TRANSPORT PROBLEMS 2019 Volume 14 Issue 2
PROBLEMY TRANSPORTU DOI: 10.20858/tp.2019.14.2.5

Keywords: human factors; fatigue; psychological testing; performance; pilot; flight simulator

Vladimir SOCHA, Lenka HANAKOVA, Michal FREIGANG,
Jakub KRAUS, Slobodan STOJIC
Czech Technical University in Prague, Faculty of Transportation Sciences
Horska 3, 128 03 Prague, Czech Republic
Lubos SOCHA*, Peter HANAK
Technical University of Košice, Faculty of Aeronautics
Rampová 7, 041 21 Košice, Slovakia
**Corresponding author*. E-mail: lubos.socha@tuke.sk

IMPACT OF PILOTS' TIREDNESS ON THE OUTCOME OF PSYCHOLOGICAL TESTING

Summary. Presented work is primarily oriented on the experimental verification of the influence of fatigue on the psychological condition of the flying personnel, using psychological and performance tests. For the evaluation of a pilot performance, the 24 hours experiment was conducted. In total, eight subjects participated in the experimental measurements. Eight participants went through several tests, including simulator flights, to investigate the effects of the fatigue on the results of psychological measurements. Measurements included workload evaluation, using NASA task load evaluation concept and performance testing, using the so-called OR-test. Significant statistical differences between measurements performed during 24 hours were not found in the case of NASA task load Scores. In the case of OR-test, Friedman ANOVA and subsequent post-hoc analysis show that the greatest decrease in performance was observed in approximately 22 hours of wakefulness, i.e. approximately in half of the measuring process. The concept of 24-hour measurements for the quantification of fatigue is not commonly used yet as well as objectivization using performance testing. As the apparent effect of fatigue is mainly on performance testing results, it can be argued that this work could serve as a basis for further studies on fatigue. Also, it could serve as a support for introducing new pilots' psychological testing procedures in the future, which could contribute to current efforts to improve aviation safety.

1. INTRODUCTION

The role of fatigue in the human organism has extraordinary significance. It can be considered as an important natural indicator of the overloading of the body that is responsible for its protection [1]. As the beginning of fatigue is slow and unnoticeable, the risk is that a pilot will not realize a decrease in his performance and abilities [2]. Fatigue is a highly subjective matter, and it manifests in different ways. An actual level of fatigue could be best evaluated by the person who feels it. It is, therefore, difficult to measure or detect it. However, its characteristics are the same for one particular person, who knows what the reasonable forms of rest and regeneration time are [1,3]. In the case of objectivization, the fatigue could be detected through behavior and facial expression or, for example, using voice analysis [4]. Fuzzy face expression along with irritability, malaise, passivity and slow reactions can be considered as typical signs of fatigue, indicating a decrease in overall human performance [3].

Fatigue, as one of the possible changes in the current physical state of the body, has a direct impact on the mental and physical performance of a pilot. One of the basic requirements for pilots is physical and mental ability to perform their duties in the cockpit [1]. Typical fatigue symptoms in the cockpit include inattention, laxity, impaired situational awareness, and mistakenness [2,4-8], and a pilot is easily disturbed, he responds irritably, his communication with other crew members is rather restrained, and his mood varies. The problems with piloting precision, fine locomotor skills, short-term memory, and seeing could be observed [2,8]. Furthermore, fatigue reduces pilot resistance to hypoxia, overload, or kinetosis [1]. In extreme cases, a workload could be so high that a pilot does not have enough capacity to deal with a complex situation. Exhaustion of the functional reserve of an organism then brings a risk of failure [9].

Pilots and crew members are constantly confronted with long working days, early departures, late arrivals, and non-standard working hours, which include night shifts and specific timetables [10,11]. Commercial pilots on long journeys, as well as military pilots, frequently pass through many time zones, which contribute to circadian disorders and sleeping problems [9].

Safety is one of the most important factors in aviation. A negative impact on the safe conduction of the flight may be related to the family relationships problems, working shifts, financial problems or problems derived from dissatisfaction from career growth. Stress and anxiety have a demonstrable impact on air incidents [12-15]. Sleeping disorders could also have an impact on pilot performance, and these are proved, in context of increased fatigue, to be the cause of some air accidents [4].

A monitoring of fatigue in aviation, especially, among commercial pilots, is based mainly on subjective quantifiers, i.e. a fatigue is monitored mainly through questionnaire survey [16-18]. The results of such surveys point to the fact that pilots report a decrease in performance mainly due to a fatigue (for example, according to the results [5] an impact of fatigue on the performance is reported by more than 84% of the pilots questioned). Standard methods for tracking the performance of the pilots include the NASA Task Load Index (NTLx). NTLx is a standardized questionnaire, i.e. a subjective pilot evaluation after performed tasks [19].

At the same time, not only in aviation, there is a general effort leading to an objectification of the data. In recent years, it has been realized that questionnaire surveys are not the most appropriate way for the quantification of fatigue and other techniques appeared. One of these techniques is performance testing. These tests are part of the standard testing performed at the Institute of Aviation Medicine.

Based on what is mentioned above, the aim of this study is an evaluation of the influence of pilot's fatigue on the outcome of psychological and performance tests. Such testing could further contribute in enhancing air transport safety, in the context of the introduction of the Fatigue Risk Management System in aviation.

2. MATERIALS AND METHODS

2.1. Participants and measuring procedure

The experiment was prepared, and the person who was the subject of the measurement was awake at least 36 hours after the measurement. This was achieved in a way that subjects got up on the measurement day at 6:00 or earlier. During the following 12 hours (until the beginning of measurements), the subjects were not allowed to carry out strenuous physical activities, use substances that cause fatigue, take alcohol, etc. On this basis and according to the predefined rule, a goal was to standardize the daily regime of the attended participants. Altogether, eight subjects took part in experiment, from the age range of 21–24 years and all being a professional pilots, with experience in Instrument Flight Rules (IFR) flying in Instrumental Meteorological Conditions (IMR) of around 100–150 flight hours. The theoretical and practical base of the pilots was comparable.

