychophysiology. A recent review of more than 20 workload assessment algorithms developed for use in several environments revealed that none of the algorithms reviewed fulfill all these requirements, and most do not generalize across individuals (Heard et al. 2018). Table 1 includes

some of the commonly encountered psychophysiological workload measures.

5 An individualized workload model

We sought to develop a workload model (Teo et al. 2016) that met the following criteria:

- It must reliably distinguish between low and high workload and must identify when high workload is reached in real time.
- *Justification:* For the adaptive aid to be useful, the system needs to identify low and high levels of operator workload and respond appropriately. Aid that does not match the workload level is not as helpful (see Teo et al. 2018).
- It must customize the set of workload measures to the individual to optimize sensitivity on an individual basis.
- *Justification:* The large inter-individual variability in physiological responses to workload precludes the use of a common set of psychophysiological workload measures and target criteria for adaptation across all individuals. Having a workload model that can specify the best set of measures for each individual will improve the system’s ability to identify excessive workload for that person.
- It must incorporate multiple measures that assess a range of psychophysiological workload responses to capture the complete workload state of the individual and increase diagnosticity.

Table 1 Common psychophysiological workload measures

Psychophysiological measures	Response of measure to high workload
Heart rate	Increases (Wilson and Eggemeier 1991)
Heart rate variability (HRV)	Decreases (Mulder et al. 2004)
Pupil diameter	Increases (Casali and Wierwille 1983; May et al. 1990; Backs and Walrathf 1992)
Eye fixation duration	Increases (Callan 1998); Decreases (Schulz et al. 2010)
No. of eye fixations	Increases (Van Orden et al. 2001)
Theta waves (from EEG)	Increases (Hankins and Wilson 1998)
Alpha waves (from EEG)	Decreases (Hankins and Wilson 1998)
Beta waves (from EEG)	Increases (Kurimori and Kakizaki 1995)
Oxygen saturation (rSO ₂)	Increases (Sassaroli et al. 2008)
Electrodermal activity	Increases (De Waard 1996)
Cerebral bloodflow velocity (CBFV)	Increases (Warm and Parasuraman 2007)

Adapted from Meister (2014)

- *Justification:* Including multiple measures that are differentially sensitive to various cognitive processes activated by the tasking should yield richer information about the source or nature of the workload for the individual, which can be used to improve the quality of the adaptive aid.

5.1 Identifying the onset of high workload

First, a robust workload manipulation which does not produce any taskload-workload dissociations (Yeh and Wickens 1988) is required in order to capture the pattern of psychophysiological responses associated with low and high workload through manipulating taskload. Data from a previous study, Study 1 (Abich et al. 2013), were used to develop the workload model (i.e., Study 1 data served as the training dataset). Study 1 manipulated low and high workload with single vs. dual tasking, which, from the performance results and subjective workload measures, was shown to consistently yield the needed workload manipulation. The scenarios from Study 1 can be found in Table 2.

The basis of the workload model lies in comparing two sets of change scores, calculated from a workload algorithm based on multiple psychophysiological indicators. The first set comprised the change scores on the algorithm value between conditions known to elicit low and high workload, i.e., the “baseline difference scores.” To determine the workload level induced by a new condition (so that adaptive aid can be rendered if workload is high), a second set of difference scores is formed. This second set consisted of the psychophysiological change between the original low workload condition and the new condition that elicits an unknown level of workload, i.e., the “test difference scores.” If the set of “baseline difference scores” and “test difference scores” are similar in magnitude and direction (i.e., they match), then the new condition is considered to have induced a workload level comparable with the original high workload condition. To illustrate, Fig. 1 depicts three hypothetical sets of difference scores. A

Fig. 1 Comparing difference scores to determine workload level elicited by a new task (y-axis shows hypothetical values of psychophysiological measures)

comparison of the sets of *Baseline difference scores* to the *Test difference scores 1* would show that the psychophysiological changes producing the difference scores are similar in magnitude and direction, indicating that New Task 1 was eliciting a similarly high workload as the dual task since the sets of *Baseline difference scores* and *Test difference scores 1* match. However, the sets of *Baseline difference scores* and *Test difference scores 2* would not match, indicating that New Task 2 did not induce high workload (see Fig. 1).

By pairing the various scenarios in Study 1, we obtained multiple sets of difference scores with which we could develop the workload model (see Table 3).

5.2 Individualization

Measures sensitive to workload changes for the individual are those on which the individual shows a large change going from a low to high workload condition. For instance, for one individual, the workload measure of “number of eye fixations” may show a large change between a low and high workload condition indicating that the workload increase could be related to heavier visual demand. However, for a second individual, the measure showing a large change could be HRV instead. Such information contributes to diagnosticity

Table 2 Manipulation of workload levels in Study 1

Scenario	No. of tasks/taskload	Workload manipulated
Scenario 1 (S1)	Single task: CD* task only	Low workload
Scenario 2 (S2)	Dual tasks: CD task at varying event rate and TD** task	High workload
Scenario 3 (S3)	Single task: TD task only	Low workload
Scenario 4 (S4)	Dual tasks: CD task and TD task at varying event rate	High workload

*CD: Change Detection task. The change detection (CD) task involved detecting changes in the icons overlaid on a map. Participants assumed the role of a soldier on a mission with an unmanned ground vehicle (UGV) robot and were informed that the icons represented enemy assets and activities. They reported any appearance, disappearance, or movement of icons by clicking on the corresponding -labeled buttons. **TD: Threat Detection task. The threat detection (TD) task required detecting threats (which the participants were first trained to identify) among other characters in a video feed from an “unmanned ground vehicle” (participant’s “robot asset”) moving through a street lined with the various characters. Participants reported threat characters with mouse clicks on the threats within the video feed.

