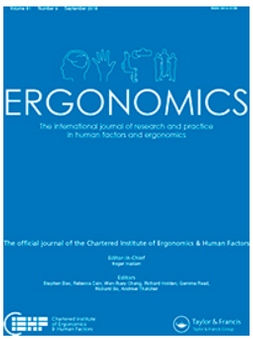

Ergonomics

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 & Zhang Xuefeng

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

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

Accepted author version posted online: 24 Apr 2020.

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

Yi Ding[1,2], Yaqin Cao[1,2*], Vincent G. Duffy[2], Yi Wang[1], Xuefeng Zhang[1]

1. *School of Management Engineering, Anhui Polytechnic University, Wuhu, P. R. China*

2. *School of Industrial Engineering, Purdue University, West Lafayette, IN, USA*

*Correspondence: Yaqin Cao. School of Management Engineering, Anhui Polytechnic University, NO. 8 Beijing middle road, Jiujiang District, Wuhu 241000, P. R. China. E-mail: emiledy@sina.com (Y. Ding), caoyaqin.2007@163.com (Y. Cao), duffy@purdue.edu (V. G. Duffy), xxr0912@gmail.com (Y. Wang).

Total words: 6934

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

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 heartbeats 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 just taking physiological indices as inputs. The present study also shows 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.

Keywords: mental workload; multi-modal 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 optimization 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 et al., 2015a; 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 characteristics of the task, of the operator, and environmental context, which is difficult to directly (Hancock and Matthews, 2019; Matthews et al., 2015a; Van Acker et al., 2018; Young et al., 2015). Although mental workload is difficult to directly observe, the previous research has suggested that it can be inferred from the measurement of physiological processes (Casali and Wierwille, 1984; Charles and Nixon, 2019; Matthews et al., 2015a). Compared to subjective measures and performance measures, physiological indices have better performances in the aspects of sensitivity, diagnostic ability, and nonintrusiveness (Parasuraman and Rizzo, 2008; Zhao, Liu, and Shi, 2018). The physiological indices used to measure mental workload involve recording electroencephalogram (EEG) activity, electrocardiographic (ECG) activity, electrodermal activity (EDA), eye movement, respiration activities, and blood pressure (Charles and Nixon, 2019; Kramer, 1991; Matthews et al., 2015a; 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 get 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). Also, from the review papers summarized by Charles and Nixon (2019), Tao et al. (2019), and Young et al. (2015), there was no a 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 mental workload (Hogervorst et al., 2014; Matthews et al., 2015b; Wickens, 2008). 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). Also, 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., 2015; Wilson and Russell, 1999, 2003; Zhao, Liu, and Shi, 2018). Linear discriminant analysis (LDA), support vector machine (SVM), classification tree and k-nearest neighbor (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 et al. (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 what extent a combination of varied indices 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 characterized by different levels of difficulty. As muscle activity can also affect cognitive attention (Stephenson et al., 2019) and related physiological signals, this kind 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 the reviews (Charles and Nixon, 2019; Tao et al., 2019; Young et al., 2015). Also, different classifiers were conducted to examine how well different indices can be used to 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. They have 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 participants' self-report of stress, nervous and mood. Responses were scored from 4 for 'definitely' to 1 for 'not at all'. All of the participants did not feel stressed, nervous or had a low mood before the experiment. Each of the participants provided written informed consent before the experiment and received financial compensation.

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 (Fig. 1). Signal recording and analysis were conducted on the ErgoLAB human-machine-environment testing cloud platform (Kingfar International Inc., 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 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.

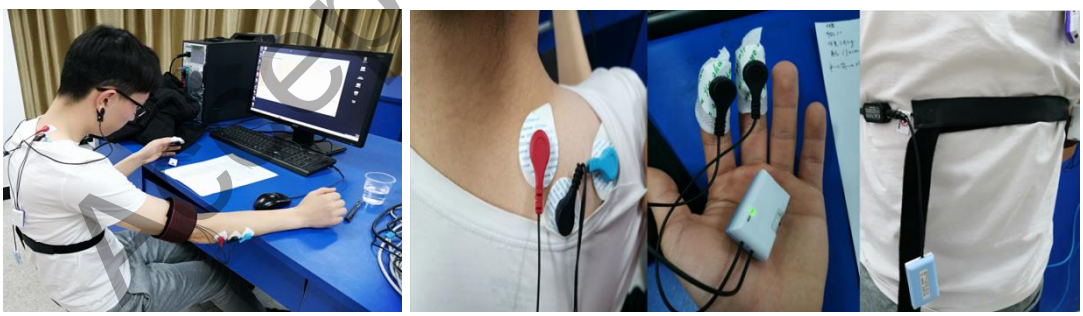

Fig. 1. Sensors placement on a subject.