The measurements, in the sense of a 24-hour experiment, began from 18:00 Central European Time in the summer months between 1. 6. 2018 and 31. 8. 2018. The subjects were instructed in detail how to use the navigation devices effectively, and they were also reminded about the controls and settings

of the onboard devices that were available. It was necessary to use an iPad device in which the JeppFD (Jeppesen, Englewood, CO, USA) application was installed. It is important to note that these air maps are currently used by many airlines, and it is not easy for a layman to understand such maps. All subjects knew these maps, as they were commonly used for flights during practical training, whether on a simulator or on an airplane. After a detailed explanation, the subjects got answers for all their questions related to the test.

Flights took place on the Beechcraft simulator, a twin-engine propeller aircraft. The cockpit of the simulator was equipped with standard flight, navigation and communication devices along with an engine and structure control equipment. Instruments using mechanical indicators, integrated indicators and electronic display EFIS were available. During the flights, the emphasis was placed on the use of instruments using mechanical indicators – mechanical speedometer, artificial horizons, barometric mechanical altimeter, CDI pointer, turn indicator in combination with the relative tilting indicator (turn coordinator), directional flywheel, mechanical variometer, VOR / LOC / GPS indicator, ADF indicator (MDI) and magnetic compass.

There were 8 measurements on the simulator for all participants (see Fig. 1), each of which lasted approximately one and a half hours. The briefing before the flight took about 20 minutes. All the measurements took place in Germany, from several airports through different radionavigation facilities to several other airports. Subjects never took off and landed to one airport, and it was always a navigational flight. All flights were under IFR for IMC. Even though the flights were not exactly the same, all were of the same nature, with the main concern to ensure the uniformity of the course of measurement. Therefore, each flight was conducted over three radionavigation points under defined conditions, which ensured the same length and character of the individual flights.

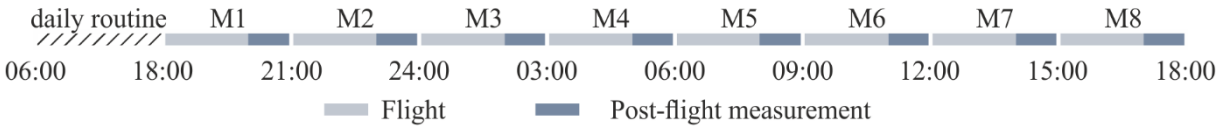

Fig. 1. Overall schedule and concept of experimental measurements

The subjects had a simulation of a real IFR flight, including engine start, requests and clearances, listening of information from the automatic terminal information service, etc. All communication with the air traffic controller was conducted in English. The surrounding traffic, including communication with them was not simulated.

After the flight, a post-flight measurement took place where the subjects were questioned through questionnaire survey, where developed software was used for the calculation of the NTLx. The subjects also underwent the performance test, the OR-test. This testing is described in the chapters below. In the presented study, the emphasis was placed on these two tests, which were carried out on the basis of the individual flights that were realized. The overall schedule and concept of experimental measurements is shown in Fig. 1.

2.2. Psychological testing

The two types of testing were used in the experiment. The first of the chosen methods was the performance test, called OR-test, based on a standard test used at the Institute of Aviation Medicine. The test was performed in approximately three-hour intervals and was repeated eight times for each subject.

The test was performed on a PC, and it was actually a program designed in C# language. The user started the program, generating a table of 25 rows and 15 columns (a total of 375 cells).

Each cell of the table contained one code consisting of 4 characters, see Fig. 2. The first two characters were letters and the other two were numbers. The combination of letters was always retained throughout the column, the column started with a random number, and increased with each row with a value of 1. If the value of 99 was exceeded within the column, the value 00 was then

followed by the value of 01. The table in terms of columns consisted of three parts. In the central part, i.e. the middle column, the letter part of the code always composed of the letters MM. The numeric code in this column then started with 01 and went further to the value of 25. The first part of the table contained codes starting with letters A–M. In individual columns, the codes were in an alphabetical order, i.e., the first column will always be the first combination in alphabetical order. There cannot be a situation where the first column begins with the letter C and the second with the letter A. Similarly, there cannot be a situation in which one column is the combination of AC and the following the combination of AA. It is clear, therefore, that in the 7th column there may be the combination of ML. The third part of the table (columns 9–15) worked on a similar principle, but with the second part of the alphabet, i.e. the letters N–Z. For this part of the table, the 9th column could be the combination of MN, and the last one may be the highest alphabetical combination, i.e., ZZ. For the purpose of the program, letters of the alphabet containing the diacritics, the letter CH and the letter W, were omitted. The basic principle of the test is shown in Fig. 2.