Table 3 Sets from Study 1 scenarios

	Scenario 1: single task (CD only)	Scenario 2: dual task (CD + TD)	Scenario 3: single task (TD only)	Scenario 4: dual task (CD + TD)
Set #1: Baseline diff. scores	✓	✓		
Set #1: Test diff. scores	✓			✓
Set #2: Baseline diff. scores		✓	✓	
Set #2: Test diff. scores			✓	✓

CD: Change Detection task; TD: Threat Detection task

as it suggests that an aid with visual processing may benefit the first individual more than the second.

There are different ways of specifying algorithms to capture with a single index the workload responses that are diagnostic for the individual. One approach is to weight responses according to their sensitivity. Another is to select only those responses that show large changes to increases in taskload. For some algorithms, the definition of “a large change” is a change of 0.5 standard deviations (SD) or more between the low and high workload conditions. Measures that show this type of change between conditions are designated as the individual’s set of workload “markers.” For other algorithms, the individual’s workload “markers” are the top few measures that register the largest changes between a low and high workload. In both approaches, only the measures most sensitive for the individual (i.e., his/her workload “markers”) will be used to compute the workload index. This approach allows the algorithms to be expressed as “rules” that are generalizable across different individuals and populations, while still accommodating inter-individual differences in psychophysiology. They are specific to the individual and yet generalizable at the same time.

5.3 Combining multiple measures

We examined a total of 26 psychophysiological measures,¹ including many listed in Table 1, i.e., EEG, ECG, CBFV,

¹ The 26 psychophysiological measures monitored: 1) Alpha Frontal mean, 2) Alpha Parietal mean, 3) Alpha Occipital mean, 4) Beta Frontal mean, 5) Beta Parietal mean, 6) Beta Occipital mean, 7) Theta Frontal mean, 8) Theta Parietal mean, 9) Theta Occipital mean, 10) Heart rate variability mean, 11) Inter-beat Interval mean, 12) rSO₂ right mean, 13) rSO₂ right SD, 14) rSO₂ left mean, 15) rSO₂ left SD, 16) CBFV right mean, 17) CBFV right SD, 18) CBFV left mean, 19) CBFV left SD, 20) Fixation duration mean, 21) Fixation duration SD, 22) No. fixations, 23) Pupil diameter mean, 24) Pupil diameter SD, 25) ICA mean and, 26) ICA SD.

rSO₂, eye fixation duration, Index of Cognitive Activity (ICA) (Marshall 2002), and pupil diameter measures as potential workload markers. These measures were selected for their sensitivity to the workload induced by the tasks used as shown in previous studies (Abich et al. 2013; Matthews et al. 2015b).

Scores from the multiple psychophysiological measures are first standardized to remove scale differences across the measures. Standardization of all scores allows a single workload index to be computed by combining multiple psychophysiological measures. The sets of “baseline difference scores” and “test difference scores” are obtained by combining the standardized values of multiple measures.

Algorithms that quantified the similarity in psychophysiological changes across multiple measures, or sets of differences scores, were generated. The algorithms combined the values on these measures in different ways to yield a single workload index. To be implemented in the adaptive system, a cutoff score is then required to specify index values that indicate when high workload is reached.

5.4 Formulation of algorithms for the workload model

Various algorithms were devised to compute the workload index that reflected individual variability in psychophysiological responses to workload and incorporated multiple measures. These either weight responses according to their sensitivity or focused only on responses that show large changes to increases in taskload. The workload index under each algorithm quantified the similarity of these responses between the high workload-inducing dual task and the new task condition by comparing the baseline difference scores and test difference scores.

In addition to the two sets formed from Study 1’s scenarios, two more sets were formed to compare algorithm performance to at-chance accuracy. These sets included the use of a separate data set comprising values drawn randomly from a theoretical normal distribution (all psychophysiological data have been standardized at this point). The random data were used as the data from a new unknown condition (see Table 4).

Unlike sets #1 and #2, in which the baseline and test difference scores are expected to reflect similar patterns of psychophysiological changes, the baseline and test difference scores in both sets #3 and #4 were not expected to match. Poor-performing algorithms may not yield workload indices that concur with these expectations.

5.4.1 Algorithm 1

In Algorithm 1, a workload index to quantify the similarity of psychophysiological changes is computed as the proportion of markers that show the same large change in workload response in the new condition. Computing the index as a

Table 4 Sets from Study 1 scenarios and random data

	Scenario 1: single task (CD only)	Scenario 2: dual task (CD + TD)	Scenario 3: single task (TD only)	Random dataset
Set #3: Baseline diff. scores	✓	✓		
Set #3: Test diff. scores	✓			✓
Set #4: Baseline diff. scores		✓	✓	
Set #4: Test diff. scores			✓	✓

CD: Change Detection task; TD: Threat Detection task

proportion ensured a fixed range of values under this algorithm from 0 to 1. The more similar the workload response elicited by the new unknown condition is to that induced by the original high workload condition, the more the workload index would approach the value of 1 since in both instances, the same measures would have registered similarly large changes from the low workload condition. Examples of how the workload index would be computed under Algorithm 1 are as follows:

Workload index under Algorithm 1

$$= \frac{\text{Markers observed in the Baseline and Test Difference scores}}{\text{Markers observed in the Baseline Difference scores.}}$$

Conceptually, the workload index under Algorithm 1

$$= \frac{\text{Wkld response common to both the new\&high wkld condition}}{\text{Wkld response in the high wkld condition}}$$