2.3 Procedure

A simulated computer-based work consisting of mental arithmetic tasks characterized by different levels of difficulty was designed to impose different mental workload 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 measures about mental workload measurement or prediction 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

Behavioral 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 analyzed. 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 normalization process was used to transform the data (Guyon et al., 2006). Here, z-score standardization was applied to the current value of an index (standardization of x is $(x_{curren}-x_{average})/$ (SD of raw data)). The signal under each level of task was divided into eight-time nodes (2 mins before the experiment, 2

min, 4 min, 6 min, 8 min, 10 min, 12 min, and 14 min). 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 analyze 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 analyzed 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. The classification accuracy is defined as the ability to correctly classify positive and negative results and showed in Eq.1.

$$\mathrm{Accuracy} = \frac{TP + TN}{P + N} \times 100 \tag{1}$$

Where TP denotes true positive correctly labeled as high mental workload at the corresponding level, TN denotes true negative correctly labeled as not high mental workload at the corresponding level, P and N denote the count of positive and negative. Precision is defined as the fraction of predictions that are accurate, and recall is defined as the fraction of instances that are accurately predicted, which are shown as Eq. 2 and Eq.3, respectively.

$$Precision = \frac{TP}{TN + FP} \times 100 \tag{2}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \times 100 \tag{3}$$

Where FN denotes false negatives and refers to data points incorrectly labeled as high mental workload at the corresponding level, and FP denotes false positives and refers to data points incorrectly labeled as not high mental workload at the corresponding level.

3. **Results**

After the experiment, three types of data were obtained, 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 analyzed. Then, the relationships between different measures of mental workload were analyzed. 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 behavioral data.

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 $F(2,34)=32.559$, $p<0.001$, $\eta^2=0.657$. Also, paired t-tests showed that there were significant differences between easy and medium ($t = -6.07$, $p < 0.001$) level tasks, between easy and difficult ($t = -6.93$, $p < 0.001$) level tasks, and between medium and difficult ($t = -3.82$, $p = 0.001$) level tasks. Fig. 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.

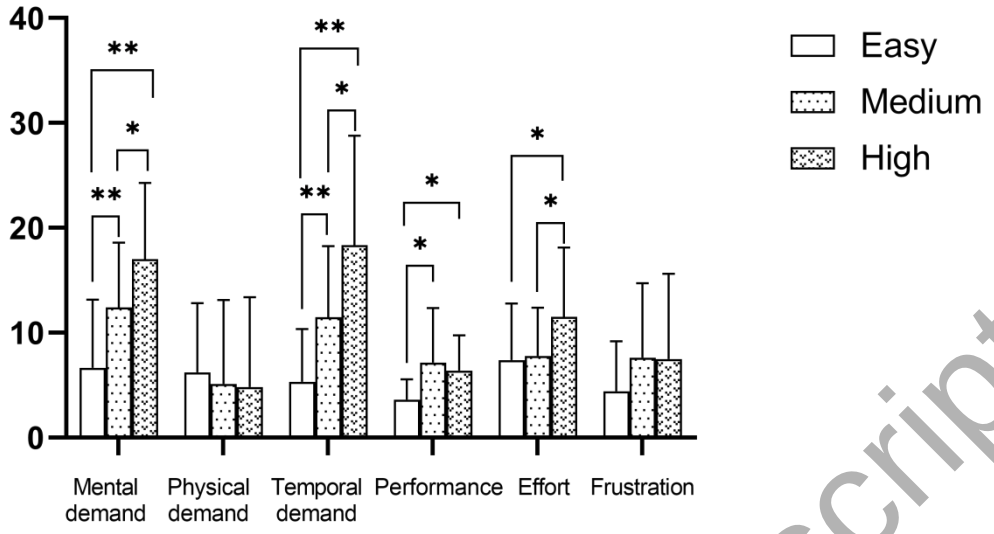

Fig. 2. Comparisons among three tasks with different levels of difficulty from the NASA-TLX subscale. Error bars represent standard error (*p<0.05, **p<0.01)

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 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 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 for 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. Fig. 3 showed that the participants were more accurate and responded faster in easy-level tasks than in medium- and difficult-level tasks and in medium-level tasks than in difficult-level tasks.