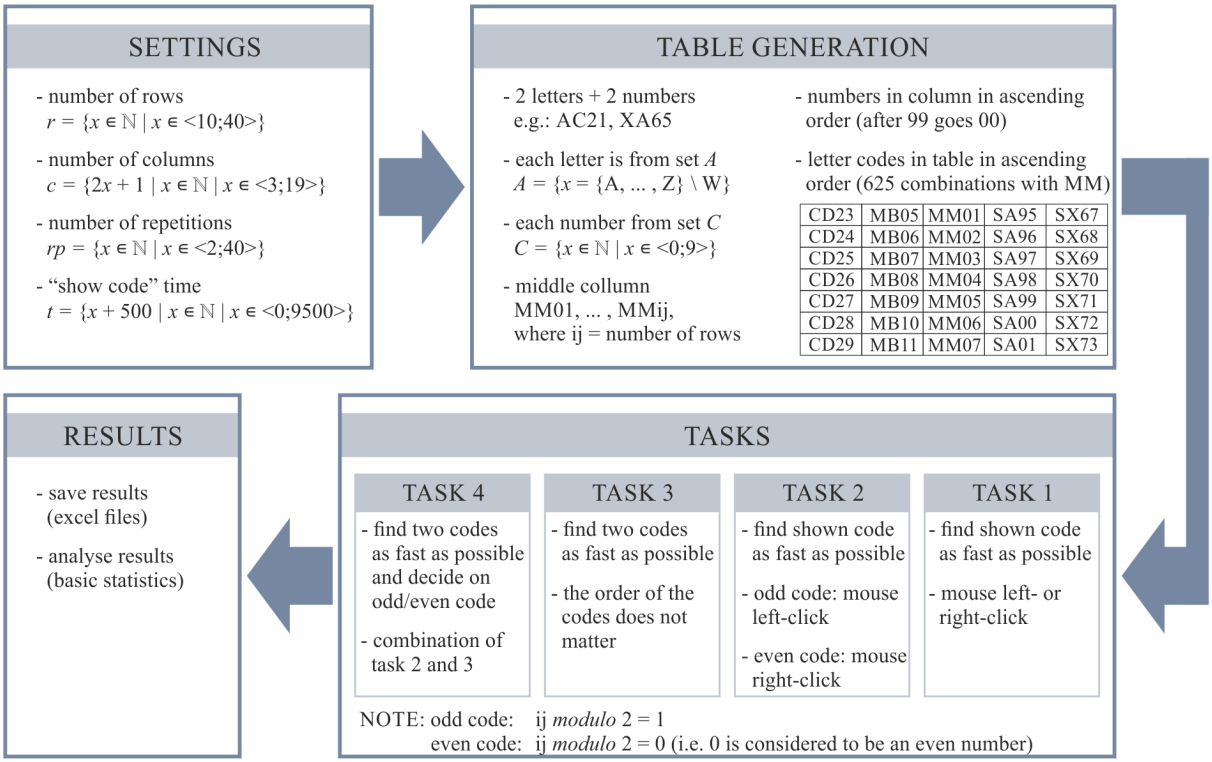

Fig. 2. Block diagram of OR-test

The experiment was divided into four parts, of which each task was repeated fifteen times. After starting the program, task instructions in English were displayed in front of each section, detailing exactly what the task is about.

The first part of the experiment included a field with two-letter and two-digit code displayed on the monitor for three seconds, see Fig. 2. Subsequently, it was the task of the subjects to locate and check this field as quickly as possible. The correctness and time from the disappearance of the text input field after clicking on any field by the user were evaluated.

The second part was the same as part one; however, the user also had to evaluate whether the number behind the letters was even or odd. If it was even, it was clicked with the right mouse button, and if it was odd, it was clicked with the left mouse button. If the 00 combination was displayed, it was considered to be an even number.

The third part had an increased demand on users compared to the previous parts. On the screen, two pairs of letters and two pairs of digits appeared for 3 seconds. The objective of the subject was to

combine the first pair of letters with the first pair of digits and the second pair of letters with the second pair of digits (e.g. SX MB 67 09). According to Fig. 2, the task was to find the SX67 and MB09 fields and check them in the table as soon as possible.

The last part of the test was based on the previous part but was further extended with classifying the numbers on even and odd. In the case shown in Fig. 2, correct number was clicked with the left mouse button on both fields, as both numbers are odd. Again, the number 00 was required to be numbered as even.

At the end of the fourth part, the results were shown to the subjects (the average response time and correctness). The data were saved in *.xlsx format for subsequent analysis. The exported data carried information about the required codes and the real codes, response time information and a list of correct/incorrect answers.

2.3. NASA Task Load Index

For purposes of subjective performance evaluation, the so-called NASA Task Load Index (NTLx) was used. Therefore, a software tool for collecting and evaluating data was created using the standard test, see also [19,20]. Subjects passed through this questionnaire after each simulated flight, where their mental demand, physical demand, temporal demand, performance, effort and frustration were evaluated on a scale from 1 to 20.

In the case of mental demand, the subjects were assessed on the basis of the individual load from the point of view of psychological actions, e.g. how much he had to think, count, navigate in space, how much he was influenced in flight by clouds, the IMC conditions, and also searching for instruments, or unclear arrangement of devices (orientation in the instrument order), communication with the controller (another accent), whether it was simple or exhausting for him.

In terms of physical demand, the subjects were determined on the basis of how much the task was physically demanding, e.g., airplane control, whether the flight was relaxing or there was a constant need for flight controls movement.

In the case of temporal demand, subjects were evaluated on the basis of the exposure to time pressure, e.g. whether they were instructed progressively and logically by the ATCO, whether there was need to hurry during the take-off, did they have time to check all on-board instruments, to perform a motor test, check the amount of fuel or pressure, or whether there was not enough time to finish all tasks in the given time frame.

Performance scores were self-ratings from an individual's perspective, i.e., how successful an individual thinks he was and how much he was generally satisfied with his performance during the flight, course of communication, navigation, take-off, and landing.

In the case of the evaluation of the effort, the subjects were evaluated on the scale from 1 to 20, i.e., how demanding it was to satisfy the required level of performance e.g. whether holding the given altitude for them was simple or complicated, or whether they were able to keep the altitude in the tolerance after being warned by the controller.

Regarding frustration, subjects were assessed on how insecure, demotivated, stressed, angry or outraged by the flight they were, no matter whether due to a complexity of the flight tasks performed or because of their own mistakes, fatigue or long stereotypical flight.

After completing this questionnaire (after evaluation), the second part followed. It represented a comparison of these six categories to each other. Each category was compared with all other categories; therefore, it was possible to prove what affected the subjects the most. The subjects gradually marked the item that affected them more from their impression. Based on this approach, it is possible to specify weight for individual items.

The result of the test, then, indicates the magnitude of the significance of each monitored category to the total Task Load Index (TLx). The evaluation procedure consists of the conversion of the first part of the questionnaire into a percentage form. The scale from one to twenty represents 5%–100% with a 5% step (see Fig. 3-A). Through the second questionnaire, it is possible to assign the weight to individual factors based on the particular summation of their designations or preferring it to another category. In total, from six to two combinations, the number of occurrences is divided by 15, which

gives weight for each category (see Fig. 3-B). If this weight is multiplied by percentages from the first questionnaire, the score is obtained. The level of significance of each of the monitored categories implies from the obtained score (see Fig. 3-C), therefore, we can talk about overall scores. If all these categories are added, a total TLx [20] can be obtained. However, this value is inherently meaningful and does not carry any information on the overall score. Therefore, overall scores will be used for further evaluation within the submitted paper.

