AIR-TRAFFIC-CONTROLLERS (ATCO): NEUROPHYSIOLOGICAL ANALYSIS OF TRAINING AND WORKLOAD ATCO

Ing. Pietro Aricò [(1)(2)(5)], Ing. Gianluca Borghini [(1)(2)(4)(5)], Dr. Ilenia Graziani[(1) (5)], Ing. Jean-Paul Imbert[(6)], Ing. Géraud Granger[(6)], Ing. Railane Benhacene[(6)], Dr. Simone Pozzi[(3)], Dr. Linda Napoletano[(3)], Ing. Gianluca Di Flumeri [(7)], Prof. Alfredo Colosimo [(7)] and Prof. Fabio Babiloni[(1)(2)(5)].

[1] Dept. Physiology and Pharmacology, University "Sapienza" of Rome, Italy

[2] IRCCS "Fondazione Santa Lucia", Rome, Italy

[3] Deep Blue srl, Rome, Italy

[4] Department of Computer Science and System, University "Sapienza", Rome, Italy

[5] BrainSigns srl, Rome, Italy

[6] Ecole Nationale de l'Aviation Civile, Toulouse, France

[7] Dept. of Anatomy, Histology, Forensic Medicine and Orthopedics, University "Sapienza" of Rome, Italy

Short title

Training and workload analyses of ATCO

Key words

EEG, ECG, EOG, cognitive learning, training assessment, mental workload, ATM, ATCO.

Financial information

This work is co-financed by EUROCONTROL on behalf of the SESAR Joint Undertaking in the context of SESAR Work Package E - NINA research project. This work is also supported by BrainSigns srl, with a grant from Regione Lazio, FILAS, named "BrainTrained". CUP F87I12002500007.

Corresponding author: Ing. Pietro Aricò, Dept. Physiology and Pharmacology, University "Sapienza" of Rome, Piazzale Aldo Moro, 5 – 00185 Rome (Italy).

email: pietro.arico85@gmail.com

ABSTRACT

OBJECTIVES

The aim of this paper is to present an extensive neurophysiological study of the Air-Traffic-Controllers (ATCOs) during *en route Air Traffic Control (ATC)* simulations. In other words, the purpose was to extract neurophysiological features suitable for the evaluation of the learning progress and for the real-time estimation of the user's workload level.

METHODS

In collaboration with the French Ecole Nationale de l'Aviation Civile (ENAC, Toulouse), it has been developed and tested a specific task for the en-route ATCO. The task's difficulty can be altered according to how many aircrafts the participant have to control, the number and type of clearances required over the time and the trajectory of other interfering aircrafts. The subjects have been asked to learn how to complete the task within a training period of a week and, in the second week, to execute it under different difficulty levels. During the experiments, the Electroencephalogram (EEG), the Electrocardiogram (ECG), the Electrooculogram (EOG), the behavioral data and the perception of the workload have been collected.

RESULTS

The results showed that the frontal theta Power Spectral Density (PSD), the parietal alpha PSD, the heart rate (HR) and the eyeblinks rate (EBR) are reliable features by which evaluating the learning progress and the user's workload.

CONCLUSIONS

It has been demonstrated that it could be possible i) to quantify how well the subjects can accomplish with a new task and ii) to compare subject's performances, in terms of cognitive resources. In addition, it has been presented iii) a system able to significantly differentiate three workload levels, and iv) how the subjective features used for the workload evaluation remain stable over the time.