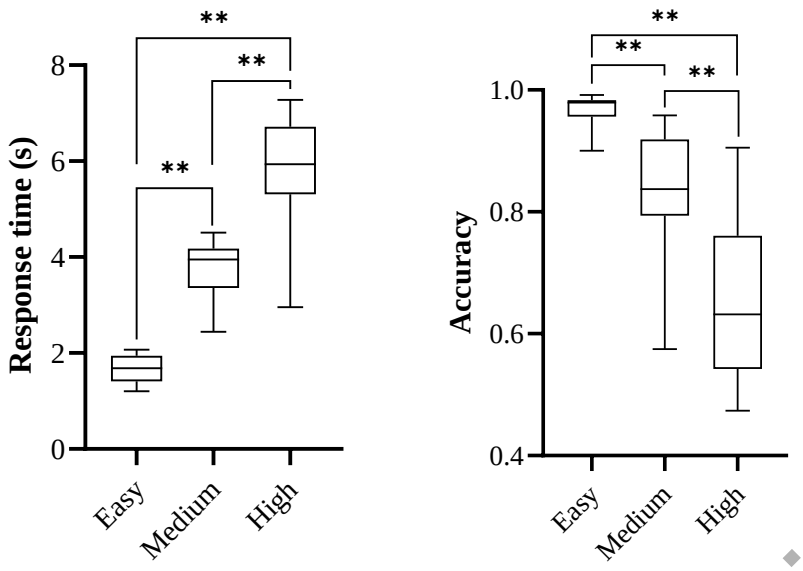

Fig. 3. The comparison of task performance evoked by three tasks with 95% confidence intervals

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 analyzed. 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.

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.

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

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

Tasks / Index	Easy level		Medium level		Difficult level	
	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

Electrodermal Activity

The EDA signals were bandpass filtered at 0.05–500 Hz and digitized at a rate of 64 Hz through the ErgoLAB platform (Beijing Kingfar Technology). A Bandstop 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 to 21 years old, M_{age} = 20.0 years, SD_{age} = 0.97). The repeated ANOVA results showed that there was a significant main effect of task level with $F(2,30) = 7.586$, $p = 0.002$, $\eta^2 = 0.336$, 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 $t(15) = -2.212$, $p < 0.043$. Also, 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 Fig. 4.

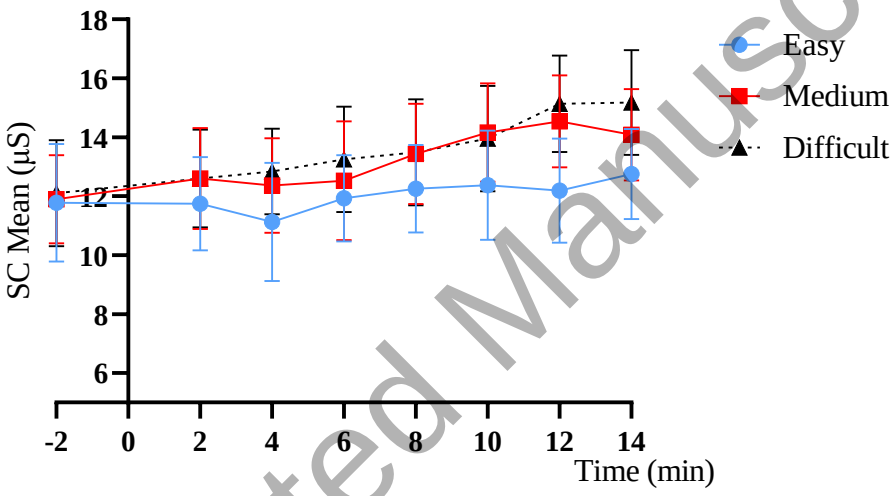

Fig. 4. The comparison of SC means evoked by three tasks

Electrocardiography (ECG) activity

Cardiac activity can be analyzed 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 digitized at a rate of 1024 Hz. The bandstop 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 $t(17) = 0.127$, $p = 0.900$. Also, 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 Fig. 5.

