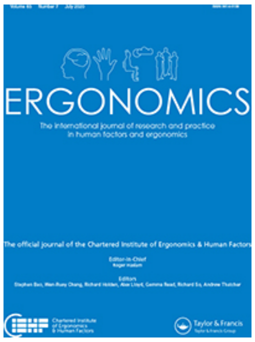

Ergonomics

ISSN: 0014-0139 (Print) 1366-5847 (Online) Journal homepage: https://www.tandfonline.com/loi/terg20

Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning

Yi Ding, Yaqin Cao, Vincent G. Duffy, Yi Wang & Xuefeng Zhang

To cite this article: Yi Ding, Yaqin Cao, Vincent G. Duffy, Yi Wang & Xuefeng Zhang (2020) Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning, Ergonomics, 63:7, 896-908, DOI: 10.1080/00140139.2020.1759699

To link to this article: https://doi.org/10.1080/00140139.2020.1759699

Published online: 07 May 2020.

ERGONOMICS
2020, VOL. 63, NO. 7, 896–908
https://doi.org/10.1080/00140139.2020.1759699

Taylor & Francis
Taylor & Francis Group

Measurement and identification of mental workload during simulated computer tasks with multimodal methods and machine learning

Yi Ding[a,b], Yaqin Cao[a,b], Vincent G. Duffy[b], Yi Wang[a] and Xuefeng Zhang[a]

[a]School of Management Engineering, Anhui Polytechnic University, Wuhu, P. R. China; [b]School of Industrial Engineering, Purdue University, West Lafayette, IN, USA

ABSTRACT
This study attempted to multimodally measure mental workload and validate indicators for estimating mental workload. A simulated computer work composed of mental arithmetic tasks with different levels of difficulty was designed and used in the experiment to measure physiological signals (heart rate, heart rate variability, electromyography, electrodermal activity, and respiration), subjective ratings of mental workload (the NASA Task Load Index), and task performance. The indices from electrodermal activity and respiration had a significant increment as task difficulty increased. There were no significant differences between the average heart rate and the low-frequency/high-frequency ratio among tasks. The classification of mental workload using combined indices as inputs showed that classification models combining physiological signals and task performance can reach satisfying accuracy at 96.4% and an accuracy of 78.3% when only using physiological indices as inputs. The present study also showed that ECG and EDA signals have good discriminating power for mental workload detection.

Practitioner summary: The methods used in this study could be applied to office workers, and the findings provide preliminary support and theoretical exploration for follow-up early mental workload detection systems, whose implementation in the real world could beneficially impact worker health and company efficiency.

Abbreviations: NASA-TLX: the national aeronautics and space administration-task load index; ECG: electrocardiographic; EDA: electrodermal activity; EEG: electroencephalogram; LDA: linear discriminant analysis; SVM: support vector machine; KNN: k-nearest neighbor; ANNs: artificial neural networks; EMG: electromyography; PPG: photoplethysmography; SD: standard deviation; BMI: body mass index; DSSQ: dundee stress state questionnaire; ANOVA: analysis of variance; SC: skin conductance; RMS: root mean square; AVHR: the average heart rate; HR: heart rate; LF/HF: the ratio between the low frequencies band and the high frequency band; PSD: power spectral density; MF: median frequency; HRV: heart rate variability; BPNN: backpropagation neural network

ARTICLE HISTORY
Received 9 September 2019
Accepted 13 April 2020

KEYWORDS
Mental workload; multimodal measures; psychophysiology; machine learning; workload classification

1. Introduction

The widespread use of computer-based systems has rendered human-computer systems more complex, and the emergence of the 'Internet of things,' 'Internet +,' and 'Intelligent Manufacturing' have increased cognitive-resource load demands. The demand for mental workload optimisation grows prominently. The current consensus is that both excessively high and excessively low levels of mental workload negatively influence work performance (Hancock and Matthews 2019; Van Acker et al. 2018). Mental workload is a subjective experience in response to a task load that cannot be directly measured and plays an important role in human performance (Hancock and Matthews 2019; Matthews, Reinerman-Jones, Barber et al. 2015; Van Acker et al. 2018). Hence, we need to develop a reliable mental-workload measuring method to fully understand the role of mental workload in human-computer interactions.

Mental workload is multidimensional and determined by the characteristics of the task, of the operator, and the environmental context, which is difficult to directly measure (Hancock and Matthews 2019; Matthews, Reinerman-Jones, Barber 2015; Van Acker et al. 2018; Young et al. 2015). Although mental workload is difficult to directly observe, previous studies

CONTACT Yaqin Cao caoyaqin.2007@163.com School of Management Engineering, Anhui Polytechnic University, No. 8 Beijing Middle Road, Jiujiang District, Wuhu 241000, P. R. China

have suggested that it can be inferred from the measurement of physiological processes (Casali and Wierwille 1984; Charles and Nixon 2019; Matthews, Reinerman-Jones, Barber et al. 2015). Compared to subjective and performance measures, physiological indices have better performance in terms of sensitivity, diagnostic ability, and nonintrusiveness (Parasuraman and Rizzo 2007; Zhao, Liu, and Shi 2018). The physiological indices used to measure mental workload involve recording electroencephalographic (EEG) activity, electrocardiographic (ECG) activity, electrodermal activity (EDA), eye movement, respiration activities, and blood pressure (Charles and Nixon 2019; Kramer 1991; Matthews, Reinerman-Jones, Barber et al. 2015; Young et al. 2015), but a question that arises when people put this into practical use is which variable(s) one should measure in order to obtain the best workload assessment. As a comparison, peripheral physiological signals provide alternative approaches for quick and practical applications than EEG activity (Gopher and Donchin 1986; Zhao, Liu, and Shi 2018). Additionally, from the review papers summarised by Charles and Nixon (2019), Tao et al. (2019), and Young et al. (2015), there was no universal solution to measuring mental workload with physiological indicators and no method was found to be superior to others.

Another key criticism of the aforementioned various physiological measures is that most empirical studies have used univariate statistical approaches to investigate the multifaceted concept of mental workload (Hogervorst, Brouwer, and Van Erp 2014; Matthews, Reinerman-Jones, Wohleber et al. 2015; Wickens 2017). Moreover, numerous studies have shown that metrics from different physiological systems are only weakly correlated at best, and the divergence between subjective ratings and objective measures is even more starkly apparent (Hancock and Matthews 2019). Furthermore, as metrics indicate that latent mental workload may differ among individuals (Hockey et al. 2009), classifier models should be used to establish workload discriminators (Baldwin and Penaranda 2012; Cinaz et al. 2013; Matthews et al. 2015a; Wilson and Russell 1999, 2003; Zhao, Liu, and Shi 2018). Linear discriminant analysis (LDA), support vector machine (SVM), classification tree and k-nearest neighbour (KNN), and artificial neural networks (ANNs) have been widely used in data classification, and among them, ANNs may be more suitable if multiple physiological features are used (Tjolleng et al. 2017; Wilson and Russell 2003). Wilson and Russell (2003) obtained an accuracy of 85.8% for ANNs trained on within-difficulty manipulation. Jimenez-Molina, Retamal, and Lira (2018) used multiple physiological features as inputs and obtained an accuracy of 93.7% by combining electrodermal activity, EEG, and photoplethysmography (PPG). Although several classification methods have been applied to classify mental workload using physiological parameters, they did not examine how well different physiological indices can be used to assess workload, and to what extent a combination of varied indices would improve performance.

Considering the above issues, the present study developed an ANN model to classify the mental workload imposed by a type of mental arithmetic task characterised by different levels of difficulty. As muscle activity can also affect cognitive attention (Stephenson et al. 2019) and related physiological signals, this type of signal was also recorded during the tasks to ascertain the same level of physical load in the different tasks. Several most commonly used peripheral physiological measures (EDA, respiration, EMG, and PPG) were recorded according to previous reviews (Charles and Nixon 2019; Tao et al. 2019; Young et al. 2015). Moreover, different classifiers were conducted to examine how well different indices can estimate workload and the performance of a combination of different indices.

