Chinese Journal of Aeronautics, (2014),27(5): 1070–1077

CSAA

Chinese Society of Aeronautics and Astronautics
& Beihang University

Chinese Journal of Aeronautics

cja@buaa.edu.cn
www.sciencedirect.com

A model for discrimination and prediction of mental workload of aircraft cockpit display interface

Wei Zongmin, Zhuang Damin *, Wanyan Xiaoru, Liu Chen, Zhuang Huan

School of Aeronautic Science and Engineering, Beihang University, Beijing 100191, China

Received 29 October 2013; revised 3 March 2014; accepted 25 March 2014
Available online 28 September 2014

KEYWORDS

Bayesian Fisher discrimination;
Cockpit;
Display interface;
Heart rate variability;
Mental workload;
NASA_TLX

Abstract With respect to the ergonomic evaluation and optimization in the mental task design of the aircraft cockpit display interface, the experimental measurement and theoretical modeling of mental workload were carried out under flight simulation task conditions using the performance evaluation, subjective evaluation and physiological measurement methods. The experimental results show that with an increased mental workload, the detection accuracy of flight operation significantly reduced and the reaction time was significantly prolonged; the standard deviation of R-R intervals (SDNN) significantly decreased, while the mean heart rate exhibited little change; the score of NASA_TLX scale significantly increased. On this basis, the indexes sensitive to mental workload were screened, and an integrated model for the discrimination and prediction of mental workload of aircraft cockpit display interface was established based on the Bayesian Fisher discrimination and classification method. The original validation and cross-validation methods were employed to test the accuracy of the results of discrimination and prediction of the integrated model, and the average prediction accuracies determined by these two methods are both higher than 85%. Meanwhile, the integrated model shows a higher accuracy in discrimination and prediction of mental workload compared with single indexes. The model proposed in this paper exhibits a satisfactory coincidence with the measured data and could accurately reflect the variation characteristics of the mental workload of aircraft cockpit display interface, thus providing a basis for the ergonomic evaluation and optimization design of the aircraft cockpit display interface in the future.

1. Introduction

With the development of flight automation, the main role played by pilot in the human–machine interaction system of aircraft cockpit has changed from the manual operator to the supervisor of aircraft operational state. The role change has significantly increased the mental workload for pilots. In particular, when encountering special situations during the flight, the pilot will face an extremely strict requirement for processing information, i.e., the pilot is required to process a

* Corresponding author. Tel.: +86 10 82338163.
E-mail address: dmzhaung@buaa.edu.cn (D. Zhuang).

Peer review under responsibility of Editorial Committee of CJA.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.cja.2014.09.002
1000-9361

large amount of flight information instantly and make a decision in response to the situation.[1,2] As a result, high mental workload or even overload may occur, thereby significantly influencing pilots' work efficiency, reliability of flight operation as well as the physiological and psychological health of pilots.[2] Relevant research on accident analysis showed that 60% to 90% of aviation flight accidents and incidents occurred in the flight task with high mental workload intensity and stress level.[3] Therefore, in the design stage of human–machine interface of aircraft cockpit, the accurate evaluation, quantitative classification and even prediction of mental workload of pilot under different display interfaces would play an essential role in optimizing the mental task design of human–machine interface and the allocation of human–machine functions, and have important practical significance in preventing aviation accident and ensuring aviation safety.

Many studies have been carried out on the measurement and evaluation of mental workload on human–machine interface of aircraft cockpit, which mainly employed the subjective evaluation method, performance measurement method (including main task evaluation and sub-task evaluation methods) and physiological evaluation method.[4] Studies have shown that different evaluation methods should be applied to different task situations and mental workload levels, and it is unrealistic to attempt to comprehensively reflect the mental workload conditions under different task situations by using one indicator. Therefore, using multiple techniques to comprehensively evaluate mental workload is a reasonable method as an alternative to the single method or index-based evaluation. Meanwhile, the multi-dimensional characteristic of mental workload also emphasizes the necessity of comprehensive evaluation. In recent years, some researchers have employed multi-index comprehensive evaluation method to study the measurement of mental workload related to flight.[5,6] These studies were mostly based on the results of single-index measurement and applied certain modeling technique to realize comprehensive evaluation.[6–8] At present, the modeling techniques commonly used in this field mainly include factor analysis, regression analysis and artificial neural network modeling, etc.[9] Compared with other modeling techniques, Bayesian Fisher discrimination analysis method can realize class discrimination and prediction, effectively preserve the selected indexes, prevent information loss, and obtain a stable discrimination result.[10] In this study we employed the Bayesian Fisher discrimination analysis method for theoretical modeling.