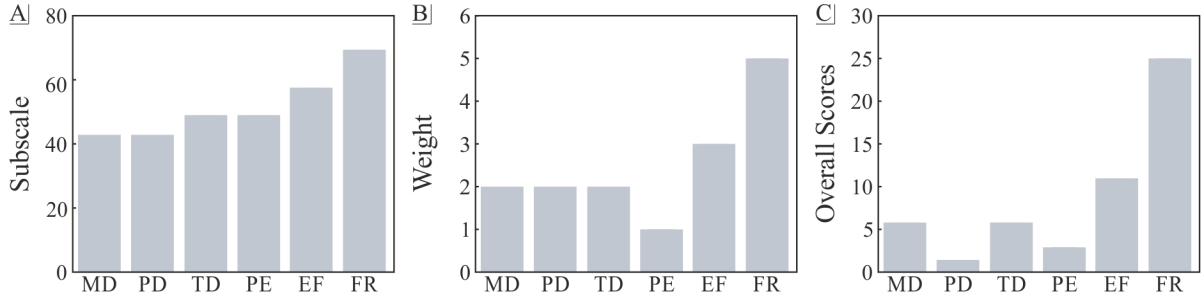

Fig. 3. Example of partial results for the Nasa Task Load Index

2.4. Data processing and statistical evaluation

The results of psychological and performance testing were further processed statistically. In view of the results of the Jarque–Bera test, in which the zero hypothesis was rejected, i.e., the data did not come from the normal distribution, nonparametric statistical testing was chosen. Based on measured data, i.e., data from eight subjects collected in eight measurements, the Friedman test (also Friedman's non-parametric one-way ANOVA with repeated measures) as a test for more than two dependent selections was chosen to test the differences between individual measurements [21]. For the post-hoc analysis, the Dunn–Sidak test [22] was used. Statistical testing was carried out using own-design software in Matlab environment (MATLAB R2017a, MathWorks, Inc., Natick, MA, USA).

3. RESULTS

3.1. NTLx results

In the case of the mental demand parameter, statistically significant differences between the distributions in the measured measurements were not found using the Friedmann test, X^2=11.72 a p=0.11. Therefore, post-hoc analysis did not show any statistically significant differences between the individual pairs of measurements at the significance level of α=5 %.

Distribution of individual measurements for mental demand is shown in Fig. 4. There was an increase in the subjective sensation of psychological stress from 1 to 3, followed by a slight decrease and a rapid increase, gradual increase to the maximum at the measurement 6. During the 7th measurement, a slight decrease was evident, followed by a renewed increase. However, the differences were not so great and the Friedman test did not show any statistically significant differences

In the case of the physical demand parameter, statistically significant differences were not found, X^2=3.61, and the resulting p-value was 0.82. Therefore, post-hoc analysis did not show any statistically significant differences between the individual pairs of measurements at the significance level of α=5 %.

No statistically significant differences were found for the temporal demand parameter, X^2=8.61 and p=0.28. Post-hoc analysis in this case also did not show any statistically significant differences between the individual pairs of measurements at the significance level of α=5 %. Despite the absence of significant differences between observed measurements, it is possible from the data presented in Fig. 4 to see a characteristic trend. The subjective score of the temporal demand given by

the candidates from the second measurement (at 24:00) continually increased to the measurement number 4 (at 06:00), and it decreased during fifth measurement and subsequently increased again continually.

No statistically significant differences were found for the performance parameter, X^2=4.86, p=0.67. The progress of this parameter is similar to the temporal demand score.

In case of effort, statistically significant differences were not found as well, as X^2=4.86 and p=0.67. Post-hoc analysis in this case also did not show any statistically significant differences between the individual pairs of measurements at the significance level of α=5%. The distribution of the results of the test related to the effort, shown in Fig. 4, shows a slight increase in the subjective feeling of effort from the 3rd (at 03:00) to the 5th measurement (at 09:00), followed by a slight decrease of the median in 6th measurement (at 12:00).

In the case of the frustration parameter, statistically significant differences were not found, X^2=2.87 and p=0.89. From the distribution of this parameter in Fig. 4, it is possible to see the biggest decrease in the level of frustration during the 7th measurement.