INTRODUCTION

Controlled airspace is divided into sectors. An *en route* sector is a region of airspace that is typically situated at least 50 km from an airport for which an associated ATCO has responsibility. ATCOs have to accept aircraft into their sector; check aircraft, issue instructions, clearances, and advice to pilots and hand aircraft off to adjacent sectors or to airports. When the aircraft leaves the airspace assigned to the ATCO, control of the aircraft passes onto ATCO controlling the next sector (or to the tower ATCO). As is typical in many real-world complex systems, this environment imposes multiple concurrent demands on the operator, in fact in the *en route* air traffic control environment, the system that confronts the air traffic controller comprises a large number of aircrafts coming from a variety of directions, at diverse speeds and altitudes, heading to different destinations [1]. ATCOs have two main goals. The primary goal is to ensure that aircraft under jurisdiction adhere to International Civil Aviation Organization (ICAO) mandated separation standards. For example, one of the most common separation standards requires that aircraft under radar control be separated by at least 1,000 feet vertically and 5 nautical miles horizontally. The secondary goal is to ensure that aircraft reach their destinations in an orderly and expeditious manner. These goals require the ATCO to perform a variety of tasks, including monitoring air traffic, anticipating loss of separation (i.e., conflicts) between aircraft, and intervening to resolve conflicts and minimize disruption to flow. (For an extensive compilation of the tasks and goals of *en route* control, see [2]). Total world airline scheduled passenger traffic in terms of passenger-kilometers is projected to grow at an annual average rate of 4.4% over the period 2002 to 2015, according to forecasts prepared by the ICAO (2004). To accommodate predicted traffic growth there is a need to increase *en route* airspace capacity through the introduction of new air traffic management systems, controller tools and procedures. But, the consensus among research and operational communities is that it is really important to understand the factors that drive mental workload if they are to improve airspace capacity [3], [4]. Most research has focused on identifying characteristics of the air traffic picture that create task demand for ATCOs (e.g., [5]–[7]). Others argue that there is no simple linear relationship between task demand and workload (e.g., [8], [9]). Several current research groups agree with Sperandio's [10] view that a relationship between task demand and workload can be better understood by considering how ATCOs use strategies to manage their resources and regulate their workload [4], [8], [11]–[14]. Factors such as skills, training, experience, fatigue and other "stressors" all mediate the relationship between task demands, safety and performance of the ATCO. Hence, it is easy to understand how quantitative information about skills level and mental states could help to evaluate the ATCO's workload level and to decide if they might need more training before working into real environments. Several studies described the correlation of spectral power of the EEG bands with the complexity of the task that the subjects are performing [15]. In fact, an increase of electroencephalographic power spectral density especially over the frontal cortex in the theta band (4 - 7 Hz) and an EEG PSD decrease in the alpha band (8 - 12 Hz) over the parietal cortex have been observed when the required mental workload, the task's complexity and the amount of information to be processed increase. Furthermore, it has been suggested that also an increased Heart Rate (HR) could be related with an increased mental workload and engagement. On the contrary the eyesblink duration and rate are inversely correlated with the increase of the mental workload and attention levels [15]. The hypotheses of the study are that i) as the EEG theta PSD increases and the EEG alpha PSD decreases with these cognitive phenomena , at the end of the training period such PSDs increment and decrement should be different from the beginning of the training period , therefore such trends could be taken as indexes of the correct acquisition of procedural skills and of less request of cognitive resources for the correct execution of the task, ii) the combined use of EEG features and HR can improve the reliability of the measure with respect of using the single information. Also, these biosignals can be used for the real-time evaluation of the operator's workload level, in a real Air Traffic Management (ATM) scenarios. Such hypotheses have been tested on a group of 6 subjects who succeeded in the 5-days-training-period and who were asked to execute the experimental task for two more weeks in order to evaluate their mental workload under three different difficulty levels.

MATERIALS AND METHODS

Experimental subjects and ATM simulation task

A group of six healthy volunteers has been selected in terms of age (21±4 years) and previous computer game skills. The subjects have been asked to learn how to execute correctly an ATM task (labyrinth, LABY), that never did before, under easy (E), medium (M) and hard (H) conditions, randomly selected and proposed to them in order to avoid any habituation and expected effects. The LABY microworld is a functional simulation of Air Traffic Control (ATC), provided by the software engineers of the French ENAC, that captures the underlying processes involved in electronic air traffic management with a simplified version of the operational human-machine interface. Microworlds are computer-based human-in-the-loop simulation environments that offer testing, behavioral/physiological measurement, and training capabilities, with the flexibility to build various scenarios [16], [17]. LABY is a dynamic environment whereby a controller must issue directional commands to guide aircrafts along a predetermined route, whilst avoiding potential conflicts and dealing concurrently with other incoming information. The LABY microworld is based upon the main task of guiding *N* plane(s) around a predetermined route, indicated by a green path (Fig. 1). Participants must input numerical values such as heading, flight level, speed, etc., in order to direct flight around the trajectory and to avoid any conflicts or obstacles which may occur during the flight-route. Penalties are applied if the aircrafts deviate off the route or if other constraints are not met. The difficulty of the task can be altered according to how many aircrafts the participant have to control, the number and type of clearances required over the time and the number/trajectory of other interfering flights. In the first week, the subjects trained for 5 consecutive days (SESSIONS T1÷T5) and their neurophysiological signals have been recorded in the first (T1), in the third (T3) and in the fifth (T5) session. The behavioral (task performance) and subjective workload perception data have been collected every day. After the training period, the subjects were asked to execute the LABY on two consecutive days in the second week and on a day after a week since the last experimental session. At the end of each experimental condition, the subjects filled the NASA-Task Load indeX (TLX, [18]) questionnaire for the evaluation of the perceived workload of the proposed task.