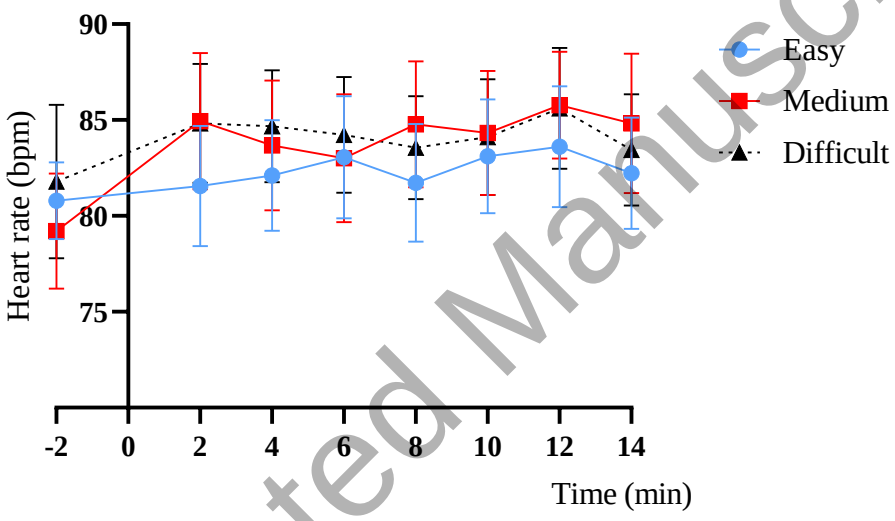

Fig. 5. The comparison of heart rate evoked by three tasks

The repeated-measures ANOVA results of LF/HF showed that there was a significant main effect of task level with $F(2,34) = 5.6$, $p = 0.008$, $\eta^2 = 0.248$. 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 $t(17) = 0.410$, $p = 0.687$. Also, 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 Fig. 6.

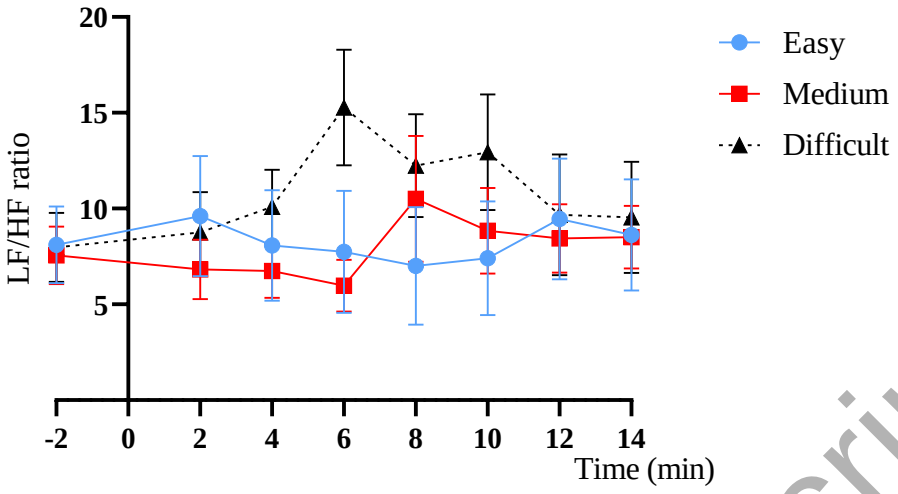

Fig. 6. The comparison of LF/HF evoked by three tasks

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 digitized at a rate of 64 Hz. The bandstop filter was used to eliminate 50 Hz power frequency interference. Baseline denoise was filtered with a cutoff 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 $t(17) = -2.299$, $p = 0.034$. Also, 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 Fig. 7.