Example 1 Workload index for an individual with 4 markers (i.e., HRV mean, interbeat-interval (IBI) mean, theta frontal mean spectral power density (SPD), left mean CBFV) when the new condition induces a *similar* workload level to that of the known high workload condition. The workload index value is higher, at 0.75:

Workload index under Algorithm 1 for Example 1

$$= \frac{\text{HRV, IBI, Theta frontal SPD}}{\text{HRV, IBI, Theta frontal SPD, Left CBFV}} = \frac{3}{4} = 0.75$$

Example 2 Workload index for an individual with 4 markers (i.e., HRV mean, IBI mean, theta frontal mean SPD, left mean CBFV) when the new condition induces a *different* workload level to that of the known high workload condition. The workload index value is low at 0.25:

Workload index under Algorithm 1 for Example 1

$$= \frac{\text{HRV}}{\text{HRV, IBI, Theta frontal SPD, Left CBFV}} = \frac{1}{4} = 0.25$$

Workload index under Algorithm 1 For this algorithm, similarity in psychophysiological response was indicated by the proportion of the individual's "markers" that had registered the response indicating large workload change in both sets of difference scores. "Test difference scores" computed from single-dual task differences (i.e., Set #1 and Set #2) would

Table 5 Algorithm 1: workload index means and std. dev.

Set	Baseline diff. scores	Test diff. scores	Similarity	Algorithm 1* workload index: M (SD)
#1	Scen. 1 and Scen. 2: single and dual task	Scen. 1 and Scen. 4: single and dual task	High	0.55 (0.24) [†]
#2	Scen. 3 and Scen. 2: single and dual task	Scen. 3 and Scen. 4: single and dual task	High	0.54 (0.25) [†]
#3	Scen. 1 and Scen. 2: single and dual task	Scen. 1 and Random: single and random	Low	0.29 (0.13) [†]
#4	Scen. 3 and Scen. 2: single and dual task	Scen. 3 and Random: single and random	Low	0.32 (0.14) [†]

*Algorithm 1: larger values indicate greater similarity between the set of baseline and test difference scores

[†] Markers were defined as the measures that registered a change of at least 0.5 SD between low and high workload-inducing conditions

Table 6 Algorithm 2: workload index means and Std. dev.

Set	Baseline diff. scores	Test diff. scores	Similarity	Algorithm 2* workload index: M (SD)
#1	Scen. 1 and Scen. 2: single and dual task	Scen. 1 and Scen. 4: single and dual task	High	4.72 (2.04)
#2	Scen. 3 and Scen. 2: single and dual task	Scen. 3 and Scen. 4: single and dual task	High	4.98 (2.18)
#3	Scen. 1 and Scen. 2: single and dual task	Scen. 1 and Random: single and random	Low	7.62 (1.77)
#4	Scen. 3 and Scen. 2: single and dual task	Scen. 3 and Random: single and random	Low	7.52 (1.55)

*Algorithm 2: smaller values indicate greater similarity (i.e., smaller distance) between the set of baseline and test difference scores

match that in “Baseline difference scores.” “Test difference scores” computed with random data (i.e., Set #3 and Set #4) would not be expected to match the “baseline difference scores” (see Table 5).

The effect size (Cohen’s d) between sets #1 and #3 (using the same baseline differences scores) is 1.347 while that between sets #2 and #4 is 1.086, indicating that the Algorithm 1 was well able to distinguish between data from an actual high workload condition from random data. These values indicate large effect sizes according to Cohen’s (1988) criteria.

5.4.2 Algorithm 2

The workload index quantifying the similarity of the two sets of change scores for Algorithm 2 is the Euclidean distance between them, with smaller distance scores denoting greater similarity. The index can be individualized by only including the individual’s own set of markers in the distance computation. Whereas Algorithm 1 seeks to select a subset of response measures from those available for each individual, Algorithm 2 incorporates information from all responses, even those that were relatively insensitive for an individual. The following shows the computation of Algorithm 2 workload index:

$$\text{Workload index, } d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Table 7 Algorithm 3: workload index means and std. dev.

Set	Baseline diff. scores	Test diff. scores	Similarity	Algorithm 3* workload index: M (SD)
#1	Scen. 1 and Scen. 2: single and dual task	Scen. 1 and scen. 4: single and dual task	High	20.11 (6.48)
#2	Scen. 3 and Scen. 2: single and dual task	Scen. 3 and Scen. 4: single and dual task	High	19.48 (5.92)
#3	Scen. 1 and Scen. 2: single and dual task	Scen. 1 and Random: single and random	Low	14.36 (3.91)
#4	Scen. 3 and Scen. 2: single and dual task	Scen. 3 and Random: single and random	Low	14.63 (3.98)

*Algorithm 3: larger values indicate greater similarity between the set of baseline and test difference scores

where x is the “baseline difference score” for a psychophysiological workload measure, y is the “test difference score” for a psychophysiological workload measure, and i is a psychophysiological metric (e.g., $i = 1$ denotes heart rate variability or HRV, $i = 2$ denotes interbeat interval or IBI, etc.)

Workload index under Algorithm 2 Similarity in psychophysiological response, for this algorithm, was reflected as the Euclidean distance between the sets of all difference scores. As before, “test difference scores” computed from single-dual task differences (i.e., Set #1 and Set #2) should match and be “nearer” (i.e., smaller distance) to the “baseline difference scores”, while the distance between “test difference scores” computed with random data (i.e., Set #3 and Set #4) and “baseline difference scores” should be larger (Table 6).

Algorithm 2 was also well able to distinguish between data from an actual high workload condition from random data. The effect size (Cohen’s d) between Sets #1 and #3 is 1.519 while that between Sets #2 and #4 is 1.343.