In the present study, we apply the multi-index comprehensive evaluation method to mental workload measurement. Meanwhile, the comprehensiveness of NASA_TLX scale in evaluating mental workload,[11] the non-invasive and instantaneous characteristics of electrocardiogram (ECG) measurement technology,[12,13] and the directness of performance measurement were taken into account.[4] In combination with the characteristics of abnormal information processing under special flight conditions, we carried out the experiment using a flight simulator, and studied the diagnostic of three types of evaluation indexes, namely performance measurement (reaction time and accuracy), subjective evaluation (NASA_TLX scale) and physiological measurement (heart rate and HRV) to the mental workload variation in the flight operation process. After that, based on the evaluation results of selected metal workload evaluation indexes, the discrimination model of mental workload state was established using Bayesian Fisher discrimination method to realize the scientific evaluation, quantities classification and prediction of the mental task in the design of human–machine interface of aircraft cockpit.

2. Experimental measurement

2.1. Subjects

Sixteen flying cadets from Beihang University were recruited to participate in the present study, ranging in age from 21 to 28 years (age = 24.6 ± 3.2 years). All subjects were healthy, right-handed, with normal or corrected vision and normal color vision. For ensuring the objectivity of experimental ECG data, all subjects were asked to refrain from caffeine, alcohol, tobacco, and drugs, and from any vigorous physical activity 12 h before the experiment. They were also required not to take any cold food or do any intense exercise, and report no subjective discomfort 1 h before the experiment. Despite of the good training at simulated flight operations, prior to performing the experimental task, each subject was given instructions about the task and completed a training session to insure that he was well acquainted with the task procedure, operation and requirements.

2.2. Equipment and environment

The subjects were tested in a flight simulator located in Beihang University. Equipment and environment set up for the experiment were shown in Fig. 1. Throughout the experiment, the cockpit door was closed and the experimental environment was kept quiet. The experimental environment inside the cockpit was with favorable lighting conditions, stable temperature (25 ± 2) °C and low noise 20–30 dB. The experiment did not begin until subjects were with a steady heartbeat after sitting for at least 5 min in the laboratory bench.

2.3. Experimental task

The subjects were asked to perform the whole dynamic process of flight simulation in a flight simulator, including take-off, climb, cruise, approach and landing. Each flight simulation task lasted for 13 min, and mental workloads were manipulated in separate conditions by adjusting the quantity of flight indicators and refresh frequencies (present time and interval

Fig. 1 Equipment and environment set up for the experiment.

Table 1 Scope set up for abnormal flight indicators.

Flight parameter	Abnormal flight indicator	Mental workload			
		Baseline	Low	Moderate	High
Pitch	Exceed 20°	0	1	1	1
Indicated air speed	Exceed 740.8 km/h	0	1	1	1
Altitude	Exceed 3.048 km	0	1	1	1
Heading	Exceed 50°	0	0	1	1
Roll	Exceed 20°	0	0	1	1
Rudder	Abnormal	0	0	1	1
Aileron	Abnormal	0	0	0	1
Landing gear	Abnormal	0	0	0	1
Engine status	Abnormal	0	0	0	1

Note: "0" denotes that under the mental workload level, the flight indicators is normal; "1" denotes that under the mental workload level, the flight indicators is randomly abnormal.

time) of abnormal information. The duration of abnormal information was 2 s and inter-stimulus interval between abnormal information was random. During the simulation flight, each subject was instructed to monitor the flight indicators presented on the head-up display, and had to detect and respond to abnormal information quickly and accurately. The scope set up for abnormal flight indicators and the monitoring requirements under different mental workloads were shown in Table 1.