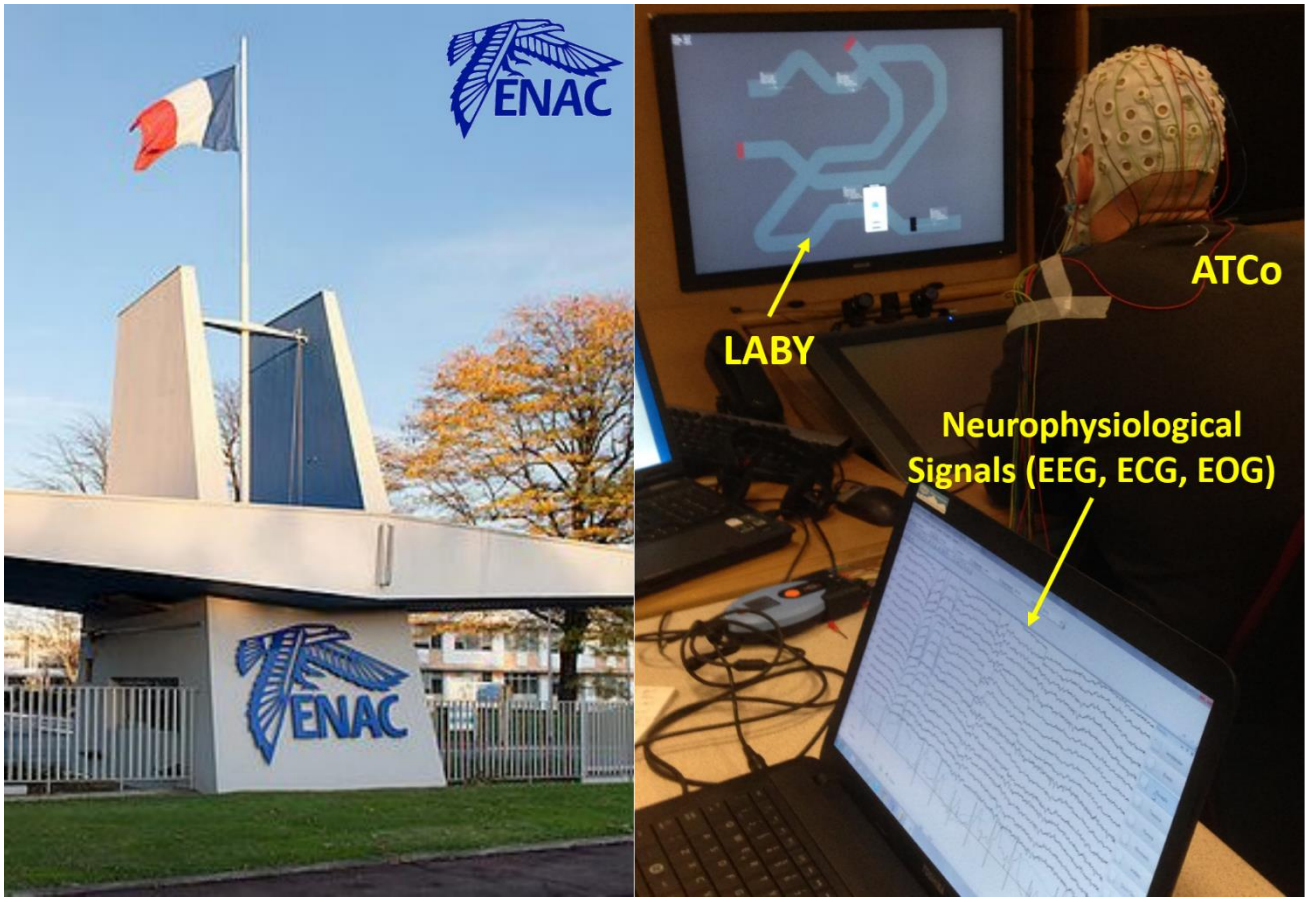


Figure 1.

Acquisition of the brain activity and of the physiological signals: Electroencephalogram (EEG) and physiological signals, including vertical electrooculogram (EOG) and electrocardiogram (ECG), have been recorded by the digital monitoring *BEmicro* system (EBNeuro system). The sixteen EEG channels (FPz, F3, Fz, F4, AF3, AF4, C3, Cz, C4, P3, Pz, P4, POz, O1, Oz and O2), the ECG and the EOG channels have been collected simultaneously with a sampling frequency of 256 (Hz). All the EEG electrodes have been referenced to both earlobes, and the impedances of the electrodes were kept below 10 (kΩ). The bipolar electrodes for the heart activity have been placed on the Erb's point, while the bipolar electrodes for the EOG have been positioned vertically on the left eye.