5.4.3 Algorithm 3

Similarity of the psychophysiological change is quantified as the number of workload measures that have signs that match for the “baseline difference score” and “test difference score.” Matched signs indicate that the direction of the change

Table 8 Algorithm 4: workload index means and std. dev.

Set	Baseline diff. scores	Test diff. scores	Similarity	Algorithm 4* workload index: M (SD)
#1	Scen. 1 and Scen. 2: single and dual task	Scen. 1 and Scen. 4: single and dual task	High	0.95 (0.77) [†] 0.99 (0.78) [‡]
#2	Scen. 3 and Scen. 2: single and dual task	Scen. 3 and Scen. 4: single and dual task	High	0.97 (0.77) [†] 0.98 (0.75) [‡]
#3	Scen. 1 and Scen. 2: single and dual task	Scen. 1 and Random: single and random	Low	0.26 (0.48) [†] 0.57 (0.69) [‡]
#4	Scen. 3 and Scen. 2: single and dual task	Scen. 3 and Random: single and random	Low	0.16 (0.44) [†] 0.50 (0.66) [‡]

*Algorithm 4: larger values indicate greater similarity between the set of baseline and test difference scores

[†] Test difference scores had to be at least the same magnitude as baseline difference scores

[‡] Test difference scores only had to be at least 0.5 that of the baseline difference scores

between conditions is similar. The more similar the change in psychophysiological workload response is between the set of “baseline difference scores” and “test difference scores,” the greater the number of matched signs compared with that which will occur by chance. The workload index is computed as the number of the measures for which the signs for the “baseline difference score” and “test difference score” matched. Since the number of psychophysiological measures used is 26, the range of values for the workload index under Algorithm 3 is 0 to 26. Like Algorithm 2, this algorithm utilizes information from all responses, but on a categorical rather than a continuous basis.

Workload index under Algorithm 3 For this algorithm, the more similar the changes in psychophysiological responses were, the greater the number of matched signs between the sets of difference scores (Table 7). Index values for Sets #1 and #2 indicate greater match between the “baseline difference scores” and “test difference scores” while values for Sets #3 and #4, which involve random data, show poorer match.

Between Sets #1 and #3, the effect size (Cohen’s d) is 1.074 while that between Sets #2 and #4 is 0.962, indicating that the Algorithm 3 was able to distinguish between data from an

actual high workload condition from random data. However, d s were somewhat smaller than those for Algorithms 1 and 2.

5.4.4 Algorithm 4

For this algorithm, the top two psychophysiological “markers” for the individual are first identified from the “baseline difference scores” (i.e., the two measures which showed the largest difference between the original low and high workload-inducing conditions). Since only the top two markers are included, the workload index would range from 0 to 2, with index values approaching 2 if the “baseline differences scores” and “test difference scores” are similar in magnitude and direction. Another derivative of this algorithm requires the change in the “test difference scores” to be in the same direction but only at least half the magnitude of that in the “baseline difference scores.” This algorithm reverts to selection of key markers on an individual basis, focusing especially on those most responsive to workload.

Workload index under Algorithm 4 Under this algorithm, similarity in psychophysiological response was the extent to which the individual’s top 2 “markers” showed the greatest

Table 9 Study 1 scenarios yielding various sets for the mock-up

Set	Baseline difference scores	Rolling test difference scores	Expected similarity between baseline and rolling test difference scores
#5	S1 and S2: single and dual tasks	S1 and S2: single and dual tasks	Very similar as both are differences between the same single and dual tasks.
#6*	S1 and S2: single and dual tasks	S1 and S4: single and dual tasks	Similar as both are differences between single and dual tasks.
#7	S1 and S2: single and dual tasks	S1 and S3: single and single tasks	Dissimilar as the baseline diff. scores are from single and dual tasks, but the rolling test diff. scores are the changes between two single tasks.
#8	S1 and S2: single and dual tasks	S1 and S1: single and single tasks	Very dissimilar as the baseline diff. scores are from single and dual tasks, but the rolling test diff. scores are not expected to show any differences.

*Set #6 = Set #1 as their baseline and test difference scores are created from the same conditions

Fig. 2 Mock-up of adaptive system with Algorithm 1

change in both sets of difference scores. Very similar sets of difference scores (i.e., Sets #1 and #2) should yield values close to 2 (Table 8).

Although Algorithm 4 was also able to distinguish between data from an actual high workload condition from random data, the effect sizes were somewhat lower than for the other algorithms. Cohen's d between Sets #1 and #3 (using the same

Fig. 4 Mock-up of adaptive system with Algorithm 2a (smaller values indicate greater similarity)

baseline differences scores) ranged from 0.57 to 1.075 while that between Sets #2 and #4 ranged from 0.679 to 1.292.

All four algorithms seemed able to distinguish the psychophysiological changes resulting from high workload from random data. However, both Algorithms 3 and 4 produce discrete values that may limit their use. Algorithm 3 defines similarity only in terms of the direction of the psychophysiological

Fig. 3 Mock-up of adaptive system with Algorithm 2 (smaller values indicate greater similarity)

Fig. 5 Mock-up of adaptive system with Algorithm 2b (smaller values indicate greater similarity)

Table 10 Signal detection outcomes from the mock-up

	Aid should be triggered (i.e., Set #8 and Set #9)	Aid should not be triggered (i.e., Set #10 and Set #11)
Aid was triggered	Hit	False alarm (FA)
Aid was not triggered	Miss	Correct rejection (CR)

change, without a criterion for change magnitude. Closer examination of the workload index values from Algorithm 4 showed that even when the “baseline” and “test” difference scores were supposed to match (i.e., both from single and dual task conditions), most of the participants had index values that did not reflect this similarity. Additionally, the range of values under Algorithm 4 is limited as it is equal to the number of “markers” considered to be “top markers”. Increasing the range of index values may result in including “markers” that are not as sensitive for the individual. The effect sizes for Algorithms 3 and 4 were also lower than those for Algorithms 1 and 2. For these reasons, only Algorithms 1 and 2 were selected for further analyses and evaluation.