2.4. Experimental procedure

In the present study, mental workloads were manipulated into three levels, including high, moderate, and low. Prior to the three levels mental workloads tasks, each subject completed a normal flight simulation task as the baseline. Within-subject factorial design was implemented in the experiment, and all the subjects completed the flight simulation task at the three levels of the mental workloads, respectively. An experimental design method, similar to that of Latin square design, was adopted to counterbalance the sequence of the flight simulation tasks to reduce the effects of sequence to the experiment results.[14] In order to record the ECG data, all the participants were asked to wear ECG electrodes throughout the study. After each session, each subject was instructed to take a 30 min rest, meanwhile completed self-report assessments of mental workload using the NASA_TLX.

2.5. Data recording and analysis

2.5.1. Performance data recording

The accurate operation rate and reaction time (the time interval between the occurrence of the abnormal information and correct responding) as indicators of performance evaluation were automatically recorded by the system through computer programming.

2.5.2. ECG data recording

FX-7402 12-channel automatic analysis of ECG machine was adopted to synchronously record the ECG signals. The ECG data recorded included the heart rates of subjects measured every 5 min, time series during R-R intervals, ECG within this period and the electrode placement arranged as lead II. The heart rate value range was 20–300 beat per minute (bpm), the heart rate detection accuracy was ± 2 bpm, the sampling frequency was 0.05–150 Hz, and the waveform recording speed was 25 mm/s.

Relevant studies showed that both heart rate and heart rate variability (HRV) indexes can effectively reflect the different levels of mental workload.[15–17] However, there was some limitation for using frequency-domain index of HRV to reflect the physiological change, because it is affected by the length of data extraction period, and the essence of physiological change reflected still needs further study.[18] Besides, relevant studies also indicated that within a certain period of time (5 min), there is a significant correlation between time-domain related indexes and frequency-domain related indexes in R-R interval.[18] Therefore, the present study identified the time series in R-R period as a key index of ECG signal measurement. It is assessed by heart rate (HR) and analyzed by HRV in the time domain, including mean heart rate (mean HR), count of normal R-R intervals (RRI count), mean of normal R-R intervals (mean RRI), maximum of normal R-R intervals (maximum RRI), minimum of normal R-R intervals (minimum RRI), the ratio of the maximum RRI and minimum RRI (max/min RRI), and standard deviation of normal R-R intervals (denoted "SDNN").

2.5.3. Subjective data recording and analysis

In order to eliminate the influence of short-term memory, the subjects were asked to complete NASA_TLX scale within 30 min after they completed each of the three (high, moderate and low) mental workload flight tasks.[19] The NASA_TLX uses six dimensions to assess mental workload, namely detailed description of the mental demand, physical demand, temporal demand, performance, effort and frustration, for each dimension, are provided.[20,21] For the convenience for the subjects to accurately and effectively complete subjective evaluation, in the present study, the NASA_TLX scale was presented in numerical value, i.e. scoring from 0 to 100, with 0 representing no effort and 100 representing maximum effort. First, a score (from 0 to 100) was obtained on each dimension according to the subjects' subjective feelings on the flight related mental workload. Then, a paired comparison task was performed for all pairs of the six dimensions, which required the subjects to choose which dimension had a greater relevance to the overall mental workload. After that, each of the six dimensions was given a specific weight according to the number of times that each dimension was chosen in pared comparison. The final mental workload score was obtained by multiplying each individual dimension scale score by its respective weight and dividing the total score of all dimensions by 15 (the total number of paired comparisons). Repetitive measure analysis of variance (ANOVA) was employed for the analysis of the above data by using SPSS 17.0 statistical package.

3. Experimental results

3.1. Result of performance measurement

At three different mental workload levels (high, moderate and low levels), the accurate operation rate and reaction time of subjects towards abnormal flight information were shown in Table 2. Single-factor ANOVA showed there were significant ($P < 0.001$) main effects of mental workload for both accurate operation rate and reaction time. As the mental workload level increased, the accuracy rate of subjects decreased successively ($P < 0.001$), and the reaction time increased successively ($P < 0.001$).

3.2. Result of subjective evaluation

Result of NASA_TLX was also shown in Table 2. Result of the single-factor repeated measure ANOVA suggested a remarkable ($P < 0.001$) main effect of mental workload. With the increase of mental workload, the scores of NASA_TLX gradually increased ($P < 0.001$). Result of paired comparison showed that the subjective mental workload score at low mental workload level was obviously lower ($P < 0.001$) than that at moderate mental workload level which was in turn obviously lower ($P < 0.001$) than that at high mental workload level.