2. Method

2.1. Participant

The experiment aimed to investigate the gauging of mental workload in computer work with non-invasive wearable sensors. We recruited 18 right-handed, healthy individuals with normal or corrected-to-normal vision. Mean (±SD) age, body mass, stature, and body mass index (BMI) of the participants were 20.1 ± 0.94 years, 67.6 ± 11.7 kg, 177 ± 4.1 cm, and 21.5 ± 3.3, respectively. The participants had no history of neurological or mental illness or organic disease, such as heart-related and skin conditions. None of the participants were allergic to the electrodes used in the experiment. Three items were used from the DSSQ scale (Dundee Stress State Questionnaire, Matthews et al. 1999) to obtain participant self-reports of stress, nervousness and mood. Responses were scored from 4 for "definitely" to 1 for "not at all". None of the participants reported experiencing stress, nervousness or negative mood before the experiment. Each of the participants provided written informed consent before the experiment and received financial compensation.

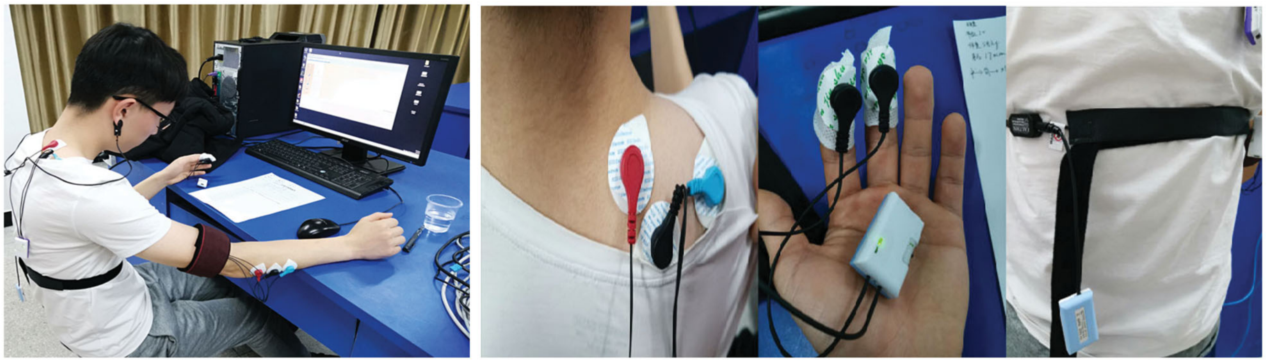

Figure 1. Sensors placement on a subject.

2.2. Apparatus

All stimuli were presented on an Acer P229HQL screen with 1920 × 1080 resolution. The experiment was controlled by E-prime 2.0 (Psychology Software Tools, Inc., Pittsburgh, PA). Participants sat at a distance of approximately 60 cm from the screen without a chin rest in a quiet room with normal light. Non-invasive wearable electrocardiography (ECG), EDA, electromyography (EMG), and respiration sensors were used for physiological signal collection (Figure 1). Signal recording and analysis were conducted on the ErgoLAB human-machine-environment testing cloud platform (Kingfar International Inc., Beijing, China). The ECG raw data were recorded by placing three electrodes below the left (negative) and right (ground) clavicle and the left costal cartilage (positive), respectively. EMG signals were collected from the finger extensor and trapezius muscles, and EDA from the index and middle fingers of the left hand. Scrubbing cream and a cotton swab were used to reduce skin impedance. Respiration activities were recorded by attaching the breath-measuring module to the subject's chest with elastic straps. Three Kangren® pre-gelled disposable AgCl electrodes with an active area of 6.15 mm^2 (Type: CH3236TD) were placed on the muscle belly and top and down of the finger extensor muscles of the right hand. The sample rate of EMG was 1024 Hz, with a bandpass filter of 5–500 Hz and a noise level of 1.6 μV. The root mean square (RMS) of the signal was determined using a time constant of 120 ms. The sample rate of EDA was 32 Hz and that of ECG was 1024 Hz with a noise level of 1.6 μV. All electrode impedances were maintained below 5 kΩ during the experiment.

2.3. Procedure

A simulated computer-based work consisting of mental arithmetic tasks characterised by different levels of difficulty was designed to impose different mental workloads on the participants to explore the relationship between mental workload, task performance, and physiological parameters. The participants were instructed to rest well and not drink caffeinated beverages the night before the experiment. Before the tasks, the participants were asked to play a training version of the stimulus task until they became familiar with the rules and controls. Then, the participants were trained to use the NASA-TLX, and after completing each task, the NASA-TLX was used to collect the subjective ratings of perceived mental load. The participants completed three tasks, with a 10 min rest between each two tasks.

A type of mental arithmetic task with three levels of difficulty was designed to elicit varying levels of mental workload. In the 'easy' level, participants were required to complete a mental arithmetic task consisting of three numbers (i.e. a + b+c) in 15 min. Each mental arithmetic operation lasted 6 s and was randomly generated. Participants were asked to respond as quickly as possible by clicking the right mouse button if the answer was between 10 and 20, and otherwise to click the left mouse button. Then, in the 'medium' level, participants were required to complete a mental arithmetic task consisting of five numbers (i.e. a + b-c + d-e) in 15 min. Each mental arithmetic operation lasted 6 s and was randomly generated. Participants were asked to respond as quickly as possible by clicking the right mouse button if the answer was between −5 and 5, and otherwise to click the left mouse button. In the 'difficult' level, participants were required to complete a mental arithmetic task consisting of seven numbers (i.e. a + b-c + d-e + f-g) in 15 min. The mental arithmetic operation was randomly generated and lasted 6 s. Participants were asked to respond as quickly as possible by clicking the right mouse button if the answer was between -4 and 6, and otherwise to click the left mouse button. The tasks were programmed and presented using E-prime professional. Six task orders can be obtained based on the three tasks with different levels of difficulty. Eighteen participants were randomly and equally divided into six groups. Finally, three participants were assigned to each task order. Before the formal experiment, the participants practiced for 1 min; participant accuracy was not required to reach a certain degree in the process.

The environmental conditions were also controlled with soft light (170 LX ± 3 LX) to eliminate the impact of light on task performance. The microclimatic environment was set at a comfortable level with a temperature of 23.6 ± 0.9 °C and a relative humidity of 36.2 ± 1.5%. Some physiological parameters used to measure or predict mental workload are sensitive to temperature, e.g. humidity, age, sex, time of day, and season (Charles and Nixon 2019; Kramer 1991). Hence, the environmental conditions during the experiment remained constant to eliminate the impact of the task environment to the extent possible.

2.4. Data processing and statistics

Behavioural data (i.e. the response times (RTs) from the task display on the computer to clicking the mouse and accuracy at each task level), subjective ratings of perceived mental workload, and physiological responses were analysed. First, physiological signals were pre-processed through ErgloLAB (Beijing Kingfar Technology). Data cleaning was conducted with wavelet denoising and high-pass, low-pass, and RMS filtering. As the signals were at different scales and some processed data did not obey the approximate normal distribution, a normalisation process was used to transform the data (Guyon and Elisseeff 2006). Here, z-score standardisation was applied to the current value of an index (standardisation of x is (x_{curren}–$x_{average}$)/(SD of raw data)). The signal under each task level was divided into eight-time nodes (2 min before the experiment, 2 min, 4 min, 6 min, 8 min, 10 min, 12 min, and 14 min during the experiment). Then, univariate repeated-measures analysis of variance (ANOVA) was used to examine the effects of change in the levels of mental demands (easy, middle, and difficult) on subjective ratings, task performance, and physiological parameters. Violation of sphericity was handled with a Greenhouse–Geisser correction, and the effect size (eta squared η^2) is reported for all ANOVAs. A paired *t*-test was used to analyse the pairwise comparisons. The data analysis was carried out using SPSS version 24.0 (IBM Corporation, Armonk, NY, USA). Statistical significance for all tests was set at $p < 0.05$. Data outliers were analysed according to a study conducted by Cao et al. (2019).