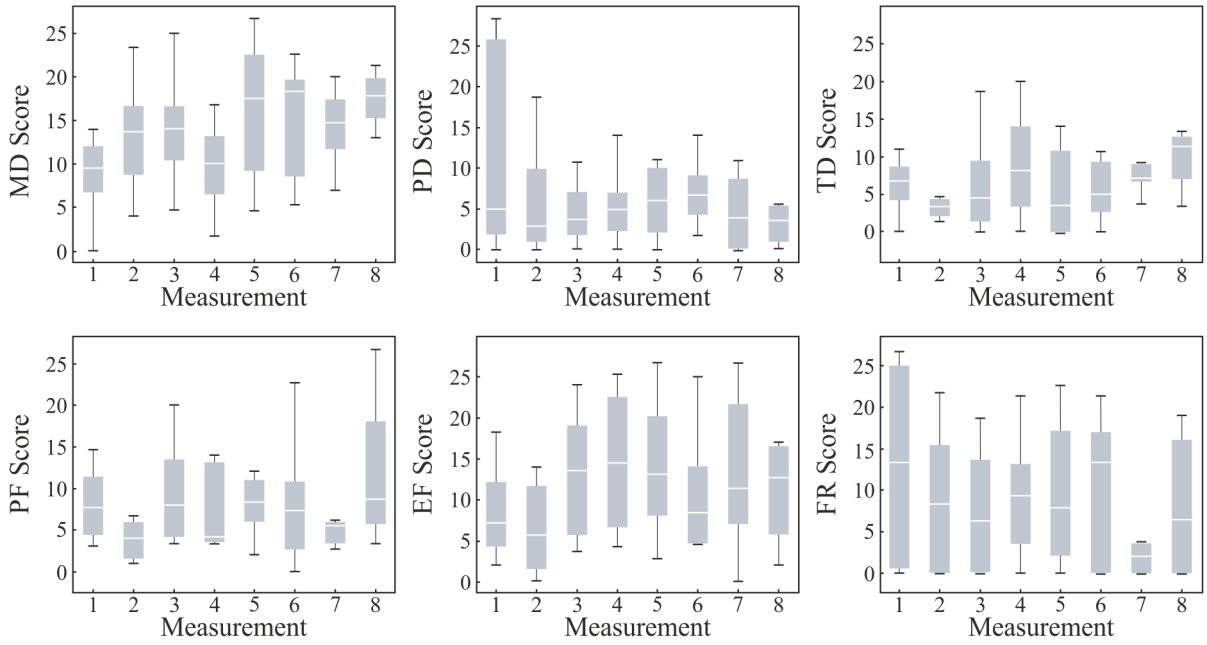

Fig. 4. Presentation of the results of overall scores for individual task load categories (MD - mental demand, PD - physical demand, TD - temporal demand, PF - performance, EF - effort, FR - frustration)

3.2. OR-Test results

For the time required to check the code in task 1, using the Friedmann test, statistically significant differences between distributions in the observed measurements were found, X^2=33.66 and p=1.99×10^{-5}. Post-hoc analysis, therefore, showed statistically significant differences between measurements 3 and 6 (p=0.03, confidence interval (6.24, 219.15)), measurements 3 and 7 (p=0.02, confidence interval (9.49, 231.67)), measurement 4 and 5 (p<0.01, confidence interval (18.36, 224.05)), measurements 4 and 6 (p<0.01, confidence interval (33.96, 246.88)), and measurements 4 and 7 (p<0.01, confidence interval (37.22, 259.40)) at the significance level of α=5 %. For graphical presentation of mentioned results, see Fig. 5-A.

For the time required to check the code in task 2, using the Friedmann test, statistically significant differences between distributions in the observed measurements were found, X^2=30.07 and p=9.20×10^{-5}. The post-hoc analysis, therefore, showed statistically significant differences between measurements 1 and 6 (p=0.03, confidence interval (3.68, 216.95)), measurements 3 and 6 (p=0.02, confidence interval (4.12, 216.55)), measurement 5 and 6 (p<0.01, confidence interval (34.77,

247.20)), measurements 5 and 7 (p=0.02, confidence interval (7.61, 229.28)), and measurements 5 and 8 (p<0.01, confidence interval (2.02, 223.69)) at the significance level of α=5 %. For graphical presentation of mentioned results, see Fig. 5-B.

In the case of the total time required for checking the code in task 3, Friedmann's test found statistically significant differences between the distributions of the measured values, X^2=12.95 and p=7.33×10^{-2}. The post-hoc analysis showed a statistically significant difference between measurements 4 and 7 (p=0.04, confidence interval (2.37, 224.55)) at the significance level of α=5 %. For graphical presentation of mentioned results, see Fig. 5-C.

In the case of the total time required to check the code in task 4, using the Friedmann test, statistically significant differences were found between the distributions of the measured values, X^2=67.28 and p=5.22×10^{-12}. Post-hoc analysis, therefore, showed statistically significant differences between measurements 1 and 5 (p=0.02, confidence interval (9.05, 214.74)), measurements 1 and 6 (p<0.01, confidence interval 83.73, 296.64)), measurements 1 and 7 (p<0.01, confidence interval (44.04, 266.21)), measurements 1 and 8 (p<0.01, confidence interval (95.13, 317.31)), measurements 2 and 6 (p<0.01, confidence interval (43.96, 256.87)), measurements 2 and 7 (p=0.03, confidence interval (4.27, 226.44)), measurements 2 and 8 (p<0.01, confidence interval (55.35, 277.53)), measurement 3 and 8 (p=0.02, confidence interval (11.15, 233.32)), measurements 4 and 6 (p<0.01, confidence interval (53.83, 266.74)), measurements 4 and 7 (p=0.01, confidence interval (14.14, 236.31)), and measurements 4 and 8 (p<0.01, confidence interval (65.23, 287.40)) at the significance level of α=5 %. Graphical presentation of results is shown in Fig. 5-D.

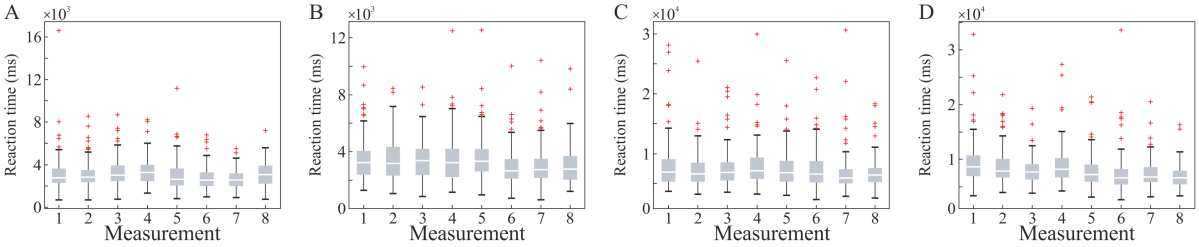

Fig. 5. Presentation of the OR-test results of total time required to check the codes in respective task (A – task 1; B – task 2; C – task 3; D – task 4)