EEG analysis: The acquired EEG signals have been digitally band-pass filtered by a 4^{th} order Butterworth filter (low-pass filter cut-off frequency: 30 (Hz), high-pass filter cut-off frequency: 1 (Hz)) and then segmented in epochs of 2 seconds, 0.125 seconds – shifted. The EOG signal has been used to remove eyes-blink artefacts from the EEG data by using the Gratton method [19]. The EEG PSD has then been estimated by using the *Fast Fourier Transform* (FFT) in the EEG frequency bands defined for each subject by the estimation of the *Individual Alpha Frequency* (IAF) value [20]. The PSDs in the theta and alpha bands have then been analyzed by estimating the *Coefficient of Determination* (r^2), or *r-square*, between the considered experimental condition and the reference condition. As $0 < r^2 < 1$ by definition, a signed r^2 has been derived by multiplying the coefficient of determination by the sign of the slope of the corresponding linear model of the regression analysis. In this way, it has been possible to obtain information not only about if the two datasets were different, but also about the direction of such difference. A Stepwise Linear Discriminant Analysis (SWLDA, [21]) has been used to select the most relevant spectral features to discriminate the mental workload levels. In particular, the classifier was trained using data from one triplet (Easy, Medium and Hard) and the extracted parameters were tested over the other remaining triplets within the same session (INTRA cross-validations) or the other sessions (INTER cross validations). Several moving average samples (N_{MA}) have been applied to the output of the classifiers ($W_{EEG,}$): $N_{MA}(1) = 0.125$ (sec), $N_{MA}(8) = 1$ (sec), $N_{MA}(16) = 2$ (sec), $N_{MA}(32) = 4$ (sec), $N_{MA}(64) = 8$ (sec). The moving average was expected

to increase the stability and the accuracy of the index with the drawback of introducing delays in the workload estimation, inducing a decrease of the workload refresh rate.

ECG and EOG analysis: As well as for the EEG, the ECG and the EOG signals have been band - pass filtered, respectively 1-8 (Hz) and 8-16 (Hz), and then segmented in epochs of 8 seconds, 0.125 seconds – overlapped. The HR and the EBR have been estimated by calculating the distance between consecutive peaks occurring in the ECG and in the EOG signals. In particular the R-peaks and the eyeblinks peaks have been used, and then they have been normalized by the z-score transformation with respect to the reference condition, in which the subjects watched the task interface without responding to them. As for the EEG, also for the HR parameter a workload index (W_{HR}) has been calculated by using the SWLDA at different output rates: $N_{MA}(1) = 0.125$ (sec), $N_{MA}(8) = 1$ (sec), $N_{MA}(16) = 2$ (sec), $N_{MA}(32) = 4$ (sec), $N_{MA}(64) = 8$ (sec).)

Fusion workload index: A Fusion workload index has been calculated as a combination of the W_{EEG} and the W_{HR} indexes. In particular, the two SWLDA classifiers outputs have been synchronized, because their different delays (EEG: 2 (sec) overlapped of 125 (msec); HR: 8 (sec) overlapped of 125 (msec)), and then a new workload index (Fusion based workload index, W_{Fusion}) has been computed as a linear combination of the W_{EEG} and the W_{HR} score (Equation 1).

$$W_{Fusion} = aW_{EEG} + bW_{HR} \quad (1)$$

The coefficients *a* and *b* of the linear combination have been estimated for each subject by means of a simple LDA performed considering the EEG and the HR score distributions (W_{EEG} and W_{HR}) calculated over the cross validations for the three different difficulty levels (Figure 2).

Classifier performance analysis: The dataset deriving from the three days of workload evaluation