The performance metrics of the classifier are classification accuracy, recall, and precision (Fawcett 2006; Jimenez-Molina, Retamal, and Lira 2018; Tjolleng et al. 2017), which were calculated to compare the performance of the classifier in this study. Classification accuracy was defined as the ability to correctly classify positive and negative results and is shown in Equation1.

$$\text{Accuracy} = \frac{TP + TN}{P + N} \times 100 \quad (1)$$

Where *TP* denotes true positives correctly labelled as high mental workload at the corresponding level, *TN* denotes true negatives correctly labelled as non-high mental workload at the corresponding level, *P* and *N* denote the count of positives and negatives. Precision was defined as the fraction of predictions that were accurate, and recall was defined as the fraction of instances that were accurately predicted, as shown in Equations 2 and 3, respectively.

$$\text{Precision} = \frac{TP}{TN + FP} \times 100 \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100 \quad (3)$$

Where *FN* denotes false negatives and refers to data points incorrectly labelled as high mental workload at the corresponding level, and *FP* denotes false positives and refers to data points incorrectly labelled as non-high mental workload at the corresponding level.

3. Results

After the experiment, three types of data were obtained, i.e., subjective ratings of mental workload, task performance, and physiological responses. First, the impact of the independent variable (i.e. task difficulty) on the dependent variables (i.e. subjective ratings of mental workload, task performance, and physiological measures) was analysed. Then, the relationships between different measures of mental workload were analysed. The final analysis used machine learning to classify mental workload based on multimodal measures of workload, and regression models were built to predict the mental workload scores based on physiological and behavioural data.

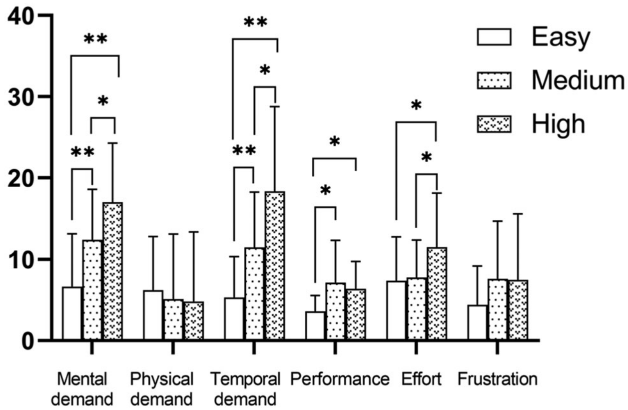

Figure 2. Comparisons among three tasks with different levels of difficulty from the NASA-TLX subscale. Error bars represent standard error (*$p < 0.05$, **$p < 0.01$).

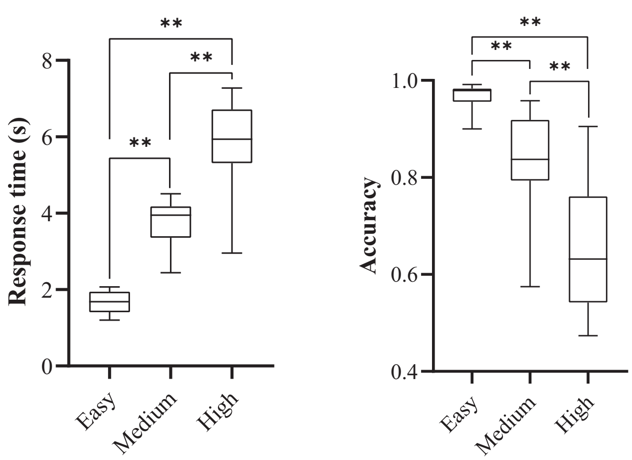

Figure 3. The comparison of task performance evoked by three tasks with 95% confidence intervals.

3.1. Self-reported mental workload levels

The NASA-TLX (Hart and Staveland 1998), including six dimensions, was used to collect participant subjective responses to assess mental workload. The subjective ratings (SR) of mental workload were 40.14 (SD = 12.64), 51.91 (SD = 14.55), and 62.79 (SD = 14.75) for the easy-, medium-, and difficult-level tasks, respectively. The statistical analysis showed that there was a main effect for task difficulty with $F(2,34)=32.559$, $p < 0.001$, $\eta^2=0.657$. Additionally, paired t-tests showed that there were significant differences between the easy and medium ($t=-6.07$, $p < 0.001$) level tasks, between the easy and difficult ($t=-6.93$, $p < 0.001$) level tasks, and between the medium and difficult ($t=-3.82$, $p=0.001$) level tasks. Figure 2 presents comparisons among tasks with different levels of difficulty for the NASA-TLX subscale values. The subjective ratings of mental workload from tasks with different levels of difficulty showed that there were no differences in physical demand and frustration among tasks.

Table 1. Description of physiological indexes measured in the experiment.

Indexes	unit	Description
AVHR	bpm	The average heartbeats per minute.
LF/HF	–	LF/HF is the ratio between the integral of Power Spectral Density (PSD) calculated in the low frequencies band (LF) and the PSD calculated in the high frequencies band (HF).
Y_{rms}	µV	The root mean square of EMG amplitude.
MF	%	The median frequency of EMG.
SC Mean	µS	The average skin conductance.
Respiration	rpm	The average respiration within a period of time.

3.2. Task performance (TP)

The accuracy and response time (i.e. RT, the time from stimulus presentation to the participant clicking the mouse) were also collected during the experiments. The accuracy of the easy-, medium-, and difficult-level tasks was 0.97 (SD = 0.02), 0.84 (SD = 0.10), and 0.65 (SD = 0.12), respectively. The response time of the easy-, medium-, and difficult-level tasks was 1.67 s (SD = 0.31), 3.76 s (SD = 0.60), and 5.86 s (SD = 1.11), respectively. The statistical analysis showed that there was a main effect of task difficulty on the accuracy and response time with $F(2,34)=397.074$, $p < 0.001$, $\eta^2=0.959$ and $F(1.51,25.72)=83.484$, $p < 0.001$, $\eta^2=0.831$, respectively. Figure 3 shows that the participants were more accurate and responded faster in easy-level tasks than in the medium- and difficult-level tasks and in the medium-level tasks than in the difficult-level tasks.

3.3. Physiological results

The physiological signals, consisting of the average of heartbeats per min, the LF/HF ratio, mean skin conductance (SC), and respiration rate were recorded and analysed. The RMS of the EMG amplitude (Y_{RMS}) and the median frequency were also collected to identify whether there was a difference in physical workload between the various tasks. The physiological indices were selected based on Charles and Nixon's review (2019) and the description of each index is shown in Table 1.

A summary of the descriptive statistics for the physiological indices is shown in Table 2. Then, the physiological features were subjected to ANOVA to search for differences in activation for the different conditions and analyse the relevance of those features for mental workload assessment.

Table 2. The mean and SD of physiological indexes evoked by different tasks.

	Tasks					
	Easy level		Medium level		Difficult level	
Index	Mean	SD	Mean	SD	Mean	SD
AVHR	82.5	8.5	84.5	9.4	84.3	8.5
LF/HF	8.24	3.86	7.98	4.37	11.22	7.12
Y_{RMS}	32.68	19.53	31.43	17.47	30.15	13.76
MF	20.08	2.56	20.36	1.78	20.53	1.78
SC Mean	12.01	7.51	13.38	7.17	13.78	7.94
Respiration	18.60	2.58	20.00	2.70	20.85	2.83

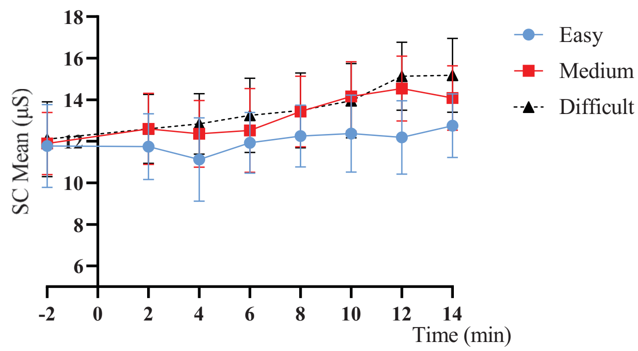

Figure 4. The comparison of SC means evoked by three tasks.