3.3. Result of ECG evaluation

Table 2 provided the results of various HR and HRV indexes of subjects at different mental workload levels (baseline, low, medium, and high levels).

Seen from Table 2, as mental workload increased, Mean HR, RRI count and max/min RRI presented an increasing trend, while maximum RRI, minimum RRI and SDNN presented a decreasing trend. Result of repeated measure ANOVA showed that at four different mental workload levels, only maximum RRI and SDNN revealed a significant difference.

For the maximum RRI index, the result of the single-factor repeated measure ANOVA showed a remarkable ($P = 0.032$) main effect of mental workload. The result of a further paired comparison suggested that the maximum RRI value at baseline mental workload level was significantly ($P = 0.032$) higher than that at high mental workload level, and the maximum RRI value at low mental workload level was significantly ($P = 0.040$) higher than that at high mental workload level. Paired comparison among maximum RRI values at other mental workload levels showed no significant ($P > 0.05$) difference.

For the SDNN index, the single-factor repeated measure ANOVA showed a remarkable ($P < 0.001$) main effect of mental workload. The result of a further paired comparison suggested that the SDNN value at baseline mental workload level was significantly higher than those at low ($P = 0.033$), moderate ($P < 0.001$) and high ($P = 0.001$) mental workload levels; the SDNN value at low mental workload level was significantly higher than those at moderate ($P = 0.001$) and high ($P = 0.006$) mental workload levels; the SDNN value at moderate mental workload level was higher than that at high mental workload level, however, no significant ($P = 0.385$) difference was observed.

Therefore, the SDNN index and maximum RRI index in HRV are indexes sensitive to mental workload change, while the SDNN index demonstrates a better diagnostic to different mental workloads than maximum RRI index and can be further used for division of different mental workload levels.

4. Theoretical modeling

4.1. Modeling methods

Based on the results of experimental measurement, the Bayesian Fisher discrimination analysis method was employed to construct the mental workload discrimination model of the aircraft cockpit display interface and determine the mental workload level of the display interface. Bayesian Fisher's linear discrimination analysis method is a typical discrimination method for data classification.[10] Based on classification and feature variables of the observations, this method aims to

Table 2 Means and standard errors of performance measures, NASA_TLX measures and physiological measures.

Measure	Mental workload			
	Baseline	Low	Moderate	High
Accuracy (%)		97.49 ± 2.80	81.09 ± 6.86	73.12 ± 6.05
Respond time (ms)		769.20 ± 63.99	969.79 ± 54.26	1045.90 ± 54.63
NASA_TLX total score		55.02 ± 10.20	65.63 ± 6.96	75.41 ± 7.05
Mean HR (bpm)	74.88 ± 10.90	74.88 ± 11.03	76.19 ± 12.76	78.06 ± 13.15
RRI count	374.00 ± 56.71	375.31 ± 55.39	382.06 ± 64.31	390.75 ± 64.45
Maximum RRI	958.50 ± 152.29	955.50 ± 151.18	925.00 ± 144.44	906.50 ± 143.56
Minimum RRI	660.00 ± 80.89	648.50 ± 119.70	642.75 ± 167.78	614.75 ± 153.48
Mean RRI	796.75 ± 105.22	801.44 ± 112.56	794.63 ± 120.76	770.38 ± 110.27
Max/min RRI	1.45 ± 0.12	1.51 ± 0.41	1.56 ± 0.67	1.60 ± 0.68
SDNN	53.38 ± 17.35	49.06 ± 18.53	43.31 ± 18.22	40.88 ± 19.34

optimize classifications and reduce the feature dimensions. In the process of analysis, it projects the observations to lower dimensional space, following the direction of maximizing the ratio of the between-class variance to the within-class variance. Its linear discrimination function is

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n \tag{1}$$

where y refers to the value of the observation in the lower dimensional space; $x_1, x_2, \ldots, x_n$ denote the feature variables of the observation; $a_1, a_2, \ldots, a_n$ refer to discrimination coefficient of each variable.