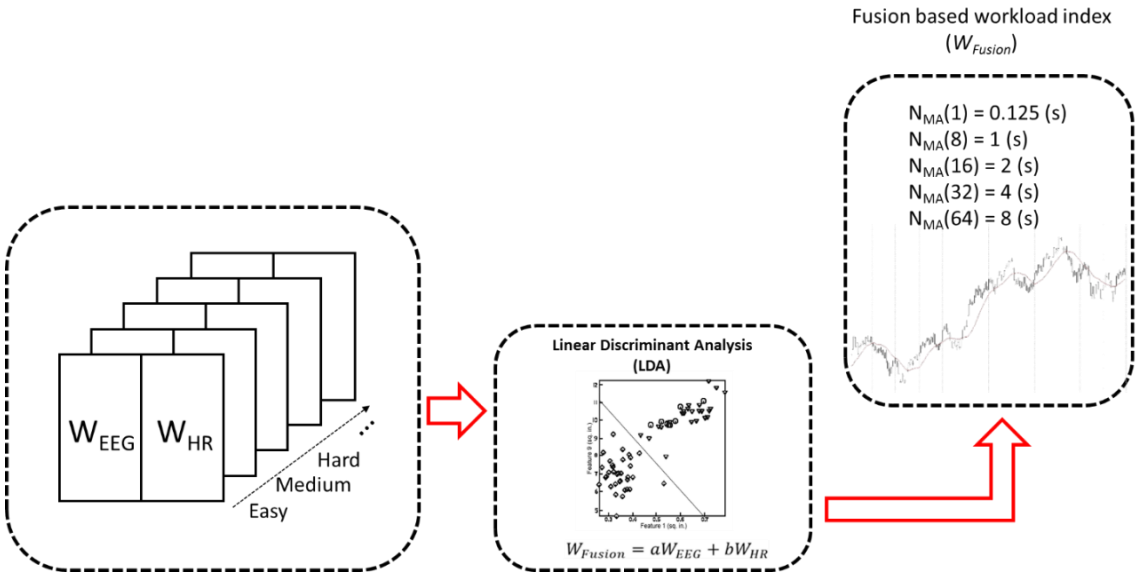


Figure 2.

have been re - organized in 15 triplets (5 triplets per session) of runs (Easy, Medium and Hard). All the possible cross-validations have been considered, training the classifier with one triplet and testing the extracted features over the remaining triplets. The values of the Area Under Curve (AUC) of the Receiver Operating Characteristic (ROC, [22]), describing the accuracy of the system, have been calculated from the outputs of the classifier (for each of the different output rates).

Workload score distributions analyses: The workload score distributions of the single conditions (Easy, Medium and Hard) has been calculated using the same approach of the AUC evaluation, thus

by training the classifier with each triplet of runs within the sessions and testing the extracted features over all the other triplets. In addition, two types of cross-validations have been defined in order to investigate how well the classifier performs; considering the training and the testing dataset of the same day (INTRA) or considering the training set of one day and the testing set of the other days (INTER). For summarize, the INTRA type refers to the cross-validations performed considering as training and testing sessions the same day. Contrariwise, the INTER type refers to the cross-validations performed considering as training session one of the three days and as testing sessions those performed in the other two days.

NASA-TLX analyses: Subjective workload perception was obtained by asking the subjects to fill the NASA-TLX questionnaire for each task condition (Easy, Medium and Hard). The NASA-TLX evaluates the perceived workload by considering six different factors: Mental Demand, Physical Demand, Temporal Demand, Frustration, Effort and Performance. The workload scores, ranged from 0 to 100, are obtained as weighted linear combination of such factors. The subjective scores of the perceived workload were then compared to the mental workload indices estimated by the system.

Statistical analyses: The results derived from the different analyses have been then validated by the statistical analyses performed by using the STATISTICA software (Statsoft). For the Training Protocol, the one-way repeated measures ANOVA (Confidence Interval, CI = .95) was used for all the neurophysiological data (dependent variables) with the SESSIONS (3 levels) as independent variable. Such factor has three levels, one for each day of the week in which the EEG recording was made (T1, T3 and T5). For the Workload Protocol, statistical analyses over the i) classifier performances, ii) workload scores distribution and iii) NASA-TLX scores have been performed.

i) Three repeated measures ANOVA (CI = .95) have been performed, one for each classifier (EEG, HR and Fusion based), using the different comparison couples of difficulty levels (Easy vs Hard, Easy vs Medium and Medium vs Hard, 3 levels), and two of the "moving average lengths" ($N_{MA}(x)$, x={1, 64}, 3 levels) as within factors and the related AUC values as dependent variable, for all the subjects. We selected only two moving average values because the number of subjects. Also, a repeated measures ANOVA has been performed by considering the three classifiers (EEG, HR and Fusion based, 3 levels) and the two moving average lengths ($N_{MA}(x)$, x={1, 64}, 3 levels) as within factors, and the AUC values averaged for the three couples of conditions (Easy vs Hard, Easy vs Medium and Medium vs Hard) as dependent variables for all the subjects. In addition, a Duncan post-hoc test has been performed in order to test the effects between all the factors.

ii) Three repeated measures ANOVA (CI = .95) have been performed, one for each classifier (EEG, HR and Fusion based), using the difficulty conditions (Easy, Medium and Hard, 3 levels) and Cross-validation type (INTRA and INTER, 2 levels), for each subject, as independent variables and the related workload index distributions (W_{EEG}, W_{HR} and W_{Fusion}) as dependent variables..

iii) A one-way ANOVA (CI=.95) was performed on the NASA-TLX scores (dependent variable) with the difficulty condition (Easy, Medium and Hard) as independent variable.