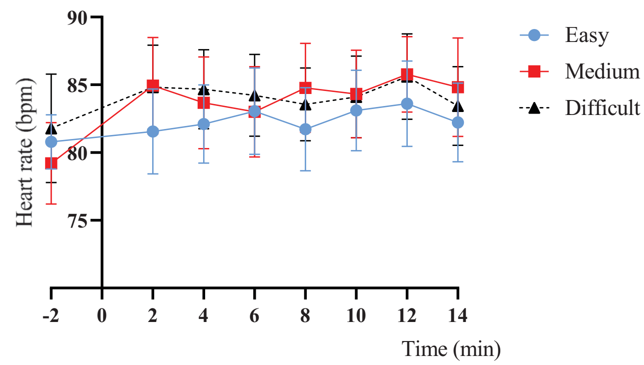

Figure 5. The comparison of heart rate evoked by three tasks.

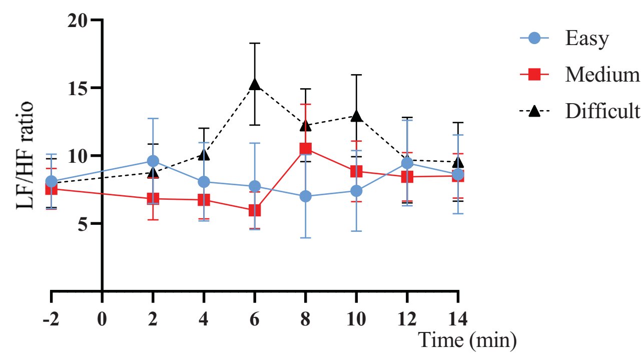

Figure 6. The comparison of LF/HF evoked by three tasks.

3.3.1. Electrodermal activity

The EDA signals were bandpass filtered at 0.05–500 Hz and digitised at a rate of 64 Hz through the ErgoLAB platform (Beijing Kingfar Technology). A band-stop filter was used to eliminate the 50 Hz power frequency interference. A moving RMS filter was used to eliminate noise with a window size of 125 ms. Data of two subjects were deleted due to missing signals during the experiment. Finally, EDA data were obtained from 16 participants (18–21 years old, M_{age} = 20.0 years, SD_{age} = 0.97). The repeated-measures ANOVA results showed that there was a significant main effect of task level with F(2,30) = 7.586, $p = 0.002$, η^2 = 0.336, and the two-sample paired *t*-test showed that the difficult-level task evoked higher mean SC than did the medium- and easy-level tasks with $t(15) = -4.955$, $p < 0.001$ and $t(15) = -5.262$, $p < 0.001$. Mean SC was higher in the medium level than in the easy level with $t(15) = -2.212$, $p < 0.043$. Furthermore, the results of the comparison showed that there were no significant differences for EDA before the tasks with $p_s > 0.05$. The variations of mean SC with time under the different tasks are shown in Figure 4.

3.3.2. ECG activity

Cardiac activity can be analysed in the time or frequency domain, in which heart rate variability (HRV) and HR are typically reported measurements (Charles and Nixon 2019). Wavelet denoising with a medium intensity level was used to eliminate white noise. Then, signals were bandpass filtered at 0.5–20 Hz and digitised at a rate of 1024 Hz. The band-stop filter was used to eliminate 50 Hz power frequency interference. Two metrics from the time and frequency domains were selected (i.e. HR and LF/HF). The repeated-measures ANOVA results of HR showed that there was no significant main effect of task level with F(2,34) = 2.645, $p = 0.086$, $\eta^2 = 0.135$. The paired *t*-test showed that the medium and difficult levels evoked almost significantly higher HR than did the easy level with $t(17) = -2.071$, $p = 0.054$ and $t(17) = -1.965$, $p = 0.066$. However, there was no significant difference between the medium and difficult levels with $t(17) = 0.127$, $p = 0.900$. Moreover, the results of the comparison showed that there were no significant differences for HR before the tasks with $p_s > 0.05$. The HR variation with time under the different tasks is shown in Figure 5.

The repeated-measures ANOVA results of LF/HF showed that there was a significant main effect of task level with F(2,34)=5.6, $p = 0.008$, η^2 =0.248. The paired *t*-test showed that the difficult level evoked higher LF/HF than did the easy and medium levels with $t(17) = -2.641$, $p = 0.017$ and $t(17) = -2.411$, $p = 0.027$, but there was no significant difference between the easy and medium levels with

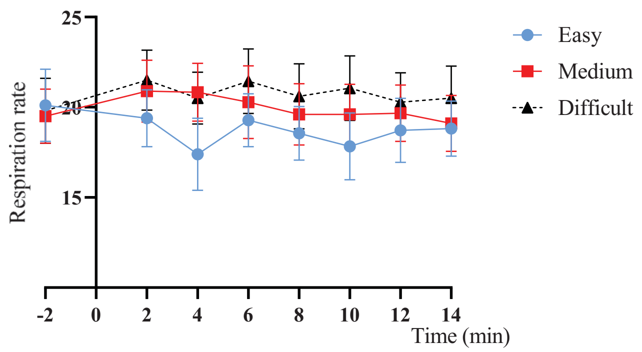

Figure 7. The comparison of respiration evoked by three tasks.

$t(17)=0.410$, $p=0.687$. Furthermore, the results of the comparison showed that there were no significant differences for LF/HF before the tasks with $p_s>0.05$. The LF/HF variation with time under the different tasks is shown in Figure 6.

3.3.3. Respiration

Wavelet denoising with a medium intensity level was used to eliminate white noise. Then, respiration signals were bandpass filtered at 0.5–20 Hz and digitised at a rate of 64 Hz. The bandstop filter was used to eliminate 50 Hz power frequency interference. Baseline denoising was performed with a cut-off frequency of 0.5 Hz. The repeated-measures ANOVA results of respiration showed that there was a significant main effect of task level with $F(2,34) = 17.624$, $p<0.001$, $\eta^2=0.509$. The paired t-test showed that the easy level evoked lower LF/HF than did the medium and difficult levels with $t(17)=-4.639$, $p<0.001$ and $t(17)=-4.879$, $p<0.001$, and the medium level evoked lower respiration than did the difficult level with $t(17)=-2.299$, $p=0.034$. Further, the results of the comparison showed that there were no significant differences in the respiration rate before the tasks with $p_s>0.05$. The respiration variation with time under the different tasks is shown in Figure 7.

3.3.4. Electromyography

EMG activity was also recorded during the experiment to identify whether there was a significant difference between the various tasks in physical load. The EMG signals were pre-amplified at the source (gain of 100) and bandpass filtered (5–500 Hz). The RMS of the muscle amplitude was determined using a time constant of 125 ms. The band-stop filter was used to eliminate 50 Hz power frequency interference. The amplitude and median frequency (MF) are the most commonly used indicators in EMG studies (Halder et al. 2018); with the physical load increasing, the muscular amplitude becomes larger and MF decreases (Halder et al. 2018; Viitasalo and Komi 1977). The repeated-measures ANOVA results of Y_{rms} and MF showed that there were no main effects of task level with $F(2,34) = 0.412$, $p=0.666$, $\eta^2=0.024$ and $F(2,34) = 0.294$, $p=0.747$, $\eta^2=0.017$, respectively. The results confirmed that there was no significant difference in physical load among the different tasks.