6 Evaluation of workload models

The workload models generated with Algorithms 1 and 2 were further subjected to a mock-up of an adaptive aiding system with Study 1 data to help select threshold values, and the evaluation of the sensitivity of those threshold values.

6.1 Mock-up of the workload model in an adaptive aiding system

In the mock-up, 2-min blocks of data were streamed into the system as “live” samples of data (i.e., 2-min “rolling” window) every 30 seconds, such that consecutive samples have a 1.5-min overlap of data. In place of a static set of “test difference scores,” there is a set of “rolling test difference scores” which is constantly updated every 30 s to reflect the individual’s psychophysiological responses during the new condition inducing an unknown level of workload. With Study 1

scenarios and data, four more sets of “baseline difference scores” and “rolling test difference scores” were generated to compare index values for conditions that matched to differing extents (see Table 9).

From the mock-up, a potential threshold or cutoff score (i.e., solid horizontal line in the figures below) was determined. This is the workload index value that differentiated similar sets of “baseline” and “rolling test” differences scores from dissimilar sets.

The mock-up with Algorithm 1 resulted in the expected order of similarity across all samples. The most similar sets of “baseline” and “rolling test” difference scores (i.e., Set #5) had the high index values, followed by the next most similar sets (i.e., Set #6), then by Set #7, followed by Set #8 which had the lowest index values denoting lowest similarity. A possible cutoff score for this Algorithm was 0.62 (see Fig. 2).

With Algorithm 2, the expected order of sets was not observed. Set #6 which comprised difference scores that should be more closely matched than that of Set #7 had index values that indicated lower similarity instead. In addition, the potential cutoff score of 7.2 may still result in misclassifications. Due to this, Algorithm 2 was eliminated from further consideration (see Fig. 3).

This result prompted two derivatives of Algorithm 2 to be formulated. Algorithm 2a included only the top 5 measures that showed the greatest magnitude of psychophysiological change between single and dual task, while Algorithm 2b included the top 10 measures in the workload index computation. For both Algorithm 2a and 2b, the expected order of set similarity was observed although the distinction between similar sets (i.e., Set #5 and Set #6), and dissimilar sets (i.e., Set #7 and Set #8) was not distinct enough for a cutoff score to be established in both of these new algorithms (see Figs. 4 and 5).

Table 11 Study 1 sets with various algorithms at proposed thresholds

	Algorithm 1 (cutoff at 0.62)	Algorithm 2a (cutoff at 3.4)	Algorithm 2b (cutoff at 4.4)
Set #5: S1 and S2 with S1 and S2			
Hit (%)	88.6	57.8	92.0
Miss (%)	11.4	42.2	8.1
Set #8: S1 and S2 with S1 and S1			
False alarm (%)	36.2	9.5	68.8
Correct rejects (%)	63.8	90.5	36.2
Sensitivity, d'	1.743	1.503	1.049

*S1 and S3 were single task conditions; S2 and S4 were dual task conditions

Table 12 Manipulation of workload levels in Study 2

Mission	No. of tasks/taskload	Workload manipulated
Mission 1 (S1)	Single task: CD* task only	Low workload
Mission 2 (S2)	Dual tasks: CD task and MT** task at low event rate, or MTlow (5 SA prompts/3 min [†])	High workload
Mission 3 (S3)	Dual tasks: CD task and MT** task at medium event rate, or MTmed (7 SA prompts/3 min [†])	Higher workload
Mission 4 (S4)	Dual tasks: CD task and MT** task at high event rate, or MThigh (9 SA prompts/3 min [†])	Highest workload

*CD: Change Detection task; **MT: Monitoring Task;

[†] The event rate for the Monitoring Task (MT) were set in accordance with that in Reinerman-Jones et al. (2010)

6.2 Sensitivity of workload models and thresholds

The adaptive system, with the appropriate cutoff score, should detect when participants are in conditions that induce high workload (i.e., dual task in this case). A signal detection paradigm can be applied to evaluate the sensitivity of the system. When the system correctly identifies the high workload-inducing condition, then the system would have made a “Hit.” “Misses” are when the system fails to identify the onset of high workload. “False Alarms” are instances when the system triggers aid during a low workload-inducing condition, and

“Correct Rejections” are when no aid is provided during low workload-inducing condition (Table 10).

The optimal cutoff score would show high sensitivity (d'), a signal detection measure, as it will maximize “Hits” and “Correct Rejections” while minimizing “Misses” and “False Alarms (FAs).” Sensitivity was computed as follows:

Sensitivity or d'

$$= Z(\text{proportion of “hits”}) - Z(\text{proportion of “FAs”})$$

With data from Study 1, hit, miss, false alarm, and correct rejection rates were computed for Set #5 (most similar psychophysiological response) and Set #8 (most dissimilar psychophysiological response) using the most plausible thresholds of Algorithms, 1, 2a, and 2b. Results favored Algorithm 1 at the 0.62 cutoff (Table 11).