RESULTS

Training improvement assessment

LABY performance analysis: Throughout the training sessions, the performance of the subjects increased continuously in terms of mean performance level and accuracy. Figure 3 shows the performance's index adopted across the different training days. By the inspection of Fig. 3 it is easy to note the simultaneous increase of the performances level and the decrease of the amplitude of the standard deviations in the learning curve. On the second day of training, all the subjects reached at a good level of performance (almost the 90%) and since the third day (T3), they could reach performance level higher than 95%. The one-way ANOVA performed on the global LABY score showed significant differences across the sessions ($F(4, 20) = 17.74$ with a $p < .00001$, $\eta^2_p = .78$). The Duncan post-hoc test showed that the first two sessions (T1 and T2) were statistically different from all the others ($p < .0001$) while the last three ones (T3, T4 and T5) were not statistically different to each other.

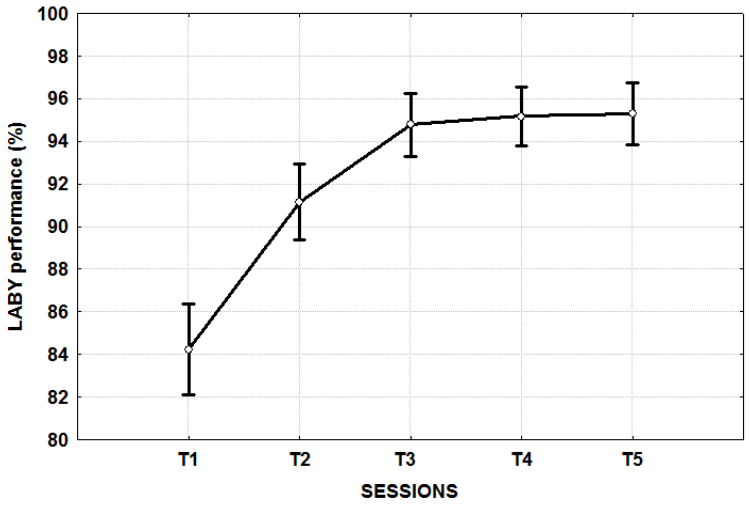


Figure 3.

Frontal PSD theta: The ANOVA results reported in Figure 4 show a statistical significant modulation of the of EEG PSD in theta band over the frontal areas (EEG channels: AF3, AF4, F3, Fz, and F4) across the different training sessions ($F(2, 10) = 4.18$, $p < .048$, $\eta^2_p = .45$) and the Duncan's post-hoc test confirmed these differences $p < .03$. It is evident that in the central session (T3), when the subjects have been supposed to have learnt how to execute correctly the task and focused the cognitive resources for improve their performances, the frontal PSD theta reached the highest increment respect all the other sessions.

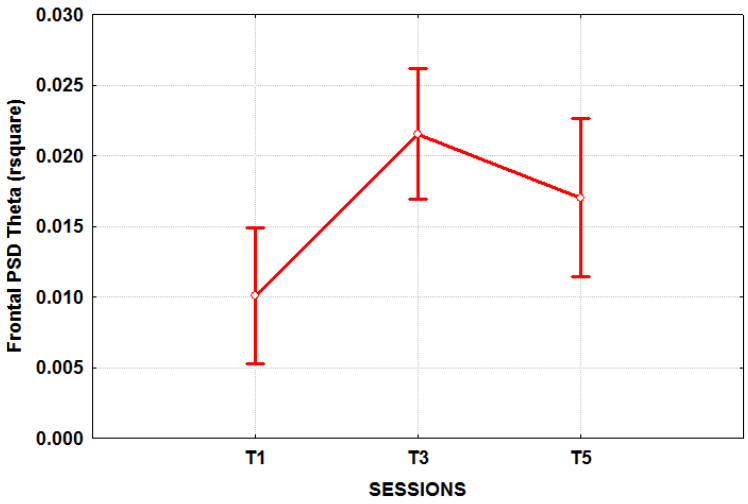


Figure 4.

Parietal PSD alpha: Figure. 5 shows the trend of the parietal EEG PSD in alpha band over the EEG channels P3, Pz and P4, represented as variation of signed r-square. Repeated measures ANOVA showed significant differences of the parietal PSD alpha ($F(2, 10)=9.95$ with an associated $p = .0042$, $\eta^2_p=.67$) and a decreasing trend of the spectral PSD from T1 to T5 across the training sessions.