4.2. Establishment of the model and instructions

In order to ensure the comprehensiveness of the discrimination, the general discrimination analysis method (all-factor analysis method) was employed in the present study, i.e., the discrimination model includes flight operation performance, NASA_TLX subjective evaluation and the time-domain index SDNN of HRV. The discriminating model constructed in this paper is shown in Fig. 2 and the discriminating functions are

$$y_1 = 1.019x_1 + 1.010x_2 + 574.625x_3 + 601.659x_4 - 568.158 \tag{2}$$

$$y_2 = 1.106x_1 + 1.196x_2 + 622.427x_3 + 571.071x_4 - 597.648 \tag{3}$$

$$y_3 = 1.174x_1 + 1.418x_2 + 633.388x_3 + 549.668x_4 - 610.753 \tag{4}$$

where y_1, y_2, y_3 represent the discriminating function value of the low, moderate and high levels of mental workloads respectively, and x_1 represents the SDNN value, x_2 the score of NASA_TLX, x_3 the reaction time of processing abnormal information, and x_4 the accuracy operation of processing abnormal information during flight. According to the values of x_1, x_2, x_3 and x_4, the values of y_1, y_2, y_3 were calculated and compared. If the y_1 value is maximum, it considers that participants are at a low level of mental workload. If the y_2 value is maximum, it means that participants are at a moderate level of mental workload. If the y_3 value is maximum, it considers that participants are at a high level of mental workload.

Fig. 2 shows the distribution of all experimental samples at the three different levels of mental workloads in the constructed Bayesian Fisher discrimination model. As seen in this figure, samples at different levels of mental workload were relatively concentrated, respectively, which suggests that the discrimination effect of the model was satisfying. By comparison, the disparity of distribution was the largest between samples at low level and high level mental workload, followed by the samples at low level and moderate level mental workload, and the distribution disparity between samples at moderate and high level mental workload was the smallest.

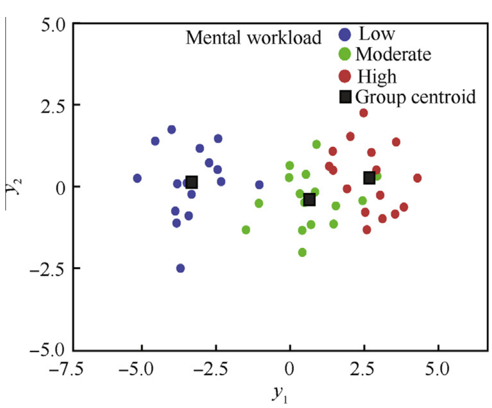

Fig. 2 Discriminating model.

4.3. Validity check of the model

There are general two methods to check the discrimination and prediction accuracy of Bayesian Fisher discrimination function, i.e., the original validation and cross validation methods. For both methods, the higher the accuracy level of discrimination and prediction, the better the constructed discrimination function.[10] To ensure the effectiveness, both the original validation and cross validation methods were employed to check the discrimination and prediction accuracy of the constructed Bayesian Fisher discrimination function. In the case of the original check method, the 48 groups of subject sample data (3 levels data from each of the 16 subjects) measured were substituted back into the constructed discrimination function to evaluate the accuracy level of discrimination and prediction, and the check results are shown in Table 3. In the case of the cross check method, the discrimination model was constructed on the basis of 47 groups of sample data, and used to predict the variable value of the rest one group of sample data, all the samples would go through the circular check once in succession, 48 times in total, and the check results are also shown in Table 3.

It could be known from the comparative results of Table 3 that when employing the general discrimination analysis method, the average discrimination and prediction accuracies of original check method and cross check method were respectively 89.58% and 85.42%. Specifically, the discrimination and prediction accuracies between low workload and other workloads were both 93.75%; the discrimination and prediction accuracies between moderate workload and other workloads

Table 3 Results of predicted mental workload accuracy.