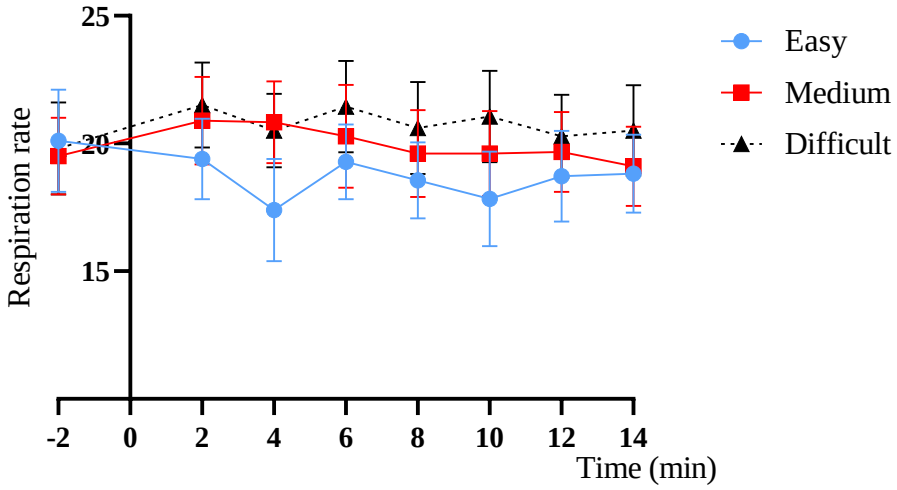

Fig. 7. The comparison of respiration evoked by three tasks

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 bandstop 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. This suggests that the tasks correspond to different levels of difficulty and demonstrates the interest of those features for later classification of the three tasks. All the above physiological indexes and performance data were taken as classifier inputs. The tasks with different levels of difficulty were taken as outputs. A total of 378 data (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 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 were presented here in Table 3. The best performance could reach about 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 decreased a lot when taking a single physiological index as input, i.e., with an accuracy of below 48%. LDA had the lowest accuracy in mental workload recognition. Also, pairwise comparison tests showed the performance by LDA is lower than other models when using all physiological indices as inputs ($p < 0.01$). When using combined signals as inputs, the accuracy of recognition can be increased by about 30% compared to a single physiological index as inputs. Also, in the three kinds of physiological indexes, EDA and ECG achieved similar performance ($p > 0.05$), and had a higher accuracy than respiration as inputs ($p < 0.05$). Also, the pairwise comparison tests showed that the classification accuracy significantly improved when taking both physiological and task performance indices as inputs ($p < 0.01$).

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%

4. Discussion

4.1 General overview of findings

The present study investigates the multimodal measurement of mental workload during simulated computer tasks and identify 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 the ANOVA analysis were selected and entered into the classification models as inputs. Also, we examine how well different variables can 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 nondynamic concept (Jafari et al., 2020; Shuggi et al., 2017). Besides, no difference in the 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 contribute 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 findings of this study, 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 et al., 2005; Finsen et al., 2001; Fournier et al., 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 taking all the physiological indices as inputs could obtain a satisfying accuracy of 78.3% by BPNN optimally. When taking response time as inputs combined with physiological indexes, the accuracy increased to 96.4%. As many researchers pointed out, 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). Also, the high classification accuracies indicate that machine learning methods have great potential to be developed for predicting the task and difficulty levels, and may be used to develop situation-aware recognition systems of the mental workload and even embedded adaptive human-computer interaction platforms.

4.2 Effective indices of mental workload measurement and classification

We proved that 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 easy and middle task difficulty levels. Differences in cognitive demand evoked by the tasks may have contributed to this disparity (Fairclough et al., 2005; Matthews et al., 2015; Mulder et al., 2004).

For EDA and respiration, both of them significantly increased with increasing task difficulty levels. Our results are consistent with the studies by Collet et al. (2014) and Engström et al. (2005), but in one of our previous studies, there was no significant difference between difficult- and medium-level tasks (Ding et al., 2019). From the review of Charles and Nixon (2019), they summarized 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. Also, the changing trend in SC (Fig. 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 et al., 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 et al., 1996; Fairclough et al., 2005; Grassmann et al., 2015). Grassmann 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 analyzed 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. Also, if a given cognitive activity requires extensive application of resources and that activity has to be carried out for long, unbroken periods of time, then it is likely that the activity should induce stress (Hancock and Warm, 1989; Warm, Parasuraman, and Matthews, 2008). From this view, the results of this study could be influenced by stress. The previous studies have pointed out that stress involves the activation of hypothalamus-pituitary-adrenocortical, and which plays an important role in the regulation of various physiological processes, especially for ECG and EDA indexes (Healey and Picard, 2005; Porges, 1991; Sun et al., 2010). According to this, the physiological responses may be affected. Also, 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 taking all physiological indexes as input, an accuracy of 78.3% could be obtained. This accuracy was significantly improved compared to taking single or two physiological indexes as input. In the applied environment, identifying people's mental workload from physiological signals largely rely on the classification accuracy, the access of data collection, and the processing method (Wilson and Russell, 2003). The 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, which means that identifying an operator's mental workload by collecting his/her physiological signals from portable equipment is far from a real

