Bio-Medical Materials and Engineering 24 (2014) 2283–2290
DOI 10.3233/BME-141041
IOS Press

Improving pilot mental workload evaluation with combined measures

Xiaoru Wanyan*, Damin Zhuang and Huan Zhang
School of Aeronautics Science and Engineering, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing 100191, China

Abstract. Behavioral performance, subjective assessment based on NASA Task Load Index (NASA-TLX), as well as physiological measures indexed by electrocardiograph (ECG), event-related potential (ERP), and eye tracking data were used to assess the mental workload (MW) related to flight tasks. Flight simulation tasks were carried out by 12 healthy participants under different MW conditions. The MW conditions were manipulated by setting the quantity of flight indicators presented on the head-up display (HUD) in the cruise phase. In this experiment, the behavioral performance and NASA-TLX could reflect the changes of MW ideally. For physiological measures, the indices of heart rate variability (HRV), P3a, pupil diameter and eyelid opening were verified to be sensitive to MW changes. Our findings can be applied to the comprehensive evaluation of MW during flight tasks and the further quantitative classification.

Keywords: Mental workload, physiological measures, comprehensive evaluation, flight simulation, ergonomics

1. Introduction

In the human-machine interaction system of modern aviation, the pilot's role is transforming from a manual operator to a monitor owing to the improved aircraft performance and automation. The pilot must often monitor various indicators simultaneously when carrying out flight missions, and thus the effective capturing of visual information greatly depends on a reasonable allocation of the pilot's limited attention resource [1]. Scholars generally consider mental workload (MW) one of the most important factors influencing the allocation of pilot's attention resources. For example, in urgent situations, some pilots may forget to perform critical tasks or omit important information due to the 'attention narrowing phenomenon' under high MW conditions [2].Therefore, in the design stage of aircraft cockpit's human-machine interface, the accurate evaluation, quantitative classification and even prediction of the pilot's MW play an essential role in optimizing the mental task design of human-machine interface and the allocation of human-machine functions, as well as have important practical significance in preventing aviation accident and ensuring aviation safety [3].

At present, the widely used MW measurements include subjective measure, primary-task performance measure, secondary-task performance measure and physiological measure [4-6]. As multiple abilities of cognitive processing are involved for flight operations, solely relying on one certain single

*Corresponding author: Xiaoru Wanyan, School of Aeronautics Science and Engineering, Beihang University, No. 37 Xueyuan Road, Haidian District, Beijing 100191, China. Tel./Fax:+8601082338163; E-mail: wanyanxiaoru@buaa.edu.cn.

index or method cannot fully reflect the pilot's MW state. Therefore, a comprehensive evaluation method is necessary. In the present study, flight simulation experiment was performed. The differences and sensitivities of indices based on behavioral performance, subjective assessment and physiological measures under different MWs were discussed in order to put forward a comprehensive MW evaluation method. The results can provide reference for the evaluation of pilot's MW and the further quantitative classification.

2. Materials and methods

2.1. Subjects

Twelve highly trained, healthy flying cadets from Beihang University were included in the present study. All of the subjects (male, aged from 23 to 25 years with a mean age of 23.8 years) were right-handed, and possessed normal or corrected to normal vision and normal hearing. All subjects were required to avoid drinking caffeinated or alcoholic beverages, smoking, taking any medication and strenuous exercise for 12 hours before the experiment. In addition, all subjects reported that they slept between 6 and 8 hours the preceding night. Written informed consents of the subjects were obtained before the experiment.

2.2. Experimental tasks

A complete dynamic process of flight simulation, consisting of three phases, taking-off, cruise and landing, was performed by each subject with a flight simulator, as shown in Figure 1. Besides, accomplishing the normal flight operations in the taking-off and landing phases, each subject was instructed to continuously monitor the flight indicators presented on the head-up display (HUD) during the cruise phase. According to the programming, abnormal information could be produced from several flight indicators, such as airspeed, pressure altitude, pitching angle, rolling angle, course angle, etc. The MWs conditions were manipulated by varying the quantity of flight indicators that pilot should monitor. High, medium, low and control MWs (quantity of flight indicators: 9, 6, 3, 0, respectively; duration of abnormal information: 1s; inter-stimulus interval between abnormal information: 2s) were set prior to the experiment. All the subjects were arranged to conduct four flight simulation tasks under the high, medium, low and control MWs respectively, with a one hour break between each two tasks.