3.4. Mental workload classification

Feature selection was conducted to improve the efficiency and time costs of the classification based on the above results. As can be seen from the ANOVA results, a total of four features were found to have significantly different distributions among the three difficulties, suggesting that the tasks corresponded to different levels of difficulty and marking the value of these features for later classification of the three tasks. All these physiological indexes and performance data were used as classifier inputs. The tasks with different levels of difficulty were used as outputs. A total of 378 data points (54*7) were extracted from the seven-time nodes for 18 participants. ANNs have been widely used in data classification, especially for highly-nonlinear and strongly-coupled relationships between multi-input and multi-output variables (Wilson and Russell 1999, 2003). A feedforward backpropagation neural network (BPNN) was used to classify the three task levels. The physiological indices were used as inputs to the network. A hidden layer of 10 nodes was used with three output nodes: low, medium, and high. Eighty percent of the data were randomly selected as the training set. The remaining 20% were used as test data to determine the accuracy of the ANN training. There are many classification models for developing predictive models, of which ANN lacks transparency and is difficult to interpret (Fong et al. 2010). LDA, SVM, classification tree, and KNN are commonly used to develop predictive nonlinear models that are suitable for physiological metrics. The LDA was employed because of the low number of samples, which sometimes can increase the problem of singular covariance matrices (Chanel et al. 2011). The classification conducted in the study underwent a 10-fold cross-validation, which is a statistical method for evaluating models (Jang et al. 2015). The comparison of different classification models is presented in Table 3. The criteria (i.e. accuracy, recall, and precision) were used to assess the classification performance of different classifiers (Fawcett 2006; Tjolleng et al. 2017). The BP network and classification models were programmed and

Table 3. Summary of classification results by taking different indexes as inputs.

Metrics	Classification method	Accuracy (%)	Recall (%)	Precision (%)
All	BPNN	96.40	96.37	96.50
	Cubic SVM	96.30	96.33	96.33
	Weighted KNN	95.20	95.00	95.33
	Medium tree	93.40	93.33	93.67
	LDA	90.70	90.67	91.00
All physiological indexes	BPNN	78.30	77.40	77.80
	Weighted KNN	77.00	76.33	75.67
	Quadratic SVM	76.70	75.67	76.33
	Fine tree	74.90	74.33	73.33
	LDA	61.90	60.67	61.00
ECG + EDA	BPNN	58.50	58.47	58.47
	Bagged tree	56.90	56.67	57.00
	Weighted KNN	55.60	55.67	55.33
	Fine Gaussian SVM	55.00	56.30	55.00
	LDA	40.00	40.30	39.67
ECG + Res.	BPNN	50.50	52.70	51.47
	Weighted KNN	50.00	50.33	50.00
	Fine Gaussian SVM	49.50	49.33	45.67
	Bagged tree	46.30	46.00	46.70
	LDA	42.90	35.16	55.56
EDA + Res.	BPNN	48.90	48.97	48.70
	RUSBoosted trees	47.60	48.33	48.00
	Medium Gaussian SVM	47.40	47.33	47.33
	Weighted KNN	46.80	47.33	47.00
	LDA	45.80	46.00	45.67
EDA	BPNN	47.40	56.33	49.73
	Coarse Gaussian SVM	46.80	47.00	46.30
	Cubic KNN	46.60	46.67	46.00
	Boosted trees	45.20	45.33	45.00
	LDA	36.80	37.00	36.67
ECG	BPNN	46.00	47.43	45.47
	Bagged tree	46.00	45.66	46.00
	Weighted KNN	45.00	45.00	44.31
	Fine Gaussian SVM	44.70	53.63	44.67
	LDA	40.50	40.30	42.67
Res.	BPNN	43.90	43.90	43.53
	Linear SVM	42.60	42.33	41.00
	LDA	42.60	42.00	40.00
	Coarse tree	41.80	42.00	35.00
	Subspace KNN	39.20	42.67	39.00

executed using the classification learner toolbox of MATLAB (2018b). The raw data, BPNN, experimental tasks, and supplementary data can be downloaded from https://share.weiyun.com/5Ed0GpT.

The accuracies of the top five classifier models with different metrics as inputs are presented in Table 3. The best performance could reach approximately 78% with all metrics as inputs using BPNN and cubic SVM. Moreover, the full model that used all types of data as inputs achieved an accuracy of 96.4% (BPNN). Classifier methods from the classification learner toolbox were compared with BPNN, in which weighted KNN could achieve an accuracy of 77%, and quadratic SVM reached an accuracy of 77.6% with all physiological indices as inputs. The other sensors had a lower level of classification accuracy, while the accuracy of classification markedly decreased when setting a single physiological index as input, i.e. with an accuracy below 48%. LDA had the lowest accuracy in mental workload recognition. Moreover, pairwise comparison tests showed the performance by LDA was lower than that by the other models when using all physiological indices as inputs ($p < 0.01$). When using combined signals as inputs, the accuracy of recognition could increase by approximately 30% compared to a single physiological index as inputs. In addition, in the three types of physiological indexes, EDA and ECG achieved similar performance ($p > 0.05$), and had a higher accuracy than respiration as inputs ($p < 0.05$). Further, the pairwise comparison tests showed that the classification accuracy significantly improved when setting both physiological and task performance indices as inputs ($p < 0.01$).

4. Discussion

4.1. General overview of findings

The present study investigated the multimodal measurement of mental workload during simulated computer tasks and identified the mental workload imposed by tasks with different levels of difficulty based on machine learning. Four categories of peripheral physiological signals (ECG, EMG, RSP, and GSR) were recorded. Features based on ANOVA were selected and entered into the classification models as inputs. Additionally, we examined how well different variables could be used to assess mental workload.

Our results showed that subjective ratings significantly changed with changing task difficulty and that performance decreased with increasing task difficulty. It must be noted that although the total score of the NASA TLX here revealed a significant difference, no differences in the physical demand and frustration dimensions were observed. A possible explanation could be that some dimensions may be less sensitive to the task requirement because they are subjective and they assess mental workload as a whole retrospective, and non-dynamic concept (Jafari et al. 2020; Shuggi et al. 2017). Besides, the lack of difference in physical demand also eliminated the effect of muscle activity on cognitive attention (Stephenson et al. 2019). Different mental workload measures account for different sources of demand during dynamic multitasking, which may have contributed to this (De Waard and Lewis-Evans 2014). Another reason for the difference may be the curvilinear relationship between mental workload and performance or subjective evaluation (Hancock and Szalma 2006; Mallat et al. 2019). There was no significant difference in muscle activity among tasks, confirming that there was the same effect of physical load on mental workload.

The main finding of this study was that indices from EDA and respiration had a significant increment as task difficulty increased. Contrary to expectations, there were no main effects of task levels for ECG metrics. There were no significant differences for AVHR and LF/HF among the tasks, which was not consistent with the results of previous studies (De Rivecourt et al. 2008; Fairclough, Venables, and Tattersall 2005; Finsen et al. 2001; Fournier, Wilson, and Swain 1999; Orlandi and Brooks 2018; van Amelsvoort et al. 2000; Veltman and Gaillard 1998). Moreover, there was a higher respiration rate with increasing task difficulty. Additionally, the classification of mental workload using different indices as inputs showed that classification models using all physiological indices as inputs could obtain a satisfying accuracy of 78.3% by BPNN. When setting response time as inputs combined with physiological indexes, the accuracy increased to 96.4%. As numerous studies have indicated, the multimodal methods used in this study can provide different perspectives and complement one another in the assessment of mental workload (Jafari et al. 2020; Ryu and Myung 2005). Furthermore, the high classification accuracies indicate that machine learning methods have great potential for the prediction of task and difficulty levels and may be used to develop situation-aware recognition systems of mental workload and even embedded adaptive human-computer interaction platforms.

4.2. Effective indices of mental workload measurement and classification

We proved the sensitivity of physiological signals as mental workload concomitant in a simulated computer work. As traditional analysis (ANOVA) showed, there were main effects of task difficulty level on ECG, EDA, and respiration indexes. However, for ECG indexes, HR and HRV, indicators of mental workload, did not return the expected results, in that they did not increase with task difficulty level. Previous studies have suggested that task demands increase is not accompanied by changes in HRV as long as the manipulation affects only structural or computational structures in the human information processing system (Jorna 1992; Mulder et al. 2004). In this view, the task difficulty levels seem not to differ significantly in this respect, especially between the easy and middle task difficulty levels. Differences in cognitive demand evoked by the tasks may have contributed to this disparity (Fairclough, Venables, and Tattersall 2005; Matthews et al. 2015; Mulder et al. 2004).