application. Although an accuracy of 96.4% when taking all physiological indexes and response time as inputs, there are many places where response time is not available in general. The limited kinds of physiological indexes extracted and data from shorter time could lead to this low classification accuracy than previous studies, for example, 42 physiological indices in total in the study of Zhao et al. (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 cause stress (Hancock and Warm, 1989; Warm, Parasuraman, and Matthews, 2008). Then physiological signals may be affected by this factor, which also leads to a low classification accuracy of mental workload. Our results also showed that LDA performed worse than the other models for taking all physiological indices as inputs. This liner method is best suitable for data with linear boundaries between different classes and is based on the assumption of the normality of data distribution in each class (Hastie et al., 2001). This may lead to the low accuracy of LDA. Another reason may be that the class boundaries are nonlinear in this classification, in which nonlinear methods could be more suitable (Kolodyazhniy et al., 2011). The classification accuracies indicate that machine-learning methods have great potential to be developed for predicting the task and difficulty levels that the participant has not yet experienced. 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 signals collection will also be affected by the real environment. This

signifies that the process and conclusions cannot be extrapolated to occupational settings. Moreover, the subjects in the experiment were all male university students. As previous studies suggested that 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 taking 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

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) from 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.

References

Van Amelsvoort L G P M, Schouten E G, Maan A C, et al. Occupational determinants of heart rate variability[J]. International Archives of Occupational and Environmental Health, 2000, 73(4):255-262.

Baldwin C L, Penaranda B N. Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification[J]. Neuroimage, 2012, 59(1):48-56.

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

Cao Y, Qu Q, Duffy V G, et al. Attention for Web Directory Advertisements: A Top-Down or Bottom-Up Process?[J]. International Journal of Human–Computer Interaction, 2019, 35(1): 89-98.

Casali J G, Wierwille W W. On the measurement of pilot perceptual workload: a comparison of assessment techniques addressing sensitivity and intrusion issues. Ergonomics, 1984, 27(10), 1033-1050.

Chanel G, Rebetez C, Bétrancourt M, Pun T. Emotion assessment from physiological signals for adaptation of game difficulty. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 2011, 41(6), 1052-1063.

Charles R L, Nixon J. Measuring mental workload using physiological measures: a systematic review[J]. Applied ergonomics, 2019, 74: 221-232.

Cinaz B, Arnrich B, Marca R L, et al. Monitoring of mental workload levels during an everyday life office-work scenario[J]. Personal and Ubiquitous Computing, 2013, 17(2): 229-239.

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

De Rivecourt M, Kuperus M N, Post W J, et al. Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight[J]. Ergonomics, 2008, 51(9): 1295-1319.

De Waard D, Lewis-Evans B. Self-report scales alone cannot capture mental workload[J]. Cognition, Technology & Work, 2014, 16(3):303-305.

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

Engström J, Johansson E, Östlund, Joakim. Effects of visual and cognitive load in real and simulated motorway driving.[J]. Transportation Research Part F: Psychology & Behaviour, 2005, 8(2):97-120.

Fairclough S H, Venables L, Tattersall A. The influence of task demand and learning on the psychophysiological response[J]. International Journal of Psychophysiology, 2005, 56(2):171-184.

Fairclough S H, Venables L. Prediction of subjective states from psychophysiology: A multivariate approach[J]. Biological Psychology, 2006, 71(1):100-110.

Fawcett T. An introduction to ROC analysis. Pattern recognition letters, 2006, 27(8), 861-874.

Finsen L, Søgaard K, Jensen C, et al. Muscle activity and cardiovascular response during computer-mouse work with and without memory demands[J]. Ergonomics, 2001, 44(14): 1312-1329.

Fong A, Sibley C, Cole A, et al. 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, 2010, 54(19):1709-1712.