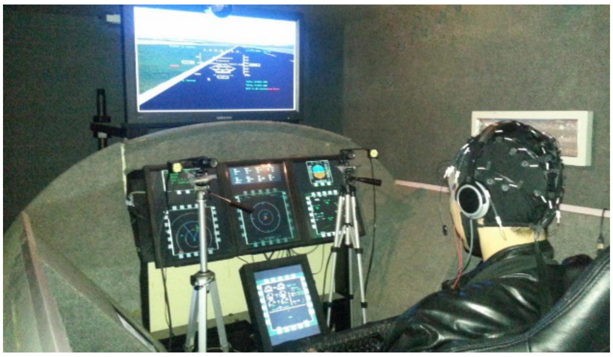

Fig. 1. MW measurement experiment.

The order of the tasks was counterbalanced across the subjects.

2.3. Data recording and processing

During the experiment, the subjects' accuracy rates and response times were automatically recorded as the evaluation indices of behavioral performance by the computer. The recorded data were subjected to one-way repeated measures analysis of variance (ANOVA) using SPSS 13.0 with the MW as the within-subjects factor. An FX-7402 12 channel automatic electrocardiogram (ECG) analysis machine was employed to record the subjects' ECG signals every five minutes. Each subject was required to wear ECG electrodes so that the heart rate (HR) and R-R intervals coefficient of variation (RRCV) could be measured. The displayed range of HR was 20~300 bpm, the detection accuracy was ±2 bpm, the sampling rate was 800 Hz, and the waveform recording speed was set as 25 mm/s. The HR and RRCV data were analyzed by one-way repeated measures ANOVA with the MW as the within-subjects factor. Meanwhile, in order to acquire event-related potential (ERP) P3a from subjects' electroencephalogram (EEG) signals, subjects were asked to wear electrode caps and headphones. Auditory stimuli were presented binaurally through headphones with a novelty oddball paradigm, in which three types of stimuli, 800 Hz frequency tones (80%, 80 dB SPL) as the standard stimuli, 1000 Hz infrequent tones (10%, 80 dB SPL) as the deviant stimuli, and rare, non-repeated and unusual tones (10%) as the novel stimuli, were included. All stimuli were presented with an exposure time of 100 ms and a stimulus onset asynchrony (SOA) of 600 ms. Subjects were instructed to pay attention to the visual task of indicators monitoring, ignoring the auditory stimuli. The EEG signals were recorded by Neuroscan Nuamps amplifier from 30 electrode sites, referenced to the tip of nose. Online band-pass filters were set to 0.1~200 Hz with a sampling rate of 1000 Hz. Impedances were kept below 5 kΩ for the electrodes of interest. After correcting the eye movements, the EEG data were segmented by the epoch of 600 ms, relative to a 100 ms pre-stimulus baseline. Epochs with artifact voltages exceeding ±150μV were rejected before averaging. The P3a-1 was obtained by subtracting the ERPs elicited by the standard stimuli from the deviant stimuli, and the P3a-2 was obtained by subtracting the ERPs elicited by the standard stimuli from the novel stimuli. Three-way repeated measures ANOVAs were conducted for the analyses of the P3a-1 and P3a-2, with the within-subject factors of MW (high; low; medium; control), laterality (left: F3/FC3/C3; middle: FZ/FCZ/CZ; right: F4/FC4/C4) and region (fronto: F3/FZ/F4; fronto-central: FC3/FCZ/FC4; central: C3/CZ/C4). In order to objectively record the eye tracking data of pupil diameter and eyelid opening, the non-contact eye tracker Smart Eye Pro system was employed in the experiment. The calibration accuracies of all the subjects were always better than 1°, and the sampling rate was 60 Hz. One-way repeated measures ANOVA was adopted to analyze the pupil diameter and eyelid opening with MW as the within-subjects factor. In order to make the subjective assessment, the subjects needed to complete the NASA Task Load Index (NASA-TLX) after each task, and the subjective evaluation results were also subjected to one-way repeated measures ANOVA with MW as the within-subjects factor.