Table 13 Study 2 scenarios yielding sets of test data with alternative workload manipulations

	Mission 1: single task (CD only)	Mission 2: dual task (CD + MTlow)	Mission 3: dual task (CD + MTmed)	Mission 4: dual task (CD + MThigh)
Set #9: Baselin- e diff. scores	✓	✓		
Set #9: Test diff. scores	✓			✓
Set #10: Baselin- e diff. scores		✓	✓	
Set #10: Test diff. scores			✓	✓
Set #11: Baselin- e diff. scores	✓	✓		
Set #11: Test diff. scores		✓		✓

CD: Change Detection task; MT: Monitoring Task

7 Testing the workload models

7.1 Robustness of models to different workload manipulations

The workload model under Algorithm 1 was next tested on a separate sample of participants. We also wanted to see if the workload model was able to identify high workload from dual tasking that was elicited by a slightly different set of tasks. In addition, we explored the use of event rate to manipulate workload.

Study 2 used the change detection (CD) task and a monitoring task (MT) to create single and dual tasking² to elicit the

² Study 2 utilized the same simulation platform as Study 1, and also had participants assume the role of a Soldier on a mission with an unmanned ground vehicle (UGV) robot. Study 2 used the same change detection (CD) task, but instead of the threat detection (TD) task, a monitoring task (MT) was paired with the CD task to create dual tasking. The monitoring task required participants to answer a series of situational awareness (SA) prompts as they monitored the same video feed used in the threat detection task. Participants monitored the feed for pre-specified targets such as vehicles, men and women. SA prompts asked about the different targets that they observed since the most recent turn in the route, e.g., “How many women did the robot pass since the last turn?”

Table 14 Testing robustness of the workload model

Set	Baseline diff. scores	Test diff. scores	What is tested
Set #9	Workload change between single and dual tasking	Workload change between single and dual tasking	Model performance on same workload manipulation as Study 1 (i.e., single-dual tasking), but with different tasks.
Set #10	Workload change between tasks differing on event rate	Workload change between tasks differing on event rate	Model performance on event rate as workload manipulation (i.e., event rate manipulation).
Set #11	Workload change between single and dual tasking	Workload change between tasks differing on event rate	Model performance on workload elicited from different workload manipulations (i.e., mixed manipulation)

low and high workload conditions. There were 3 levels of the monitoring task that differed on event rate. The scenarios in Study 2 were as follows (see Table 12):

The scenarios were combined to create the following sets of baseline and test difference scores (see Table 13):

These sets tested the workload model in following ways in Table 14.

The workload index based on Algorithm 1 was computed with these sets using data from the Study 2 participants (see Table 15).

Comparing the values from these sets to values from Sets #1 to #4 (i.e., Table 5), the workload model generalized to a different sample, and to slightly different tasks; so long as the same single-dual tasking workload manipulation was used. The model performed less well with the event rate manipulation of workload or with mixed manipulation. This is probably because the psychophysiological responses are different for different workload manipulations (Matthews et al. 2015b).

7.2 Distribution of workload index values

The distribution and range of workload index values obtained with Algorithm 1 showed that it was able to sufficiently identify workload changes from single-dual tasking. The distribution generated with data where both the “baseline” and “test” difference scores were from single-dual task manipulations (i.e., graphs for sets #1, #2, #5, or the filled-in circles) and

were distinct from that which involved random data (i.e., graphs for sets #3 and #4 or the open circles). Furthermore, 50% of the workload index values from matched conditions (i.e., both “baseline” and “test” difference scores were changes between single and dual task conditions) were at least 0.57 (solid arrow), while 90% of the values from unmatched conditions involving random data were below 0.50 (dotted arrow) (Fig. 6).

Such distributions indicate that the workload index under Algorithm 1 would be sufficiently able to identify when high workload is reached. In a separate study (Teo et al. 2018), this workload model (i.e., based on Algorithm 1 with the cutoff of 0.62) was implemented in an adaptive aiding system that was driven by workload-related psychophysiological changes. Results of that study indicated that compared with those whose aid was not adaptive, those who received adaptive aid showed greater performance improvements.

8 Future work and conclusions

An individualized workload model was developed to drive adaptive aiding. The methodology used enabled various psychophysiological measures with different scale properties and sampling rates to be combined into a single workload index, which was formulated to accommodate the inter-individual variability in psychophysiological responses that is a major

Table 15 Algorithm 1: workload index values with alternative workload manipulations

Set	Baseline diff. scores	Test diff. scores	Wkld manipulation	Algorithm 1* workload index: M (SD)
#9	Msn. 1 and Msn. 2: single and DLow	Msn. 1 and Msn.4: single and DHigh	Single-dual	0.51 (0.22) [†]
#10	Msn. 2 and Msn. 3: DLow and DMed	Msn. 3 and Msn. 4: DMed and DHigh	Event rate	0.12 (0.14) [†]
#11	Msn. 1 and Msn. 2: single and DLow	Msn. 2 and Msn. 4: DLow and DHigh	Mixed	0.10 (0.13) [†]

Msn, mission; DLow, dual task at low level; DMed, dual task at medium level; DHigh, dual task at high level

*Algorithm 1: larger values indicate greater similarity between the set of baseline and test difference scores

[†] Markers were defined as the measures that registered a change of at least 0.5 SD between low and high workload-inducing conditions

Fig. 6 Distribution of workload index values for Algorithm 1 (Sets #1 through #4 from Study 1 data, Set#5 from Study 1 data)

challenge in workload modeling. Comparisons of workload index values generated from random data provided a means to evaluate algorithm performance against chance level, while the sensitivity analysis provided a way to assess the selected threshold level. Generalizability of the workload model was assessed with alternative workload manipulations. This methodology resulted in a viable model that incorporated multiple workload measures and accommodated individual variability in psychophysiological workload responses. The model was used with some success in an adaptive aiding system (Teo et al. 2018). Nevertheless, follow-on work is needed to improve the generalizability of the model to other workload manipulations as well as model sensitivity and specificity. It is also important to develop adaptive aiding that is robust when task demands change dynamically and unpredictably.