For the time required to check the first code in task 3, using the Friedmann test, statistically significant differences were found between distributions in the measured values, X^2=148.48 and p=8.43×10^{-29}. Post-hoc analysis, therefore, showed statistically significant differences between measurements 1 and 3 (p<0.01, confidence interval (-274.77, -50.73)), measurements 1 and 4 (p<0.01, confidence interval (-352.43, -128.39)), measurements 1 and 5 (p<0.01, confidence interval (-406.80, -182.76)), measurements 1 and 6 (p<0.01, confidence interval (-251.28, -27.24)), measurements 1 and 7 (p<0.01, confidence interval (-250.16, -26.12)), measurements 2 and 3 (p<0.01, confidence interval (-325.90, -101.86)), measurement 2 and 4 (p<0.01, confidence interval (-403.57, -179.53)), measurements 2 and 5 (p<0.01, confidence interval (-457.93, -233.89)), measurements 2 and 6 (p<0.01, confidence interval (-302.41, -78.37)), measurements 2 and 7 (p<0.01, confidence interval (-301.30, -77.26)), measurements 3 and 5 (p<0.01, confidence interval (-244.04, -20.01)), measurements 4 and 8 (p<0.01, confidence interval (76.88, 300.92)), measurements 5 and 6 (p<0.01, confidence interval (43.51, 267.54)), measurements 5 and 7 (p<0.01, confidence interval (44.61, 268.65)), and measurements 5 and 8 (p<0.01, confidence interval (131.24, 355.28)) at the significance level of α=5 %. Results are graphically presented in Fig. 6-A.

For the time required to check the first code in task 4, using the Friedmann test, statistically significant differences were found between distributions in the monitored measurements, X^2=86.28 and p=7.14×10^{-16}. Post-hoc analysis, therefore, showed statistically significant differences between measurements 1 and 4 (p<0.01, confidence interval (-262.37, -39.26)), measurements 1 and 5 (p<0.01, confidence interval (-264.03, -40.92)), measurements 2 and 3 (p<0.01, confidence interval (-317.28, -94.16)), measurements 2 and 4 (p<0.01, confidence interval (-364.58, -141.46)), measurement 2 and 5

(p<0.01, confidence interval (-366.24, -143.12)), measurements 2 and 6 (p<0.01, confidence interval (-309.02, -85.91)), measurements 2 and 8 (p=0.04, confidence interval (-225.36, -2.25)), measurements 4 and 7 (p<0.01, confidence interval (40.01, 263.12)), measurements 4 and 8 (p<0.01, confidence interval (27.66, 250.77)), measurements 5 and 7 (p<0.01, confidence interval (41.67, 264.78)), and measurements 5 and 8 (p<0.01, confidence interval (29.31, 252.43)) at the significance level of α=5 %. Results are graphically presented in Fig. 6-B.

In case of mistakenness in task 1, Friedmann's test did not show statistically significant differences between distributions in the monitored measurements, X^2=13.62 and p=0.06. Therefore, post-hoc analysis did not show any statistically significant differences between the individual pairs of measurements at the significance level of α=5 %.

In the case of mistakenness in task 2, Friedmann's test did not show statistically significant differences between distributions in the monitored measurements, X^2=10.77 and p=0.15. Therefore, post-hoc analysis did not show any statistically significant differences between the individual pairs of measurements at the significance level of α=5 %.

In both cases, errors were minimal, i.e., as the extremes of respective distributions. The graphical presentation of these results would, therefore, show a common distribution median, minimum, maximum, and quartiles, along with extraordinary values. Therefore, these results are not graphically presented.

In the case of mistakenness in task 3, Friedmann's test did not show statistically significant differences between the distributions in the measured measurements, X^2=5.02 and p=0.65. Therefore, post-hoc analysis did not show any statistically significant differences between the individual pairs of measurements at the significance level of α=5 %. Results are graphically presented in Fig. 6-C.

In the case of mistakenness in task 4, Friedmann's test did not show statistically significant differences between distributions in the monitored measurements, X^2=7.81 a p=0.35. Therefore, post-hoc analysis did not show any statistically significant differences between the individual pairs of measurements at the significance level of α=5 %. Results are graphically presented in Fig. 6-D.

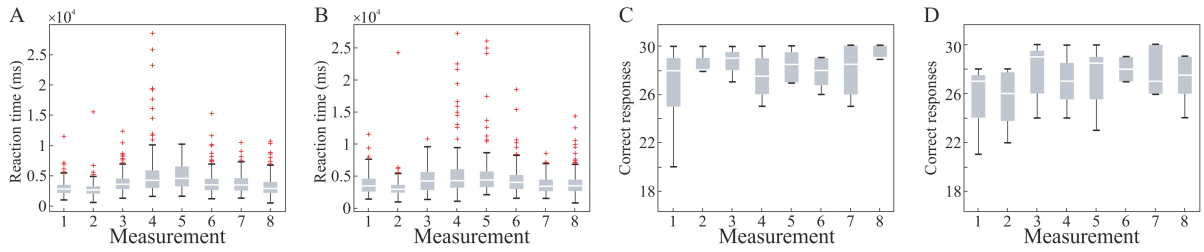

Fig. 6. Presentation of the OR-test results of time required to check first code in task 3 (A) and task 4 (B) along with distribution of correct answers in task 3 (C) and task 4 (D)

4. DISCUSSION

Results of psychological testing through NTLx showed no statistically significant differences. The subjects were most affected by the frustration, the least by time pressure. In the case of performance tests, statistically significant differences were observed from the time perspective, but not from the correctness of the answers. It is, therefore, likely that the subjects have the priority to tick the correct field over the priority of filling the box as soon as possible. With increasing experiment time, pilot performance in the test did not change much, and the scores of individual subjects were almost unchanged throughout the measurement. Therefore, the effect of fatigue on the correctness of responses in these tests was not observed.

Success in OR-test was generally high, the best subject achieved an average score of 88.17 out of a total score of 90 possible points, and the least successful subject got 76.80 out of 90 possible. The average score was 83.98 points out of a total of 90.

The average response time for task 1 across all measurements was 3040 ms, with the highest average time of 3382 ms being found in the 4th measurement, and the shortest average time was 2673 ms, detected at the 7th measurement. These results can be attributed to the fact that subjects were awake after the 4th measurement for about 20 hours in the early morning hours, and thus, considerable fatigue with impaired ability to perceive and process information was observed. In the case of the 7th measurement, daylight was already present, and the subjects were well informed about the test. This can also be supported by a statistical test, the results of which show a statistically significant difference between the 4th and the 5th measurements, with the measurement 5 taking place at the time of transition from night to day.