Table 1

Accuracy rates and response times under different MWs (average±standard deviation)

MW	Accuracy Rate/%	Response Time/s
High	53.12±16.38	0.76±0.07
Medium	60.71±10.11	0.72±0.05
Low	85.56±6.41	0.69±0.06

3. Results

3.1. Behavioral performance

Table 1 showed the average accuracy rates and response times of the subjects under different MWs. The One-way repeated measures ANOVA indicated significant main effect of MW ($p<0.001$) on accuracy rate at $\alpha=0.05$ significance level. As the MW changed from high, medium, to low, the accuracy rates increased accordingly. Post hoc tests showed significant differences between the average accuracy rates under high and medium MWs ($p=0.029$), as well as medium and low MWs ($p<0.001$). The main effect of MW on response time was also significant ($p<0.001$). This effect manifested as that the response time to abnormal information under high MW was obviously longer than those under medium and low MWs ($p=0.032$, $p=0.001$).

3.2. Subjective assessment

According to NASA-TLX, subjective ratings of the tasks' difficulties under the high, medium, low and control MWs were 74.39±3.69 points, 69.51±4.34 points, 64.75±7.30 points and 60.02±8.53 points, respectively. One-way repeated measures ANOVA showed that the effect of MW on this rating was significant ($p<0.001$) at $\alpha=0.05$ significance level. Specifically, subjective ratings on task difficulties decreased gradually with the decrease of MWs from high, medium, low, to control ($p<0.001$, $p=0.007$, $p=0.026$).

3.3. Physiological measures

Table 2 showed the average values of HR, RRCV-the time domain index of heart rate variability (HRV), pupil diameter and eyelid opening under different MWs. The data were analyzed at $\alpha=0.1$ significance level. The main effect of MW on HR was not significant ($p=0.252$), and HR did not show a regular trend of variation with the MWs. Meanwhile, RRCV showed a decline trend with the increase of MW. The variance analysis indicated that the main effect of MW was significant ($p=0.019$), showing that the RRCV under high MW was obviously lower than the ones under the other MWs ($p=0.053$, $p=0.096$, $p=0.009$). In addition, the RRCV under medium MW was significantly lower than that under the control MW ($p=0.044$). However, the differences among RRCVs under other MWs were not significant ($p>0.1$).

The main effect of MW on pupil diameter was significant ($p=0.076$). The pupil diameter under the medium MW was significantly higher than the ones under the low and control MWs ($p=0.043$, $p=0.002$), but no significant difference was found among those under the other MWs. According to Table 2, with the increase of MW, the pupil diameter increased firstly and then decreased. As for the index of eyelid opening, the main effect of MW on it was also significant ($p=0.097$). The eyelid opening under the high MW was higher than the ones under the low and control MWs ($p=0.053$, $p=0.099$), so was the eyelid opening under the medium MW ($p=0.025$, $p=0.077$). The differences among the eyelid opening under the other MWs were not significant ($p>0.1$). The changing trend of eyelid opening was similar to that of pupil diameter, and that is the eyelid opening increased firstly and then decreased with the increase of MW, as shown in Table 2.

Table 2

Measurement results of physiological indices under different MWs (average±standard deviation)

MWs	ECG Data		Eye Tracking Data	
	HR/bmp	RRCV/%	Pupil Diameter/mm	Eyelid Opening/mm
High	67.62±9.13	4.89±0.65	3.11±0.86	12.53±1.39
Medium	68.85±8.38	5.43±0.84	3.29±0.71	12.78±1.50
Low	67.23±9.36	5.84±1.36	3.10±0.72	12.22±1.42
Control	68.31±8.44	6.34±1.19	3.03±0.66	10.86±3.59