Fournier L R, Wilson G F, Swain C R. Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: manipulations of task

difficulty and training[J]. International Journal of Psychophysiology, 1999, 31(2): 129-145.

Gopher D, Donchin E. Workload: An examination of the concept. in Handbook of Perception and Human Performance, Vol. 2: Cognitive Processes and Performance, K. R. Boff, L. Kaufman, and P. Thomas, Eds. Oxford, U.K.: Wiley, 1986, pp. 1–49.

Grassmann M, Vlemincx E, Leupoldt A V, et al. Respiratory Changes in Response to Cognitive Load: A Systematic Review[J]. Neural Plasticity, 2016, 2016:1-16.

Grassmann M, Vlemincx E, Leupoldt A V, et al. The role of respiratory measures to assess mental load in pilot selection[J]. Ergonomics, 2015, 59(6):1-9.

Guyon I, Elisseeff A. An Introduction to Feature Extraction [M]. In Feature Extraction; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–25.

Hancock P A. A dynamic model of stress and sustained attention. Human factors. 1989, 31(5):519-37.

Hancock PA, Matthews G. Workload and performance: Associations, insensitivities, and dissociations[J]. Human factors. 2019, 61(3):374-92.

Hancock P A, Szalma J L. Stress and neuroergonomics [M]. In Parasuraman R, Rizzo M (Eds.), The brain at work. Oxford, UK: Oxford University Press, 2006, 195-206.

Halder A, Gao C, Miller M, et al. Oxygen uptake and muscle activity limitations during stepping on a stair machine at three different climbing speeds[J]. Ergonomics, 2018, 61(10): 1382-1394.

Hart S G, Staveland L E. Development of NASA-TLX: Results of empirical and theoretical research [M]. In Hancock P A, Meshkati N. (Eds.), Human mental workload. Elsevier Science, Amsterdam,1998, 139–183.

Hastie T, Tibshirani R, Friedman J. (2009). The elements of statistical learning: data mining, inference, and prediction. Berlin: Springer Science & Business Media.

Healey J A, Picard R W. Detecting stress during real-world driving tasks using physiological sensors. IEEE Transactions on intelligent transportation systems, 2005, 6(2), 156-166.

Hockey G R J, Nickel P, Roberts A C, et al. Sensitivity of candidate markers of psychophysiological strain to cyclical changes in manual control load during simulated process control[J]. Applied Ergonomics, 2009, 40(6):1011-1018.

Hogervorst M A, Brouwer A M, Van Erp J B. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload[J]. Frontiers in Neuroscience, 2014, 8, 322.

Jafari M J, Zaeri F, Jafari A H, et al. Assessment and monitoring of mental workload in subway train operations using physiological, subjective, and performance measures[J]. Human Factors and Ergonomics in Manufacturing & Service Industries, 2020, first published online. DOI: https://doi.org/10.1002/hfm.20831.

Jang E H, Park B J, Park M S, et al. (2015). Analysis of physiological signals for recognition of boredom, pain, and surprise emotions[J]. Journal of physiological anthropology, 34(1), 25.

Jimenez-Molina A, Retamal C, Lira H. Using Psychophysiological Sensors to Assess Mental Workload During Web Browsing[J]. Sensors, 2018, 18,458. doi:10.3390/s18020458

Jorna P G. Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. Biological psychology, 1992, 34(2-3), 237-257.

Kolodyazhniy, V., Kreibig, S. D., Gross, J. J., Roth, W. T., & Wilhelm, F. H. (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.

Kramer A F. Physiological metrics of mental workload: A review of recent progress [M]. In Damos D L (Ed.) Multiple-Task Performance, Taylor and Francis, 1991, 279–327.

Lean Y, Shan F. Brief review on physiological and biochemical evaluations of human mental workload[J]. Human Factors and Ergonomics in Manufacturing & Service Industries, 2012, 22(3):177-187.

Mallat C, Cegarra J, Calmettes C, Capa RL. A curvilinear effect of mental workload on mental effort and behavioral adaptability: an approach with the pre-ejection period[J]. Human factors. 2019, first published online.

Matthews G, Joyner L, Gilliland K, et al. Validation of a comprehensive stress state questionnaire: Towards a state big three. Personality psychology in Europe [J], 1999, 7: 335-350.