The average response time for task 2 across all measurements was 3330 ms, with the highest average time of 3583 ms being found in the 5th measurement, and the shortest average time was 2877 ms found in the 6th measurement. This is supported by statistical testing, the results of which show a statistically significant difference between these two measurements.

The average cumulative time for checking the two fields in task 3 was 7442 ms, with the highest average time of 7922 ms being found in the 4th measurement and the shortest average cumulated time was 6915 ms found in the 8th measurement. Statistically significant differences were found between measurements 4 and 7. It is, therefore, obvious that in the case of the 4th measurement, subjects were experiencing the greatest fatigue manifestations.

The average cumulative time for checking the two fields in task 4 was 8204 ms, with the highest average time of 9324 ms being found for the 1st measurement, and the shortest cumulative average time was 7088 ms for the 8th measurement. In our opinion, in this task, the ability to learn this test had manifested, and therefore, the best performance in the case of the last measurement was the worst in the case of measurement 1. Based on statistical testing a statistically significant difference was found between these measurements.

The average time for checking one field in task 3 was 3738 ms, and in task 4, it was 4106 ms. From an overall view of the results, it is clear that the subjects experienced the greatest fatigue in the morning, with no daylight yet.

Despite the effort to suppress the learning process that was developed during the OR-test preparation, a certain degree of learning was observed and a system for a successful solution was obviously found. The correctness of the marked fields was probably greatly influenced by the reduced ability to perceive and understand the information. NTLx results do not show differences between measurements. However, the results show that the least influential factor was the time load and the biggest one was frustration, which is one of the contributing factors of air accidents [23, 24].

Looking at the results of performance tests, it is clear that with increasing fatigue, performance is decreasing. At the same time, it is obvious that fatigue is not only related to the total wakefulness time, but pilots, like other people, are also affected by the daily cycle, especially with regard to the presence of daylight and also other factors influencing the circadian rhythm, body temperature, heart rhythm, blood pressure, sensory or adrenal activity [2,25,26]. This may be the basis for degraded performance scores, in the early morning hours.

Overall, the results indicate that measurement using objective methods, i.e. performance tests, seems to be appropriate for pilots' fatigue monitoring and can provide a new insight into this issue, thus enhancing the knowledge of new information not only through subjective surveys.

5. CONCLUSION

The aim of this article was to evaluate the impact of fatigue on the outcome of psychological and performance tests. For the purpose of the study, a 24-hour measurement experiment with a predefined pre-flight mode was performed in order to cause fatigue. One of the simplest possible reasons for inducing fatigue was the disruption of the normal daily rhythm with the absence of night sleep. Due to the cockpit environment, the pilot occupation and the measurement needs, this choice seemed to be convenient. The study was attended by 8 students of the Faculty of Transportation Sciences, department of Air Transport, CTU in Prague with comparable theoretical and practical knowledge.

During the course of the measurements, the subjects underwent 8 simulated IFR flights. After each flight, they have undergone psychological testing through a standardized NASA Task Load Index (NTLx) and performance testing through the OR-test, which is similar to a test used at the Institute of Aviation Medicine. During the preparation of this test, the goal was to reduce the learning process as much as possible. For the purpose of this study, both of these tests were implemented in their own-developed software. The results of both tests were then statistically evaluated.

From the results of testing, it is obvious that the greatest fatigue subjects achieved in the early morning hours. In the case of these measurements, which were at the transition from night to day light, there was an increase in the time required to perform the performance test tasks, and, at the same time, an increase of error rate of the subjects was observed, i.e., a decrease in performance. In the case of the standardized NTLx questionnaire, the results show that the largest number of subjects was burdened with frustrations, whereas the smallest number of subjects was burdened with time constraints, which can be attributed to the fact that no time limits were set for the subjects to meet individual flight tasks. No statistically significant differences between the measurements were found in NTLx results. Thus, it is clear that objective testing that was performed in this case through performance testing is able to provide more information than standard questionnaires that are represented by NTLx. At the same time, it is obvious that not only the time of vigilance but also the time of day and the presence of daylight play a role in the case of fatigue, and it is also necessary to take this factor into consideration when planning flights.

One of the limitations of this study is the lack of subjects involved in the measurement. This was, however, caused by time-consuming experimental setup that was a limiting factor for subjects' participation. Moreover, it was also necessary to provide additional testing personnel for the duration of the measurement, to act as air traffic controllers, to communicate with the crew of the airplane, and to monitor the uniformity of the subjects' regime and the measurement in the context of activities performed outside the simulated flights. Therefore, the high personnel and time requirements for the measurement are obvious, which are the basis for above-mentioned limitation. Another limitation could be the measurement length with respect to the age of the subjects. As the experiment was attended by relatively young persons (aged 21–24 years), it can be assumed that such one-time fatigue did not have to be sufficiently limiting for them.

On the basis of what is mentioned above, it can be stated that it would be further advisable to extend the sample of the measured subjects. In addition, it would be advisable to perform more measurements to induce longer fatigue that could be much more observable during the test. It should be also taken into consideration possibilities for the extension of the tests performed during the experiment, which could contribute to a further understanding of the pilot fatigue.

Although this is a pilot study on a relatively small sample of subjects, the apparent effect of fatigue is mainly on performance testing results. It can, therefore, be argued that this work could serve as a basis for further studies to provide a more detailed description of fatigue and could serve as a support for introducing new pilots' psychological testing procedures in the future, which could contribute to current efforts to improve aviation safety.

Acknowledgements

This work was also supported by the Slovak Research and Development Agency under the grant No. APVV-17-0167 "Application of the Self-regulatory techniques for the Flight Crew Preparation".