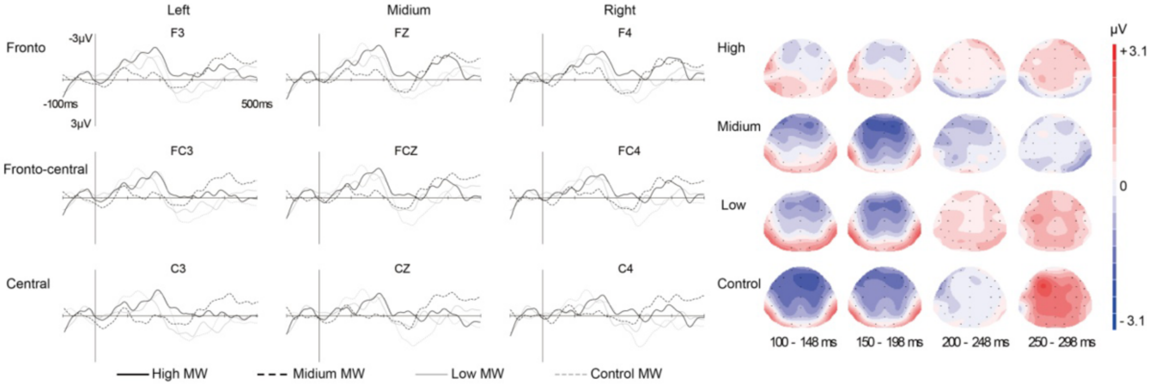

Fig. 2. Grand average waveforms and 2D scalp topographic distributions of P3a-1 under different MWs.

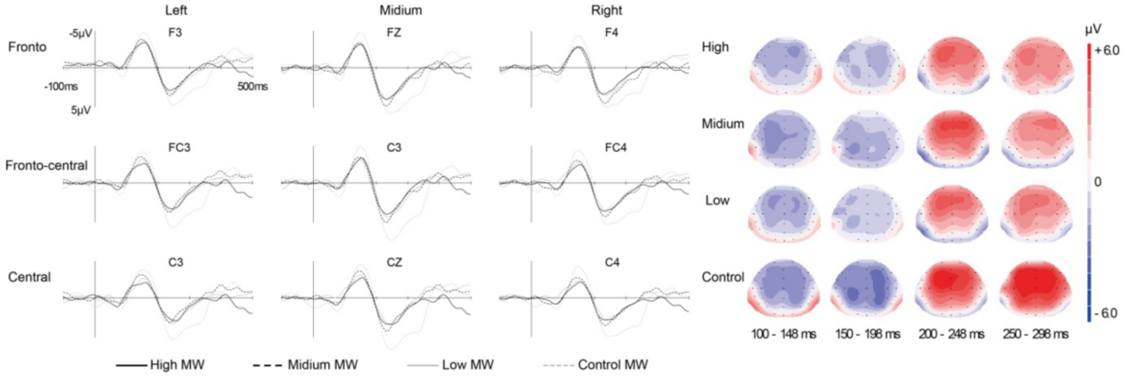

Fig. 3. Grand average waveforms and 2D scalp topographic distributions of P3a-2 under different MWs.

The grand average waveforms and two-dimensional (2D) scalp topographic distributions of P3a-1 under different MWs were presented in Figure 2. The three-way repeated measures ANOVAs revealed that the main effect of MW on the peak amplitude was significant (p=0.049) at α=0.1 significance level, showing that the peak amplitudes of P3a-1 was higher under the medium (1.830 μV, p=0.094), low (2.344 μV, p=0.019) and control (1.882 μV, p=0.050) MWs than those under the high (0.450 μV) MW. The differences among the peak amplitudes under the other MWs were not significant (p>0.1). The main effect of region and laterality were also not noteworthy (p>0.1). The grand average waveforms and 2D scalp topographic distributions of P3a-2 under different MWs were presented in Figure 3. The three-way repeated measures ANOVAs revealed that the main effect of MW on the peak amplitude was significant (p=0.089), showing that the peak amplitudes of P3a-2 was higher under the control (6.727 μV) MW than those under the high (4.154 μV, p=0.075) and low (4.162 μV, p=0.059) MWs. The differences among the peak amplitudes under the other MWs failed to reach significance (p>0.1). The main effect of region on this indicator was also significant (p=0.042), reflecting a fronto-central distribution for the fronto (4.922 μV), fronto-central (5.171 μV) and central (4.574 μV) regions. The main effect of laterality was significant (p<0.001), and showed a middle line distribution, with values of 4.507 μV, 5.495 μV and 4.665 μV for the left hemisphere, middle line and right hemisphere, respectively. No interactive effects were found in the selected time window (p>0.1).

3.4. Correlation analysis

Correlation analyses between the results of subjective measures, performance measures and physiological measures were performed respectively at α=0.05 significance level. The site FCZ with the maximum peak amplitude in the fronto-central region was selected as the representative electrode to test the relativity. Statistical analysis revealed positive correlations between NASA-TLX score and RRCV (r = 0.365, p=0.031), as well as between NASA-TLX score and peak amplitude of P3a-1 (r =0.396, p=0.019). Accuracy rate was negatively correlated with response time (r=-0.624, p<0.001), as well as the peak amplitude of P3a-1 (r=-0.391, p=0.036) and P3a-2 (r=-0.403, p=0.041). In addition, RRCV was positively correlated with the peak amplitude of P3a-1 (r=0.497, p=0.043), while eyelids opening was positively correlated with the peak amplitude of P3a-2 (r=0.401, p=0.038). All of the other correlation analysis results failed to reach significance (p>0.05).