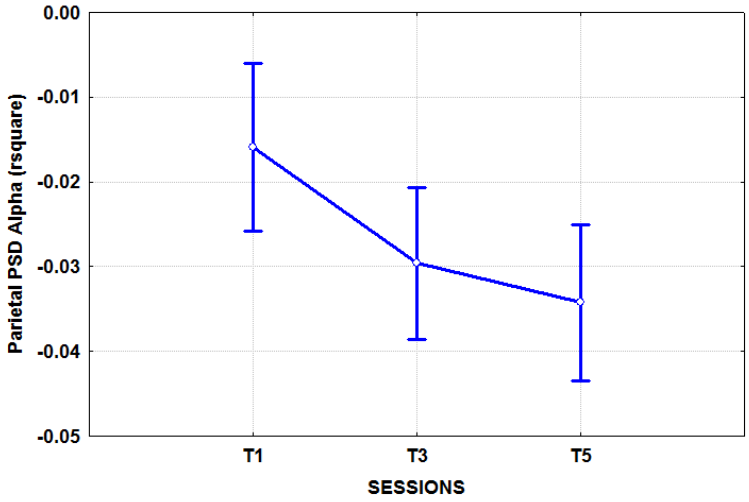


Figure 5.

Heart Rate and Eyeblinks Rate analysis: Figure 6 and 7 show the results of the ANOVA analysis of the autonomic parameters of HR and of EBR, respectively. The HR shows that the subjects were emotively engaged in correspondence of the central training session (T3), as the HR in T3 was the highest one, and that at the end of the training period they were more relaxed with the experimental task, as both the HR and the EBR decreased and increased, respectively.

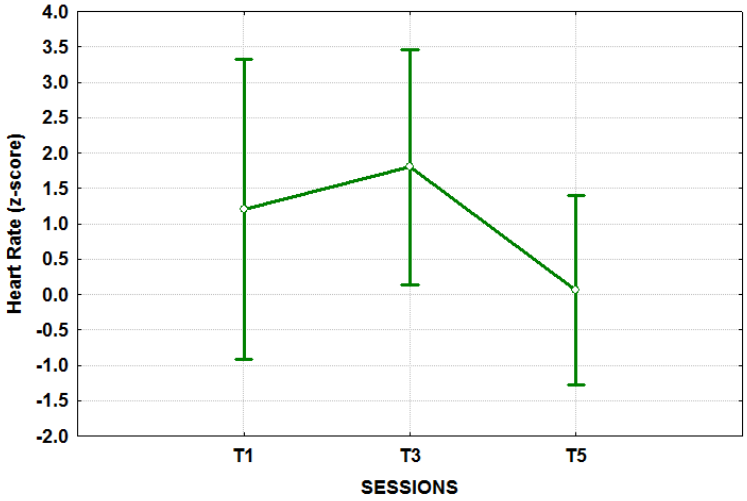


Figure 6.

In fact, the Duncan's post-hoc tests reported significant ($p<.01$) differences between the HR and EBR values of the first (T1) and last (T5) training session. In addition, the EBR z-score shows how the subjects kept to pay attention to the task, as it was negative even at the end of the training.

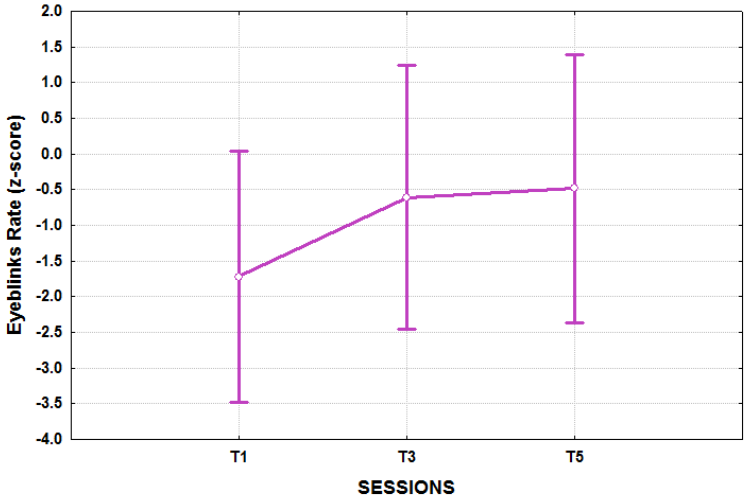


Figure 7.

Perceived mental workload: NASA-TLX analysis: The one-way ANOVA for the NASA-TLX data (Figure 8) shows significant differences among the training sessions ($F(4, 20)=7.67$ and $p< .00065$, $\eta^2_p=.61$). The post-hoc test allowed to found out that the average scores of the NASA-TLX were statistically different up to the fourth session (T4), whereas the T4 and T5 sessions were perceived similar in terms of perceived workload.