References

1. Abd-Elfattah, H.M. & Abdelazeim, F.H. & Elshennawy, S. Physical and cognitive consequences of fatigue: A review. *Journal of Advanced Research*. 2015. Vol. 6. No. 3. P. 351-358.
2. *Human performance and limitation: JAA ATPL training*. *Second Edition.* Neu-Isenburg: Jeppesen. 2007.

3. Sundelin, T. & Lekander, M. & Kecklund, G. & Van Someren, J. W. & Olsson, A. & Axelsson, J. Cues of Fatigue: Effects of Sleep Deprivation on Facial Appearance. *Sleep*. 2013. Vol. 36. No. 9. P. 1355-1360.
4. Creeeley, H. & Nesthus, T. Predicting fatigue using voice analysis. *Aviation, Space, and Environmental Medicine*. 2007. Vol. 78. No. 7. P. 730-743.
5. Kozuba, J. & Pil'a, J. Aircraft automation systems versus pilot situational awareness (SA) - Selected aspects. In: *Proceedings of 19th International Conference "Transport Means"*. Kaunas: Kaunas University of Technology. 2015. P. 688-693.
6. Kozuba, J. & Piľa, J. Chosen aspects of pilots situational awareness. *Nase More*. 2015. Vol. 62. No. SI. P. 175-180.
7. Madej, K. & Kozuba, J. Technology as a capability enhancement in the air training. In: Proceedings of 6th International Conference on "Military Technologies". Brno: University of Defence. 2017. P. 467-476.
8. Morris, M.B. & Wiedbusch, M.D. & Gunzelmann, G. Fatigue Incident Antecedents, Consequences, and Aviation Operational Risk Management Resources. *Aerospace Medicine and Human Performance*. 2018. Vol. 89. No. 8. P. 708-716.
9. Bennett, S.A. Pilot workload and fatigue on short-haul routes: an evaluation supported by instantaneous self-assessment and ethnography. *Journal of Risk Research*. 2016. Vol. 21. No. 5. P. 645-677.
10. Weiland, M. & Nesthus, T. & Compatore, C. & Popkin, S. & Mangie, J. & Thomas, L. C. & Flynn-Evans, E. Aviation fatigue: Issues in developing fatigue risk management systems. In: *Proceedings of the Human Factors and Ergonomics Society.* San Diego. 2013. P. 1-5.
11. Avers, K. & Hauck, E.L. & Blackwell, L.V. & Nesthus, T.E. A qualitative and quantitative analysis of fatigue countermeasures training in the aviation industry. *International Journal of Applied Aviation Studies*. 2010. Vol. 10. No. 2. P. 51-66.
12. Caldwell, J.A. Fatigue in aviation. *Travel Medicine and Infectious Disease*. 2005. Vol. 3. No. 2. P. 85-96.
13. Mehta, R.K. & Peres, S.C. & Steege, L.M. & Potvin, J.R. & Wahl, M. & Stanley, L.M. & Nesthus, T.E. Fatigue monitoring and management across different industries. In: *Proceedings of the Human Factors and Ergonomics Society.* Washington. 2016. P. 993-996.
14. Antoško, M. & Sabo, J & Hovanec, M. & Aviation, Sspace, P. & Sekelová, M. How to evaluate the actual psychological readiness of ATCO. In: *Proceedings of 21st International Conference "Transport Means"*. Juodkrante: Kaunas University of Technology. 2017. P. 1062-1065.
15. Antoško, M. & Pil'a, J. & Korba, P. & Lipovský, P. Psychological readiness of air traffic controllers for their job. *Nase More*. 2014. Vol. 61. No. 1-2. P. 5-8.
16. Gregory, K.B. & Winn, W. & Johnson, K. & Rosekind M.R. Pilot fatigue survey: exploring fatigue factors in air medical operations. *Air medical journal*. 2010. Vol. 29. No. 6. P. 309-319.
17. Bourgeois-Bougrine, S. & Carbon, P & Gounelle, C. & Mollard, R. & Coblentz, A. Perceived fatigue for short-and long-haul flights: a survey of 739 airline pilots. *Aviation, space, and environmental medicine*. 2003. Vol. 74. No. 10. P. 1072-1077.
18. Powell, D.M. & Spencer, M.B. & Holland, D. & Broadbent, E. & Petrie, K.J. Pilot fatigue in short-haul operations: effects of number of sectors, duty length, and time of day. *Aviation, Space, and Environmental Medicine.* 2007. Vol. 78. No. 7. P. 698-701.
19. Hart, S.G. Nasa-Task Load Index (NASA-TLX): 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 2006. Vol. 50. No. 9. P. 904-908.
20. Hart, S.G. *NASA Task load Index (TLX) Volume 1.0. Paper and Pencile Package*. 1986. Available at: http://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX.pdf
21. Gibbons, J.D. *Nonparametric statistics: An introduction*. Newbury Park: Sage Publications. 1993. 96 p.
22. Dunn, O.J. Multiple comparisons using rank sums. *Technometrics*. 1964. Vol. 6. No. 3. P. 241-252.

23. Loewenthal, K.M. & Eysenck, M. & Harris, D. & Lubitsh, G. & Gorton, T. & Bicknell, H. Stress, distress and air traffic incidents: job dysfunction and distress in airline pilots in relation to contextually-assessed stress. Stress Medicine. 2000. Vol. 16. No. 3. P. 179-183.
24. Grandjean, E.P. & Wotzka, G. & Schaad, R. & Gilgen, A. Fatigue and stress in air traffic controllers. Ergonomics. 1971. Vol. 14. No. 1. P. 159-165.
25. Graeber, R.C. Aircrew fatigue and circadian rhythmicity. 1988. *Human factors in aviation*. P. 305-344.
26. Hursh, S.R. & Redmond, D.P. & Johnson, M.L. & Thorne, D.R. & Belenky, G. & Balkin, T.J. & Storm, W.F. & Miller, J.C., Eddy, D.R. Fatigue models for applied research in warfighting. *Aviation, space, and environmental medicine*. 2004. Vol. 75. No. 3. P. 44-53.

Received 03.12.2017; accepted in revised form 03.06.2019