Method	Mental workload	Predicted mental workload accuracy (%)			
		Low	Moderate	High	Total
Original	Low	93.75	6.25	0	100
	Moderate	0	87.50	12.50	100
	High	0	12.50	87.50	100
Cross-validated	Low	93.75	6.25	0	100
	Moderate	12.50	75.00	12.50	100
	High	0	12.50	87.50	100

were respectively 87.50% and 75.00%; the discrimination and prediction accuracies between high workload and other workloads were both 87.50%; the discrimination and prediction accuracy between moderate workload and high workload was slightly lower than that on low workload, which are consistent with the single-factor repeated measure ANOVA conducted in the previous section. Such discrimination results are consistent with the study conclusions drawn by Fishel et al.[22]

5. Discussion

5.1. Differentiation of the three types of evaluation indexes among different mental workloads

In the present study, the difficulty of the flight task was changed to control the mental workload level and measure the three types of evaluation indexes of the subjects under various mental workload levels, i.e., flight operation performance (including accuracy and reaction time), physiological indexes (mean HR and six indexes of HRV) and subjective evaluation (NASA_TLX scale). The relationship between mental workload and various evaluation indexes was also explored on such basis, the results of which show that four indexes, i.e., flight operation accuracy, reaction time, SDNN and NASA_TLX scale, were significantly sensitive to the change of flight task-related mental workload.

Some studies demonstrate that both HR and HRV-related indexes can effectively reflect the mental workload level of flight.[23,24] However, the results of the present study show that HR detection might not be able to effectively reflect the mental workload, while HRV detection could be able to effectively reflect the mental workload, which is consistent with the conclusions drawn by Muth et al.[17] This might have been caused by the fact that the main factor influencing HR is physical workload, while the experimental task of the present study is to induce the occurrence of mental workload. The present study also explores the differentiation of the forementioned six time-domain indexes of HRV among different flight task-related mental workloads, the results of which showed that only the time-domain index SDNN was significantly sensitive to the change of flight task-related mental workload, as specifically demonstrated by the progressive decrease of the SDNN value with the increase of mental workload. Such a conclusion is consistent with the results obtained by both DiDomenico and Nussbaum et al. in studying the influence of various flight operation tasks on mental workload and performance,[12] and also consistent with the results obtained by Lehrel et al. in their studies on Boeing 737-800 simulator,[11] which suggests that the time-domain index SDNN can effectively evaluate the mental workload. However, compared with the two studies, the present study focuses on mental workload, and comprehensively compares 6 time-domain indexes of HRV to confirm the effectiveness of SDNN in evaluating the mental workload.

5.2. Comparison of single assessment index and multidimensional synthetic assessment

Each single measurement index of the subjects was extracted in three different mental workload states to discriminate the mental workload, and the comprehensive evaluation model based on the three types of measurement indexes was also employed for the discrimination of mental workload. The Bayesian Fisher discrimination analysis method was employed to determine the discrimination and prediction accuracies of mental workload level by both approaches respectively under the corresponding experimental conditions. All the results are shown in Table 4.

The verification by the original check method shows that the comprehensive evaluation model had the highest discrimination and prediction accuracy (89.58%), followed in succession by reaction time index (81.25%), accuracy index (77.08%), NASA_TLX scale (64.58%) and physiological indexes SDNN (39.58%). The verification by the cross check method shows that the comprehensive evaluation model had the highest discrimination and prediction accuracy (85.42%), followed in succession by reaction time index (79.17%), accuracy index (77.08%), NASA_TLX scale (64.58%) and physiological indexes SDNN (39.58%). As indicated by the comparative results of the two check methods, the multi-index-based comprehensive evaluation model had a higher overall accuracy in the discrimination and prediction of mental workload level than that of any single index, which suggests that the multidimensional comprehensive evaluation model is more effective than any single index in the discrimination and prediction of mental workload level. However, when the three types of single indexes were employed respectively, the reaction time index had the highest accuracy in the discrimination and prediction of mental workload level.

Table 4 Results of single assessment index and multidimensional synthetic assessment.