4. Discussion

In this study, as expected, good operation performance was observed under the relatively low MW due to the sufficient available cognitive resources. In contrast, under the relatively high MW, a significant reduction of mean accuracy rate was observed with a prolonged reaction time.

Generally, subjective measures have obvious advantages compared with the other MW measurements. For example, evaluation processes are easy to perform by using evaluation gauges or questionnaires, the data can be easily analyzed, and there is almost no need for assistant equipment. In the present study, NASA-TLX was adopted because of its high sensitivity and diagnosticity [7,8]. The investigation showed that the scores of six dimensions of NASA-TLX (including mental demand, physical demand, temporal demand, effort, performance, frustration level) increased accordingly with the increase of task difficulty, and the main effect of MW on the scores was significant. With increasing amount of target information, the subjects needed to deal with much more information within limited time, which elevated the difficulties of the tasks. Therefore, the subjects' mental, physical and temporal demands increased and more efforts were required. Meanwhile, as the tasks became harder, the subjects also became more frustrated by their unsatisfied operation performances.

Relevant studies show that physical activity is closely related to human heart rate, which rises with the increase of physical exercise intensity. Currently, with the fast development of automatization of modern cockpit, physical fatigue caused by flight control operations has undergone a reduction. In this situation, flight operations, which are mainly based on indicator monitoring, have little influence on pilot's HR, thus making HR a more suitable indicator for evaluating pilots' physical workload when they control the aircraft actively or switch autopilot mode into manual-fly. In the present study, the experiment results indicated an insignificant change of HR under different MWs, which was in agreement with earlier studies [9]. Also, HRV is an important index for assessing the function of autonomic nervous system. As a time domain index, RRCV reflects the distraction of heart rate variability. A series of studies show that HRV can reflect the stress reaction of pilot when experiencing higher MW [10,11]. In our study, HRV was measured during the flight simulation process. The recorded HRV revealed a gradual downward trend of it with the increase of MW in general. This experiment result was consistent with our expectation as well as the research results of related studies [11,12].

In recent years, the eye-tracking method for assessing MW has received great attention. One of the most popular eye-related indices that map mental states is the pupil diameter, which shows a strong link to the task difficulty. In addition, being related to cognitive and emotional states, eyelids opening can also be applied to the study of mental activities [13]. According to the experiment results, both the pupil diameter and eyelid opening tended to increase first and then decrease with the increase of MW. Related studies on the relationship between workload and pupil diameter state that the operator's pupil will enlarge with the workload increase, but possibly shrink in overload state [14]. The decreases of pupil diameter and eyelid opening appeared under the high MW was probably attributable to that the subjects were already experiencing fatigue to some extent. Furthermore, external environment such as lighting condition also has impact on pupil size. Therefore, eye-related indices should be combined with other indices during MW evaluation.

During the flight task, in order to obtain information timely, accurately and comprehensively, the pilot's voluntary and involuntary attentions coexist, and the 'top-down' and 'bottom-up' information processing mechanisms complement each other. Moreover, as involuntary attention often occurs when the surrounding environment changes (e.g. suddenly appeared visual target or alarm sound), the information processing mechanism based on human involuntary attention contributes to the protection of the operator from accidental injury. As an electrophysiological indicator of involuntary attention, P3a is closely related to the automatic orienting response of attention [15]. Moreover, P3a also represents the preferential allocation of cognitive resources to potentially significant events. In the present study, with the increase of MW, the peak amplitude of P3a tended to decline, implying the deterioration of the operator's ability of involuntary attention. In this case, the operator's detection ability of dangerous signal was weakened, thus resulting in potential safety risk.