Matthews G, Reinerman-Jones L E, Barber D J, et al. The Psychometrics of Mental Workload: Multiple Measures Are Sensitive but Divergent[J]. Human Factors, 2015a, 57(1):125-143.

Matthews G, Reinerman-Jones L, Wohleber R et al. Workload is multidimensional, not unitary: what now?[C]. In: Schmorrow DD, Fidopiastis CM (eds) Foundations of augmented cognition: 9th international conference, AC 2015, held as part of HCI International 2015, Los Angeles, CA, USA, August 2–7, 2015, proceedings. Springer International Publishing, Cham, 2015b, pp 44–55.

Miyake S, Yamada S, Shoji T, et al. Physiological responses to workload change. A test/retest examination[J]. Applied Ergonomics, 2009, 40(6):987-996.

Mulder L J M, de Waard D, Brookhuis K A. Estimating mental effort using heart rate and heart rate variability [M]. In Stanton N A, Hedge A, Brookhuis K, et al. (Eds.), Handbook of human factors and ergonomics methods, CRC Press, 2004, 201–210.

Nixon J, Charles R. Understanding the human performance envelope using electrophysiological measures from wearable technology[J]. Cognition Technology and Work, 2017, 19(5):1-12.

Orlandi L, Brooks B. Measuring mental workload and physiological reactions in marine pilots: Building bridges towards redlines of performance[J]. Applied Ergonomics, 2018, 69:74-92.

Parasuraman, R., & Rizzo, M. (Eds.). (2008). Neuroergonomics: The brain at work. Oxford University Press.

Roscoe A H. Assessing pilot workload: Why measure heart rate, HRV and respiration?[J]. Biological Psychology, 1992, 34(3):259-87.

Ryu K, Myung R. 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, 2005, 35(11), 991-1009.

Shuggi I M, Oh H, Shewokis P A, Gentili R J. Mental workload and motor performance dynamics during practice of reaching movements under various levels of task difficulty. Neuroscience, 2017, 360, 166-179.

Stephenson ML, Ostrander AG, Norasi H, Dorneich MC. Shoulder muscular fatigue from static posture concurrently reduces cognitive attentional resources[J]. Human factors. 2019, first published online, DOI: 10.1177/0018720819852509.

Sun F T, Kuo C, Cheng H T, Buthpitiya S, et al. Activity-aware mental stress detection using physiological sensors. In International conference on Mobile computing, applications, and services, 2010 (pp. 282-301). Springer, Berlin, Heidelberg.

Tao, D., Tan, H., Wang, H., Zhang, X., Qu, X., & Zhang, T. (2019). A Systematic Review of Physiological Measures of Mental Workload. International journal of environmental research and public health, 16(15), 2716.

Tjolleng A, Jung K, Hong W, et al. Classification of a Driver's cognitive workload levels using artificial neural network on ECG signals. Applied ergonomics, 2017, 59, 326-332.

Van Acker B B, Parmentier D D, Vlerick P, Saldien J. Understanding mental workload: from a clarifying concept analysis toward an implementable framework. Cognition, Technology & Work, 2018, 20(3), 351-365.

Veltman J A, Gaillard A W K. Physiological workload reactions to increasing levels of task difficulty.[J]. Ergonomics, 1998, 41(5):656-669.

Viitasalo J H T, Komi P V. Signal characteristics of EMG during fatigue[J]. European Journal of Applied Physiology and Occupational Physiology, 1977, 37(2):111-121.

Wickens C D. Mental workload: assessment, prediction and consequences[C]// International Symposium on Human Mental Workload: Models and Applications. Springer, Cham, 2017: 18-29.

Wilson G F, Russell C A. Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. [J]. Human Factors, 2003, 45(4):635-643.

Wilson G F, Russell C A. Operator Functional State Classification Using Neural Networks with Combined Physiological and Performance Features[J]. Human Factors & Ergonomics Society Annual Meeting Proceedings, 1999, 43(20):1099-1102.

Young M S, Brookhuis K A, Wickens C D, et al. State of science: mental workload in ergonomics[J]. Ergonomics, 2015, 58(1):1-17.

Zhao G, Liu Y J, Shi Y. Real-time assessment of the cross-task mental workload using physiological measures during anomaly detection. IEEE Transactions on Human-Machine Systems, 2018, 48(2), 149-160.