The present work touches on several issues concerning workload and system design. For one, the relationship between workload and performance is hardly a straightforward one and can be difficult to characterize. Operators' behavioral or compensatory strategies can result in different workload-performance relationships (i.e., associations, dissociations, insensitivities, linear, non-linear) (Yeh and Wickens 1988; Hancock and Matthews 2019). Secondly, different psychophysiological measures operate at different intrinsic frequencies which can affect the temporal resolution of workload characterization. For example, changes in EEG can be measured in milliseconds while changes in heart rate are detected in seconds (Hancock and Matthews 2019). Designers of system aiding behaviors must also consider the effects of the aid and other task changes since operator workload is susceptible to hysteresis effects (Cox-Fuenzalida 2007; Hancock and Matthews 2019).

A workload model that provides insight into individual operators' workload responses during various tasks offers a valuable opportunity for designing all manner of individualized technological aids and interventions. Although there is

much work still to be accomplished towards this end, the present work provides some impetus for the continuation of effort towards this vision.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Abich J IV, Reinerman-Jones L, Taylor GS (2013) Investigating workload measures for adaptive training systems. In: Proceedings of the human factors and ergonomics society annual meeting. SAGE Publications Sage CA, Los Angeles, CA, pp 2091–2095
- Ang DSC, Lang CC (2008) The prognostic value of the ECG in hypertension: where are we now? *J Hum Hypertens* 22:460–467
- Backs RW, Walrathf LC (1992) Eye movement and pupillary response indices of mental workload during visual. *Appl Ergon* 23:243–254
- Bailey NR, Scerbo MW, Freeman FG, Mikulka PJ, Scott LA (2006) Comparison of a brain-based adaptive system and a manual adaptable system for invoking automation. *Hum Factors* 48:693–709
- Baldwin CL, Penaranda BN (2012) Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *NeuroImage* 59:48–56. <https://doi.org/10.1016/j.neuroimage.2011.07.047>
- Bill A, Linder J (1976) Sympathetic control of cerebral blood flow in acute arterial hypertension. *Acta Physiol* 96:114–121
- Birren JE, Casperson RC, Botwinick J (1950) Age changes in pupil size. *J Gerontol* 5:216–221
- Brookhuis KA, Waard D (2001) Assessment of drivers' workload: performance, subjective and physiological indices, edited by: P.