Both EDA and respiration significantly increased with increasing task difficulty levels. Our results are consistent with those of Collet, Salvia, and Petit-Boulanger (2014) and Engström, Johansson, and Östlund (2005), but in one of our previous studies, there was no significant difference between difficult- and medium-level tasks (Ding, Cao, and Wang 2019). Charles and Nixon (2019) summarised in their review that EDA is sensitive to sudden but not gradual changes in mental workload. The reason may be that the tasks in this study involved arithmetic problems of increasing difficulty but the previous study employed tasks of office work gradually increasing in difficulty. Moreover, the changing trend in SC (Figure 4) was consistent with the findings of Fairclough and Venables (2006) and Miyake et al. (2009). The respiration rate was found to increase with increasing task difficulty (Charles and Nixon 2019; Fairclough, Venables, and Tattersall 2005; Nixon and Charles 2017) and is the most useful among the respiratory measures (Roscoe 1992). In our experiment, a higher respiration rate was evoked by increasing mental demand.

The result was consistent with those of previous studies (Brookings, Wilson, and Swain 1996; Fairclough, Venables, and Tattersall 2005; Grassmann, Vlemincx, von Leupoldt, and Van den Bergh 2016). Grassmann, Vlemincx, von Leupoldt, Mittelstädt, et al. (2016) claimed in their review that the respiration rate increases as stress and workload increase, but that this measure is highly related to physical activity. EMG was also measured and analysed on Y_{rms} and MF in our experiment, and there was no significant difference among the various tasks in EMG signals. This verified that the physical activity effect was at the same level in the three tasks.

It should be noted that participants were asked to conduct varied tasks in a limited amount of time in this experiment, which could cause stress. Further, if a given cognitive activity requires extensive application of resources and that activity has to be carried out for long, uninterrupted periods of time, then it is likely that the activity would induce stress (Hancock 1989; Warm, Parasuraman, and Matthews 2008). Accordingly, the results of this study could have been influenced by stress. The previous studies have indicated that stress involves the activation of the hypothalamus-pituitary-adrenocortical axis, which plays an important role in the regulation of various physiological processes, especially for ECG and EDA indexes (Healey and Picard 2005; Sun et al. 2010); therefore, the physiological responses may be affected, and the classification results may be influenced.

Our result also showed that multimodal measures can be used to monitor mental workload in simulated computer works. The classification results showed that when using all physiological indexes as input, an accuracy of 78.3% could be obtained. This accuracy was significantly improved compared to when setting a single or two physiological indexes as input. In the applied environment, identifying mental workload from physiological signals largely relies on classification accuracy, the access of data collection, and the processing method (Wilson and Russell 2003). Previous studies have suggested that within-level classification accuracy between two task difficulty levels must approach 95% to be acceptable (Wilson and Russell 2003; Zhao, Liu, and Shi 2018). In our paper, an overall accuracy of 78.3% was achieved for within-level classifications by BPNN, signifying that identifying an operator's mental workload by collecting his/her physiological signals from portable equipment could not be practically applied at present. Although we achieved an accuracy of 96.4% when setting all physiological indexes and response time as inputs, response time is not generally available in several settings. The limited types of physiological indexes extracted and the data from shorter time could have led to this lower classification accuracy than those in previous studies; for example, there were 42 physiological indices in total in the study of Zhao, Liu, and Shi (2018) and 43 features in the study of Wilson and Russell (2003). Another reason may be that tasks were completed in a limited time, which could have caused stress (Hancock 1989; Warm, Parasuraman, and Matthews, 2008) and the physiological signals could have been affected by this factor, which could also lead to a low classification accuracy of mental workload. Our results also showed that LDA performed worse than the other models when setting all physiological indices as inputs. This linear method is best suited for data with linear boundaries between different classes and is based on the assumption of the normality of data distribution in each class (Hastie, Tibshirani, and Friedman, 2009); these factors could have led to the low accuracy of LDA. Another reason may be that the class boundaries are nonlinear in this classification, for which, nonlinear methods could be more suitable (Kolodyazhniy et al. 2011). The classification accuracies indicate that machine-learning methods have great potential for predicting the task and difficulty levels. With the aid of wearable and mobile sensors, an operator's mental workload can be monitored in real-time. If the predicted mental workload is underload or overload, we can provide suggestions in time.

4.3. Limitations and future directions

The study was conducted in a laboratory with the advantage of the experimental conditions being carefully controlled. Unlike the manipulation of task difficulty under simulated conditions, it is difficult to identify task demand in the actual environment and physiological signal collection will also be affected by the real environment. This signifies that the process and conclusions cannot be extrapolated to occupational settings. Moreover, the participants in the experiment were all male university students. As previous studies have suggested, the peripheral and EEG indices should be combined in the prediction of mental workload. In the future, different types of tasks should be considered, and how to define the 'red zone' at which workload becomes unacceptable should be investigated. Finally, how to promote effective interventions to reduce effort demand and at the same time retain work performance are two critical issues. It is possible that in the near future, smart offices can measure and identify mental workload automatically, then provide suggestions for a break when mental workload approximates the 'red zone.'

5. Conclusion

In this study, we measured the changes in physiological activity as the level of task difficulty varied and attempted to validate indicators for estimating mental workload. Muscle activity was also recorded during the tasks to ensure the same level of physical load among tasks. Mental workload was treated as a subjective experience in response to a task load. Physical load would also affect participant subjective feelings to some extent, and this additional load was rarely considered in previous studies. Thus, multiple measures were used in this study to perform a thorough and precise measurement of mental workload, and the classification performance of BPNN was compared with those of the most commonly used models using multiple measures as inputs. The main findings of this study are that: (1) using limited physiological signals from portable equipment can obtain an acceptable classification accuracy; (2) ECG and EDA signals have more discriminating power compared to respiration for mental workload classification; (3) LDA performed worse than other models when using all physiological indexes as inputs. The methods used in this study could be applied to office workers, and the findings provide preliminary support and theoretical exploration for follow-up early mental workload detection systems, whose implementation in the real world could beneficially impact worker health and company efficiency.

Acknowledgments

We thank Kingfar International Inc. (China) for providing related equipment and scientific and technological support. We thank the editor and anonymous reviewers for their valuable comments and advice, which help further to improve the quality of this paper.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Natural Science Foundation of China [71801002, 71701003, 71802002]; the Humanities and Social Science Fund of Ministry of Education of China [18YJC630023]; the Anhui Natural Science Foundation Project [1808085QG228]; and Ministry of Education Industry-University Cooperation Collaborative Education Project [201901024006].

ORCID

Yi Ding http://orcid.org/0000-0002-0010-8279

References

Baldwin, C. L., and B. N. Penaranda. 2012. "Adaptive Training Using an Artificial Neural Network and EEG Metrics for within- and Cross-Task Workload Classification[J]." *Neuroimage* 59 (1): 48–56. doi:10.1016/j.neuroimage.2011.07.047.

Brookings, J. B., G. F. Wilson, and C. R. Swain. 1996. "Psychophysiological Responses to Changes in Workload during Simulated Air Traffic Control[J]." *Biological Psychology* 42 (3): 361–377. doi:10.1016/0301-0511(95)05167-8.

Cao, Y., Q. Qu, V. G. Duffy, and Y. Ding. 2019. "Attention for Web Directory Advertisements: A Top-Down or Bottom-Up Process." *International* Journal of Human–Computer Interaction 35 (1): 89–98. doi:10.1080/10447318.2018.1432162.

Casali, J. G., and W. W. Wierwille. 1984. "On the Measurement of Pilot Perceptual Workload: A Comparison of Assessment Techniques Addressing Sensitivity and Intrusion Issues." *Ergonomics* 27 (10): 1033–1050. doi:10.1080/00140138408963584.