In available related studies [16], the inconsistent conclusions on the sensitivities of various MW evaluation indices probably resulted from the different experiment tasks designed to induce MWs. Moreover, most of those studies showed a lack of aviation background. Due to the complexity and risk of flight tasks and environment, the actual applicability of such research conclusions needs to be verified in an in-flight environment. In the present study, we focused on the indicator-monitoring task based on the HUD under flight simulation condition. Our work, compared to most related, but less comprehensive studies, combined a total of three measurement methods while at the same time expanded the physiological approach by concurrently using ECG, ERP and eye tracking data to conduct a multi-dimensional and comprehensive evaluation of the MW. Furthermore, P3a, rather than the commonly used P3b, was adopted as MW evaluation index in the current study, thereby helping us to understand the relationship between automatic information processing and MW to some extent. In addition, correlations between subjective measures, performance measures and physiological measures were analyzed in the present study, and based on the correlations, a quantitative classification model of pilot's MW could be further established according to the factor analysis.

In conclusion, the present study showed that the subjective measure based on NASA-TLX and performance measure could reflect the changes of MW ideally. For physiological measures, the indices of HRV, event related potential P3a, pupil diameter and eyelid opening were verified to be sensitive to

MW changes. The proposed comprehensive evaluation method can provide effective reference for the assessment and further quantitative classification of MW related to flight task.

Acknowledgement

This study is supported by National Natural Science Foundation of China (Program Grant No. 71301005) and National Basic Research Program of China (Program Grant No. 2010CB734104).

References

[1] X.R. Wanyan, D.M. Zhuang, H.Y. Wei and J.H. Song, Pilot attention allocation model based on fuzzy theory, Computers & Mathematics with Applications **62** (2011), 2727–2735.

[2] J.B. Noel, K.W. Bauer Jr. and J.W. Lanning, Improving pilot mental workload classification through feature exploitation and combination: A feasibility study, Computers & Operations Research **32** (2005), 2713–2730.

[3] Z.M. Wei, D.M. Zhuang, X.R. Wanyan, C. Liu and H. Zhang, A model for discrimination and prediction of mental workload of aircraft cockpit display interface, Chinese Journal of Aeronautics (2014).(in press)

[4] P. Lehrer, M. Karavidas, S.E. Lu, E. Vaschillo, B. Vaschillo and A. Cheng, Cardiac data increase association between self-report and both expert ratings of task load and task performance in flight simulator tasks: An exploratory study, International Journal of Psychophysiology **76** (2010), 80–87.

[5] K. Ryu and R. Myung, Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic, International Journal of Industrial Ergonomics **35** (2005), 991–1009.

[6] B.Z. Allison and J. Polich, Workload assessment of computer gaming using a single-stimulus event-related potential paradigm, Biological Psychology **77** (2008), 277–283.

[7] S. Rubio, E. Díaz, J. Martín and J.M. Puente, Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods, Applied Psychology **53** (2004), 61–86.

[8] Y.H. Lee and B.S. Liu, Inflight workload assessment: Comparison of subjective and physiological measurements, Aviation, Space, and Environmental Medicine **74** (2003), 1078–1084.

[9] M.A. Bonner and G.F. Wilson, Heart rate measures of flight test and evaluation, The International Journal of Aviation Psychology **12** (2002), 63–77.

[10] J.A. Veltman, A comparative study of psychophysiological reactions during simulator and real flight, The International Journal of Aviation Psychology **12** (2002), 33–48.

[11] M.D. Rivecourt, M.N. Kuperus, W.J. Post and L.J.M. Mulder, Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight, Ergonomics **51** (2008), 1295–1319.

[12] G.F. Wilson, An analysis of mental workload in pilots during flight using multiple psychophysiological measures, The International Journal of Aviation Psychology **12** (2002), 3–18.

[13] L.L.D. Stasi, A. Antolí, M. Gea and J.J. Cañas, A neuroergonomic approach to evaluating mental workload in hypermedia interactions, International Journal of Industrial Ergonomics **41** (2011), 298–304.

[14] W.Y. Kang, X.G. Yuan, Z.Q. Liu and D.Y. Dong, Analysis of relations between changes of pupil and mental workloads, Space Medicine & Medical Engineering **5** (2007), 364–366.

[15] L. Jing, L. Zhao, Gong Jingjing, Chen Changsheng and Miao Danmin, Even-related potential based evidence of cognitive dysfunction in patients during the first episode of depression using a novelty oddball task, Psychiatry Research: Neuroimaging **182** (2010), 58–66.

[16] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia and F. Babiloni, Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness, Neuroscience & Biobehavioral Reviews **44** (2014), 58–75.