- Hancock, P. Desmond. Lawrence Erlbaum Associates, Mahwah, NJ, USA
- Callan DJ (1998) Eye movement relationships to excessive performance error in aviation. In: Proceedings of the Human Factors and Ergonomics Society annual meeting. SAGE Publications Sage CA, Los Angeles, CA, pp 1132–1136
- Carmody MA, Gluckman JP (1993) Task specific effects of automation and automation failure on performance, workload and situational awareness. In: Proceedings of the Seventh International Symposium on Aviation Psychology. Citeseer, Princeton, pp 167–171
- Casali JG, Wierwille WW (1983) A comparison of rating scale, secondary-task, physiological, and primary-task workload estimation techniques in a simulated flight task emphasizing communications load. *Hum Factors* 25:623–641
- Cohen J (1988) The effect size index: d. *Statistical power analysis for the behavioral sciences* 2:284–288
- Cox-Fuenzalida L-E (2007) Effect of workload history on task performance. *Hum Factors* 49:277–291
- De Waard D (1996) The measurement of drivers' mental workload. Groningen University, Traffic Research Center Netherlands
- Dorneich MC, Rogers W, Whitlow SD, DeMers R (2016) Human performance risks and benefits of adaptive systems on the flight deck. *Int J Aviat Psychol* 26:15–35. <https://doi.org/10.1080/10508414.2016.1226834>
- Endsley MR, Kiris EO (1995) The out-of-the-loop performance problem and level of control in automation. *Hum Factors* 37:381–394
- Freeman FG, Mikulka PJ, Scerbo MW, Prinzel LJ, Cloutier K (2000) Evaluation of a psychophysiological controlled adaptive automation system, using performance on a tracking task. *Appl Psychophysiol Biofeedback* 25:103–115
- Hancock P, Caird JK (1993) Experimental evaluation of a model of mental workload. *Hum Factors* 35:413–429
- Hancock PA, Matthews G (2019) Workload and performance: associations, insensitivities, and dissociations. *Hum Factors* 61:374–392. <https://doi.org/10.1177/0018720818809590>
- Hancock PA, Warm J (1989) A dynamic model of stress and sustained attention. *Hum Factors* 31:519–537
- Hancock PA, Meshkati N, Robertson MM (1985) Physiological reflections of mental workload. *Aviat Space Environ Med* 56:1110–1114
- Hankins TC, Wilson GF (1998) A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviat Space Environ Med* 69:360–367
- Heard J, Harriott CE, Adams JA (2018) A survey of workload assessment algorithms. *IEEE Trans Hum-Mach Syst*:1–18. <https://doi.org/10.1109/THMS.2017.2782483>
- Hockey GRJ (2003) Operator functional state: the assessment and prediction of human performance degradation in complex tasks. IOS Press
- Johannes B, Gaillard AWK (2014) A methodology to compensate for individual differences in psychophysiological assessment. *Biol Psychol* 96:77–85. <https://doi.org/10.1016/j.biopsycho.2013.11.004>
- Kantowitz BH, Casper PA (2017) Human workload in aviation. *Human Error in Aviation*. Routledge, In, pp 123–153
- Karwowski W (2012) A review of human factors challenges of complex adaptive systems: discovering and understanding chaos in human performance. *Hum Factors J Hum Factors Ergon Soc* 54:983–995. <https://doi.org/10.1177/0018720812467459>
- Knight, W. (2017). There's a big problem with AI: Even its creators can't explain how it works. MIT Technology Review.
- Kurimori S, Kakizaki T (1995) Evaluation of work stress using psychological and physiological measures of mental activity in a paced calculating task. *Ind Health* 33:7–22
- Marshall SP (2002) The index of cognitive activity: measuring cognitive workload. In: Human factors and power plants, 2002. proceedings of the 2002 IEEE 7th conference on. IEEE, pp 7–7
- Matthews G, Reinerman-Jones L (2017) Workload assessment: how to diagnose workload issues and enhance performance. *Human Factors and Ergonomics Society*
- Matthews G, Reinerman-Jones L, Wohleber R et al (2015a) Workload is multidimensional, not unitary: what now? In: Schmorow DD, Fidopiastis CM (eds) *Foundations of augmented cognition*. Springer International Publishing, Cham, pp 44–55
- Matthews G, Reinerman-Jones LE, Barber DJ, Abich J IV (2015b) The psychometrics of mental workload: multiple measures are sensitive but divergent. *Hum Factors* 57:125–143
- May JG, Kennedy RS, Williams MC, Dunlap WP, Brannan JR (1990) Eye movement indices of mental workload. *Acta Psychol* 75:75–89
- Meister D (2014) *Human factors testing and evaluation*. Elsevier, Amsterdam
- Mittelstadt BD, Allo P, Taddeo M et al (2016) The ethics of algorithms: mapping the debate. *Big Data Soc* 3:205395171667967. <https://doi.org/10.1177/2053951716679679>
- Mulder LBJ, de Waard D, Brookhuis KA (2004) Estimating mental effort using heart rate and heart rate variability. *Handbook of human factors and ergonomics methods*. CRC Press, In, pp 227–236
- Parasuraman R, Molloy R, Singh IL (1993) Performance consequences of automation-induced 'complacency'. *Int J Aviat Psychol* 3:1–23
- Pierce TW, Watson TD, King JS, Kelly SP, Pribram KH (2003) Age differences in factor analysis of EEG. *Brain Topogr* 16:19–27
- Prinzel LJ, Freeman FG, Scerbo MW, Mikulka PJ, Pope AT (2000) A closed-loop system for examining psychophysiological measures for adaptive task allocation. *Int J Aviat Psychol* 10:393–410. https://doi.org/10.1207/S15327108IJAP1004_6
- Reinerman-Jones L, Barber D, Lackey S, Nicholson D (2010) Developing methods for utilizing physiological measures. *Adv Underst Hum Perform Neuroergonomics Hum Factors Des Spec Popul CRC Press*, Boca Raton
- Ribeiro MT, Singh S, Guestin C (2016) Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1135–1144
- Roscoe AH (1993) Heart rate as a psychophysiological measure for in-flight workload assessment. *Ergonomics* 36:1055–1062
- Sassaroli A, Zheng F, Hirshfield LM et al (2008) Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy. *J Innov Opt Health Sci* 1:227–237
- Schulz CM, Schneider E, Fritz L et al (2010) Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment. *Br J Anaesth* 106:44–50
- Teo G, Reinerman-Jones L, Matthews G et al (2016) Augmenting robot behaviors using physiological measures of workload state. In: Schmorow DD, Fidopiastis CM (eds) *Foundations of augmented cognition: neuroergonomics and operational neuroscience*. Springer International Publishing, Cham, pp 404–415
- Teo G, Reinerman-Jones L, Matthews G, Szalma J, Jentsch F, Hancock P (2018) Enhancing the effectiveness of human-robot teaming with a closed-loop system. *Appl Ergon* 67:91–103. <https://doi.org/10.1016/j.apergo.2017.07.007>
- Van Orden KF, Limbert W, Makeig S, Jung T-P (2001) Eye activity correlates of workload during a visuospatial memory task. *Hum Factors* 43:111–121
- Warm JS, Parasuraman R (2007) Cerebral hemodynamics and vigilance. *Neuroergonomics Brain Work*:146–158
- Wilson GF, Eggemeier FT (1991) Psychophysiological assessment of workload in multi-task environments. *Multi-Task Perform* 329360
- Wilson GF, O'Donnell RD (1988) Measurement of operator workload with the neuropsychological workload test battery. *Advances in Psychology*. Elsevier, In, pp 63–100
- Wilson GF, Russell CA (2003a) Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Hum Factors* 45:381–389

- Wilson GF, Russell CA (2003b) Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Hum Factors* 45:635–644
- Winn B, Whitaker D, Elliott DB, Phillips NJ (1994) Factors affecting light-adapted pupil size in normal human subjects. *Invest Ophthalmol Vis Sci* 35:1132–1137
- Yeh Y-Y, Wickens CD (1988) Dissociation of performance and subjective measures of workload. *Hum Factors J Hum Factors Ergon Soc* 30: 111–120. <https://doi.org/10.1177/001872088803000110>
- Yeo MVM, Li X, Shen K, Wilder-Smith EPV (2009) Can SVM be used for automatic EEG detection of drowsiness during car driving? *Saf Sci* 47:115–124. <https://doi.org/10.1016/j.ssci.2008.01.007>
- Young MS, Stanton NA (2002) Attention and automation: new perspectives on mental underload and performance. *Theor Issues Ergon Sci* 3:178–194. <https://doi.org/10.1080/14639220210123789>
- Young MS, Stanton NA (2005) Mental workload. *Handb Hum Factors Ergon Methods*:39–31
- Young MS, Brookhuis KA, Wickens CD, Hancock PA (2015) State of science: mental workload in ergonomics. *Ergonomics* 58:1–17. <https://doi.org/10.1080/00140139.2014.956151>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.