Chanel, G., C. Rebetez, M. Bétrancourt, and T. Pun. 2011. "Emotion Assessment from Physiological Signals for Adaptation of Game Difficulty." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 41 (6): 1052–1063. doi:10.1109/TSMCA.2011.2116000.

Charles, R. L., and J. Nixon. 2019. "Measuring Mental Workload Using Physiological Measures: A Systematic Review[J]." *Applied Ergonomics* 74: 221–232. doi:10.1016/j.apergo.2018.08.028.

Cinaz, B., B. Arnrich, R. La Marca, and G. Tröster. 2013. "Monitoring of Mental Workload Levels during an Everyday Life Office-Work Scenario[J]." *Personal and Ubiquitous Computing* 17 (2): 229–239. doi:10.1007/s00779-011-0466-1.

Collet, C., E. Salvia, and C. Petit-Boulanger. 2014. "Measuring Workload with Electrodermal Activity during Common Braking Actions[J]." *Ergonomics* 57 (6): 886–896. doi:10.1080/00140139.2014.899627.

De Rivecourt, M., M. N. Kuperus, W. J. Post, and L. J. M. Mulder. 2008. "Cardiovascular and Eye Activity Measures as Indices for Momentary Changes in Mental Effort during Simulated Flight[J]." *Ergonomics* 51 (9): 1295–1319. doi:10.1080/00140130802120267.

De Waard, D., and B. Lewis-Evans. 2014. "Self-Report Scales Alone Cannot Capture Mental Workload[J." *Cognition, Technology and Work* 16 (3): 303–305. doi:10.1007/s10111-014-0277-z.

Ding, Y., Y. Cao, and Y. Wang. 2019. *Physiological Indicators of Mental Workload in Visual Display Terminal Work[C]// International Conference on Applied Human Factors and Ergonomics*, , 86–94. Cham: Springer.

Engström, J., E. Johansson, and J. Östlund. 2005. "Effects of Visual and Cognitive Load in Real and Simulated Motorway Driving.[J." *Transportation Research Part F: Traffic Psychology and Behaviour* 8 (2): 97–120. doi:10.1016/j.trf.2005.04.012.

Fairclough, S. H., L. Venables, and A. Tattersall. 2005. "The Influence of Task Demand and Learning on the Psychophysiological Response[J]." *International Journal of Psychophysiology* 56 (2): 171–184. doi:10.1016/j.ijpsycho.2004.11.003.

Fairclough, S. H., and L. Venables. 2006. "Prediction of Subjective States from Psychophysiology: A Multivariate Approach[J]." *Biological Psychology* 71 (1): 100–110. doi:10.1016/j.biopsycho.2005.03.007.

Fawcett, T. 2006. "An Introduction to ROC Analysis." *Pattern Recognition Letters* 27 (8): 861–874. doi:10.1016/j.patrec.2005.10.010.

Finsen, L., K. Søgaard, C. Jensen, V. Borg, and H. Christensen. 2001. "Muscle Activity and Cardiovascular Response during Computer-Mouse Work with and without Memory Demands[J]." *Ergonomics* 44 (14): 1312–1329. doi:10.1080/00140130110099065.

Fong, A., C. Sibley, A. Cole, C. Baldwin, and J. Coyne. 2010. "A Comparison of Artificial Neural Networks, Logistic Regressions, and Classification Trees for Modeling Mental Workload in Real-Time[J]." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 54 (19): 1709–1712. doi:10.1177/154193121005401973.

Fournier, L. R., G. F. Wilson, and C. R. Swain. 1999. "Electrophysiological, Behavioral, and Subjective Indexes of Workload When Performing Multiple Tasks: manipulations of Task Difficulty and Training[J." *International Journal of Psychophysiology* 31 (2): 129–145. doi:10.1016/S0167-8760(98)00049-X.

Gopher, D., and E. Donchin. 1986. "Workload: An Examination of the Concept." In *Handbook of Perception and Human Performance, Vol. 2: Cognitive Processes and Performance*, edited by K. R. Boff, L. Kaufman, and P. Thomas, 1–49. Oxford, U.K.: Wiley.

Grassmann, M., E. Vlemincx, A. von Leupoldt, and O. Van den Bergh. 2016. "The Role of Respiratory Measures to Assess Mental Load in Pilot Selection[J]." *Ergonomics* 59 (6): 745–749. doi:10.1080/00140139.2015.1090019.

Grassmann, M., E. Vlemincx, A. von Leupoldt, J. M. Mittelstädt, and O. Van den Bergh. 2016. "Respiratory Changes in Response to Cognitive Load: A Systematic Review[J]." *Neural Plasticity* 2016: 1–16. doi:10.1155/2016/8146809.

Guyon, I., and A. Elisseeff. 2006. "An Introduction to Feature Extraction." In Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L.A. (Eds.). *Feature Extraction: Foundations and applications studies in fuzziness and soft computing*, 1–25. Berlin/Heidelberg, Germany: Springer.

Halder, A., C. Gao, M. Miller, and K. Kuklane. 2018. "Oxygen Uptake and Muscle Activity Limitations during Stepping on a Stair Machine at Three Different Climbing Speeds[J]." *Ergonomics* 61 (10): 1382–1394. doi:10.1080/00140139.2018.1473644.

Hancock, P A., and J L. Szalma. 2006. "Stress and Neuroergonomics [M]." In *The Brain at Work*, edited by R. Parasuraman, and M. Rizzo, 195–206. Oxford, UK: Oxford University Press.

Hancock, P A. 1989. "A Dynamic Model of Stress and Sustained Attention." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 31 (5): 519–537. doi:10.1177/001872088903100503.

Hancock, P. A., and G. Matthews. 2019. "Workload and Performance: Associations, Insensitivities, and Dissociations[J]." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 61 (3): 374–392. doi:10.1177/0018720818809590.

Hart, S. G., and L. E. Staveland. 1998. "Development of NASA-TLX: Results of Empirical and Theoretical Research [M]." In *Human Mental Workload*, edited by P. A. Hancock and N. Meshkati, 139–183. Amsterdam: Elsevier Science.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: data Mining, Inference, and Prediction*. Berlin: Springer Science & Business Media.

Healey, J. A., and R. W. Picard. 2005. "Detecting Stress during Real-World Driving Tasks Using Physiological Sensors." *IEEE Transactions on Intelligent Transportation Systems* 6 (2): 156–166. doi:10.1109/TITS.2005.848368.

Hockey, G. R. J., P. Nickel, A. C. Roberts, and M. H. Roberts. 2009. "Sensitivity of Candidate Markers of Psychophysiological Strain to Cyclical Changes in Manual Control Load during Simulated Process Control[J]." *Applied Ergonomics* 40 (6): 1011–1018. doi:10.1016/j.apergo.2009.04.008.

Hogervorst, M. A., A. M. Brouwer, and J. B. Van Erp. 2014. "Combining and Comparing EEG, Peripheral Physiology and Eye-Related Measures for the Assessment of Mental Workload[J]." *Frontiers in Neuroscience* 8: 322. doi:10.3389/fnins.2014.00322.

Jafari, M.-J., F. Zaeri, A. H. Jafari, A. T. Payandeh Najafabadi, S. Al-Qaisi, and N. Hassanzadeh-Rangi. 2020. "Assessment and Monitoring of Mental Workload in Subway Train Operations Using Physiological, Subjective, and Performance Measures[J]." *Human Factors and Ergonomics in Manufacturing and Service Industries* 30 (3): 165–175. doi:10.1002/hfm.20831.

Jang, E.-H., B.-J. Park, M.-S. Park, S.-H. Kim, and J.-H. Sohn. 2015. "Analysis of Physiological Signals for Recognition of Boredom, Pain, and Surprise Emotions[J]." *Journal of Physiological Anthropology* 34 (1): 25. doi:10.1186/s40101-015-0063-5.