Validate method	Assessment index	Predicted mental workload accuracy (%)			
		Low	Moderate	High	Average
Original	Multidimensional	93.75	87.50	87.50	89.58
	SDNN	37.50	12.50	68.75	39.58
	NASA_TLX total score	68.75	43.75	81.25	64.58
	Respond time	100.00	68.75	75.00	81.25
	Accurate	93.75	62.50	75.00	77.08
Cross-validated	Multidimensional	93.75	75.00	87.50	85.42
	SDNN	37.50	12.50	68.75	39.58
	NASA_TLX total score	68.75	43.75	81.25	64.58
	Respond time	93.75	68.75	75.00	79.17
	Accurate	93.75	62.50	75.00	77.08

5.3. Implications and limitations of this research

By means of setting the abnormal posture recovery task of the human–machine interface of aircraft cockpit in the dynamic flight process, the present study is devoted to conducting the flight simulator-based experiment and studying the differentiation of the three types of evaluation indexes, i.e., performance measurement (including accuracy and reaction time), subjective evaluation (NASA_TLX scale) and physiological measurement (HR and HRV), among different mental workloads of pilots in the flight operation process; then the screened mental workload evaluation indexes were employed to construct the discrimination and prediction model of mental workload state based on the Bayesian Fisher discrimination method. The significance of the present study lies in screening out the corresponding sensitive indexes through experimental measurement and then employing the Bayesian Fisher discrimination method to gradually establish the comprehensive discrimination and prediction method of mental workload change in the flight operation process. The method used in the present study can be helpful for the relatively accurate classified quantification of the evaluation and prediction of mental task design in the human–machine interface design of aircraft cockpit. However, the discrimination of mental workload evaluation indexes and their characteristics of change may vary with the nature of operations and present different situations. So, when employing the discrimination and prediction modeling method proposed in the present study in practice, the discrimination and prediction model constructed here should be correspondingly adjusted according to the nature of the flight operation task concerned.

However, there are still some limitations in the present study. Firstly, there are some differences between our subjects and experienced pilots. Secondly, there are certain differences between a simulated flight and a real flight, and between flight stress task setting and a real flight situation. Given that the above factors may all influence the precision of the prediction model to certain extents in subsequent studies more realistic flight environments and flight tasks should be adopted for mental workload measurement of pilots to establish a more rational model for the discrimination and prediction of mental workload of the aircraft cockpit display interface.

6. Conclusions

At present, the common modeling methods for comprehensive evaluation of mental workload in flight include factor analysis, regression analysis and artificial neural network modeling, etc. This study constructs a new comprehensive evaluation model based on the Bayesian Fisher discrimination and classification method, by designing flight simulation tasks at different levels of mental workloads. This model is rarely used in this field before, and our study proved that it could effectively discriminate and predict the levels of mental workload, preserve the selected indexes by avoiding information loss, and obtain a stable discrimination result. The specific conclusions are shown as follows:

(1) During the dynamic process of flight simulation experiments, mean HR is not sensitive to the change of mental workload. Among all its six indexes (RRI count, maximum RRI, minimum RRI, mean RRI, max/min RRI, and SDNN) of HRV, only SDNN is sensitive to the flight related mental workload change, which is significantly decreased as the mental workload increased.
(2) The dynamic flight simulation experiments show that four indexes, i.e., flight operation accuracy, reaction time, SDNN and NASA_TLX scale, are significantly sensitive to the change of flight task-related mental workload.
(3) The verification by the original check method show that the comprehensive evaluation model has the highest discrimination and prediction accuracy (89.58%), followed in succession by reaction time index (81.25%), accuracy index (77.08%), NASA_TLX scale (64.58%) and physiological indexes (39.58%).
(4) The verification by the cross check method show that the comprehensive evaluation model has the highest discrimination and prediction accuracy (85.42%), followed in succession by reaction time index (79.17%), accuracy index (77.08%), NASA_TLX scale (64.58%) and physiological indexes (39.58%).

Acknowledgements

The authors would like to thank the anonymous reviewers for their critical and constructive review of the manuscript. This study was supported by the National Basic Research Program of China (No. 2010CB734104).