Jimenez-Molina, A., C. Retamal, and H. Lira. 2018. "Using Psychophysiological Sensors to Assess Mental Workload during Web Browsing[J]." *Sensors* 18 (2): 458. doi:10.3390/s18020458.

Jorna, P. G. 1992. "Spectral Analysis of Heart Rate and Psychological State: A Review of Its Validity as a Workload Index." *Biological Psychology* 34 (2–3): 237–257. doi:10.1016/0301-0511(92)90017-O.

Kolodyazhniy, V., S. D. Kreibig, J. J. Gross, W. T. Roth, and F. H. Wilhelm. 2011. "An Affective Computing Approach to Physiological Emotion Specificity: Toward Subject Independent and Stimulus Independent Classification of Film Induced Emotions." *Psychophysiology* 48 (7): 908–922. doi:10.1111/j.1469-8986.2010.01170.x.

Kramer, A. F. 1991. "Physiological Metrics of Mental Workload: A Review of Recent Progress [M]." In *Multiple-Task Performance*, edited by D. L. Damos, 279–327. London: Taylor and Francis.

Mallat, C., J. Cegarra, C. Calmettes, and R. L. Capa. 2019. "A Curvilinear Effect of Mental Workload on Mental Effort and Behavioral Adaptability: An Approach with the Pre-Ejection Period[J]." *Human Factors*. doi:10.1177/0018720819855919.

Matthews, G., L. Joyner, and K. Gilliland., et al. 1999. "Validation of a Comprehensive Stress State Questionnaire: Towards a State Big Three." *Personality Psychology in Europe [J]* 7: 335–350.

Matthews, G., L. E. Reinerman-Jones, D. J. Barber, and J. Abich. 2015. "The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent[J]." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57 (1): 125–143. doi:10.1177/0018720814539505.

Matthews, G., L. Reinerman-Jones, R. Wohleber., et al. 2015. "Workload is Multidimensional, Not Unitary: what Now?[C]." In *Foundations of Augmented Cognition: 9th International Conference, AC 2015, Held as Part of HCI International 2015*, edited by D. D. Schmorrow, and C. M. Fidopiastis, 44–55. Los Angeles, CA, USA; Cham: Springer International Publishing.

Miyake, S., S. Yamada, T. Shoji, Y. Takae, N. Kuge, and T. Yamamura. 2009. "Physiological Responses to Workload Change. A Test/Retest Examination[J]." *Applied Ergonomics* 40 (6): 987–996. doi:10.1016/j.apergo.2009.02.005.

Mulder, L. J. M., D. de Waard, and K. A. Brookhuis. 2004. "Estimating Mental Effort Using Heart Rate and Heart Rate Variability [M]." In *Handbook of Human Factors and Ergonomics Methods*, edited by N. A. Stanton, A. Hedge, and K. Brookhuis, 201–210. London: Taylor and Francis.

Nixon, J., and R. Charles. 2017. "Understanding the Human Performance Envelope Using Electrophysiological Measures from Wearable Technology[J]." *Cognition Technology and Work* 19 (5): 1–12.

Orlandi, L., and B. Brooks. 2018. "Measuring Mental Workload and Physiological Reactions in Marine Pilots: Building Bridges towards Redlines of Performance[J]." *Applied Ergonomics* 69: 74–92. doi:10.1016/j.apergo.2018.01.005.

Parasuraman, R., and M. Rizzo (Eds.). 2007. *Neuroergonomics: The Brain at Work*. New York: Oxford University Press.

Roscoe, A. H. 1992. "Assessing Pilot Workload: Why Measure Heart Rate, HRV and Respiration?[J]." *Biological Psychology* 34 (2–3): 259–287. doi:10.1016/0301-0511(92)90018-P.

Ryu, K., and R. Myung. 2005. "Evaluation of Mental Workload with a Combined Measure Based on Physiological Indices during a Dual Task of Tracking and Mental Arithmetic[J]." *International Journal of Industrial Ergonomics* 35 (11): 991–1009. doi:10.1016/j.ergon.2005.04.005.

Shuggi, I. M., H. Oh, P. A. Shewokis, and R. J. Gentili. 2017. "Mental Workload and Motor Performance Dynamics during Practice of Reaching Movements under Various Levels of Task Difficulty." *Neuroscience* 360: 166–179. doi:10.1016/j.neuroscience.2017.07.048.

Stephenson, M. L., A. G. Ostrander, H. Norasi, and M. C. Dorneich. 2019. "Shoulder Muscular Fatigue from Static Posture Concurrently Reduces Cognitive Attentional Resources[J]." *Human Factors* first published online, doi:10.1177/0018720819852509.

Sun, F T., C. Kuo, H T. Cheng, S. Buthpitiya, P. Collins, and M. Griss. 2010. "Activity-Aware Mental Stress Detection Using Physiological Sensors.". In *International Conference on Mobile Computing, Applications, and Services*, 282–301. Berlin, Heidelberg: Springer.

Tao, D., H. Tan, H. Wang, X. Zhang, X. Qu, and T. Zhang. 2019. "A Systematic Review of Physiological Measures of Mental Workload." *International Journal of Environmental Research and Public Health* 16 (15): 2716. doi:10.3390/ijerph16152716.

Tjolleng, A., K. Jung, W. Hong, W. Lee, B. Lee, H. You, J. Son, and S. Park. 2017. "Classification of a Driver's Cognitive Workload Levels Using Artificial Neural Network on ECG Signals." *Applied Ergonomics* 59: 326–332. doi:10.1016/j.apergo.2016.09.013.

Van Acker, B. B., D. D. Parmentier, P. Vlerick, and J. Saldien. 2018. "Understanding Mental Workload: From a Clarifying Concept Analysis toward an Implementable Framework." *Cognition, Technology and Work* 20 (3): 351–365. doi:10.1007/s10111-018-0481-3.

van Amelsvoort, L. G. P. M., E. G. Schouten, A. C. Maan, C. A. Swenne, and F. J. Kok. 2000. "Occupational Determinants of Heart Rate Variability[J]." *International Archives of Occupational and Environmental Health* 73 (4): 255–262. doi:10.1007/s004200050425.

Veltman, J. A., and A. W. K. Gaillard. 1998. "Physiological Workload Reactions to Increasing Levels of Task Difficulty.[J." *Ergonomics* 41 (5): 656–669. doi:10.1080/001401398186829.

Viitasalo, J. H. T., and P. V. Komi. 1977. "Signal Characteristics of EMG during Fatigue[J]." *European Journal of Applied Physiology and Occupational Physiology* 37 (2): 111–121. doi:10.1007/BF00421697.

Warm, J. S., R. Parasuraman, and G. Matthews. 2008. "Vigilance requires hard mental work and is stressful[J]". *Human Factors* 50 (3): 433–441. doi:10.1518/001872008X312152.

Wickens, C. D. 2017. *Mental Workload: assessment, Prediction and Consequences[C]//International Symposium on Human Mental Workload: Models and Applications*, 18–29. Cham: Springer.

Wilson, G. F., and C. A. Russell. 1999. "Operator Functional State Classification Using Neural Networks with Combined Physiological and Performance Features[J]." *Proceedings* of the Human Factors and Ergonomics Society Annual Meeting 43 (20): 1099–1102. doi:10.1177/154193129904302009.

Wilson, G. F., and C. A. Russell. 2003. "Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. [J]." *Human Factors* 45 (4): 635–643. doi:10.1518/hfes.45.4.635.27088.

Young, M. S., K. A. Brookhuis, C. D. Wickens, and P. A. Hancock. 2015. "State of Science: mental Workload in Ergonomics[J]." *Ergonomics* 58 (1): 1–17. doi:10.1080/00140139.2014.956151.

Zhao, G., Y. J. Liu, and Y. Shi. 2018. "Real-Time Assessment of the Cross-Task Mental Workload Using Physiological Measures during Anomaly Detection." *IEEE Transactions on Human-Machine Systems* 48 (2): 149–160. doi:10.1109/THMS.2018.2803025.