References

1. Wei HY, Zhuang DM, Wanyan XR, Wang Q. An experimental analysis of situation awareness for cockpit display interface evaluation based on flight simulation. *Chin J Aeronaut* 2013;**26**(4):884–9.
2. Wei ZM, Wanyan XR, Zhuang DM. Measurement and evaluation of mental workload for aircraft cockpit display interface. *J Beijing Univ Aeronaut Astronaut* 2014;**40**(1):86–91 [Chinese].
3. Hollands JG, Wickens CD. *Engineering psychology and human performance*. New Jersey: Prentice Hall; 1999. p. 459–71.
4. Cain B. A review of the mental workload literature. Toronto: Defence Research and Development Toronto (Canada); 2007 Jul. Report No.: ADA474193.
5. Miyake S. Factors influencing mental workload indexes. *J Uoeh* 1997;**19**(4):313–25.
6. 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. *Int J Ind Ergon* 2005;**35**(11):991–1009.
7. Noel JB, Bauer KW, Lanning JW. Improving pilot mental workload classification through feature exploitation and combination: a feasibility study. *Comp Oper Res* 2005;**32**(10):2713–30.
8. Grootjen M, Neerincx M, Veltman J. Cognitive task load in a naval ship control centre: from identification to prediction. *Ergonomics* 2006;**49**(12–13):1238–64.
9. Li JB, Xu BH. Synthetic assessment of cognitive load in human–machine interaction process. *Acta Psychol Sin* 2009;**41**(1):35–43 [Chinese].
10. Yang WZ, Zhang T. *SPSS statistical analysis and industry application cases explanation*. Beijing: Tsinghua University Press; 2011. p. 163–7 [Chinese].
11. Lehrer P, Karavidas M, Lu SE, Vaschillo E, Vaschillo B, Cheng A. Cardiac data increase association between self-report and both

expert ratings of task load and task performance in flight simulator tasks: an exploratory study. *Int J Psychol* 2010;**76**(2):80–7.
12. DiDomenico A, Nussbaum MA. Effects of different physical workload parameters on mental workload and performance. *Int J Ind Ergon* 2011;**41**(3):255–60.
13. Porges SW, Byrne EA. Research methods for measurement of heart rate and respiration. *Biol Psychol* 1992;**34**(2):93–130.
14. Ding G, Li T. *SPSS statistical tutorial—from study design to data analysis*. Beijing: China Machine Press; 2005. p. 166–9 [Chinese].
15. Dahlstrom N, Nahlinder S. Mental workload in aircraft and simulator during basic civil aviation training. *Int J Aviat Psychol* 2009;**19**(4):309–25.
16. Backs RW. Going beyond heart rate: autonomic space and cardiovascular assessment of mental workload. *Int J Aviat Psychol* 1995;**5**(1):25–48.
17. Muth ER, Moss JD, Rosopa PJ, Salley JN, Walker AD. Respiratory sinus arrhythmia as a measure of cognitive workload. *Int J Psychophysiol* 2012;**83**(1):96–101.
18. Fu CY. Research on physiological and eye movement characteristics of driver under fatigue condition [dissertation]. Harbin: Harbin Institute of Technology; 2011 [Chinese].
19. Karavidas MK, Lehrer PM, Lu SE, Vaschillo E, Vaschillo B, Cheng A. The effects of workload on respiratory variables in simulated flight: a preliminary study. *Biol Psychol* 2010;**84**(1):157–60.
20. Hart SG, Staveland LE. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock N, Meshkati N, editors. *Human mental workload*. Amsterdam: North-Holland; 1988. p. 139–83.
21. Rubio S, Díaz E, Martín J, Puente JM. Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods. *Appl Psychol* 2004;**53**(1):61–86.
22. Fishel SR, Muth ER, Hoover AW, Gugerty LJ. Determining the resolution of a real-time arousal gauge. In: Gardner PJ, Fountain III AW, editors. *Chemical and Biological Sensing VII. Proceedings of SPIE 6218*; 2006. p. 286–97.
23. Wilson GF. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *Int J Aviat Psychol* 2002;**12**(1):3–18.
24. Wilson GF, Caldwell JA, Russell CA. Performance and psychophysiological measures of fatigue effects on aviation related tasks of varying difficulty. *Int J Aviat Psychol* 2007;**17**(2):219–47.

Wei Zongmin is a Ph.D. candidate at the School of Aeronautic Science and Engineering, Beihang University, Beijing, China. His main research interests are aviation accident analysis, human factors airworthiness evaluation, and human–machine interface design on flight decks.

Zhuang Damin is a professor and Ph.D. supervisor at the School of Aeronautic Science and Engineering, Beihang University, Beijing, China. He received Ph.D. degree from the Kagoshima University, Kagoshima, Japan. His current research interests are human–machine engineering, computer simulation of human–machine interface and computational fluid dynamics.