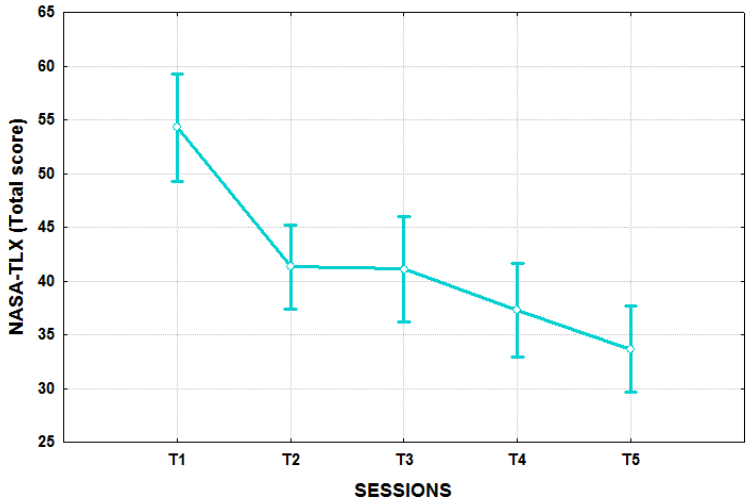
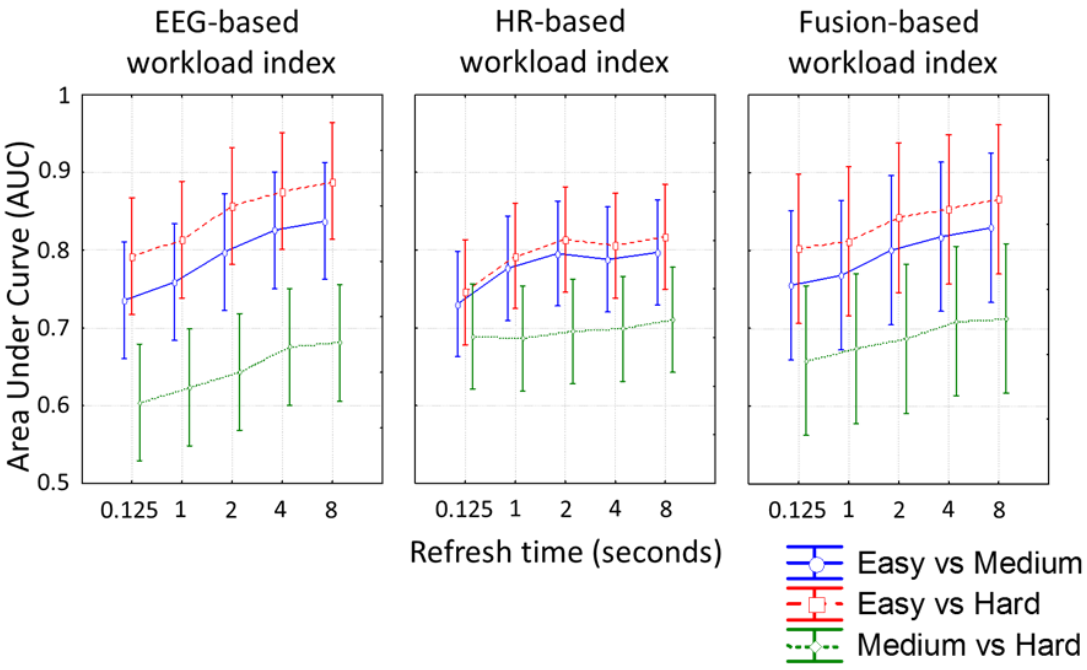


Figure 8.

Workload evaluation and classification

Classifier performance analyses: The ANOVA analyses (Figure 9) revealed a significant increment in the classifier' performance at higher moving average values for both the EEG ($F(1, 5)=372.14$, $p=0.00001$, $\eta^2_p=.99$) and the Fusion based classifiers ($F(1, 5)=18.13$, $p=0.008$, $\eta^2_p=.78$), while no differences have been highlighted for the HR based classifier ($F(1, 5)=1.34$, $p=.29$, $\eta^2_p=.21$). In addition, the post-hoc test showed that AUC values calculated using all the three classifiers in the "Medium vs Hard" couple were significantly lower (all $p<.001$) than the other two ones. Finally, the ANOVAs highlighted a significant interaction between the Classifiers and the Moving averages ($F(2, 10)=25.71$, $p=0.0001$, $\eta^2_p=.84$). In particular, post-hoc test showed that the fusion based classifier (W_{Fusion}) showed higher AUC values than the EEG based classifier (W_{EEG}) at short refresh times (0.125s, $p<.001$), and higher AUC values than both the EEG and the HR classifiers at long refresh times (8s, $p<.05$). (Figure 9).



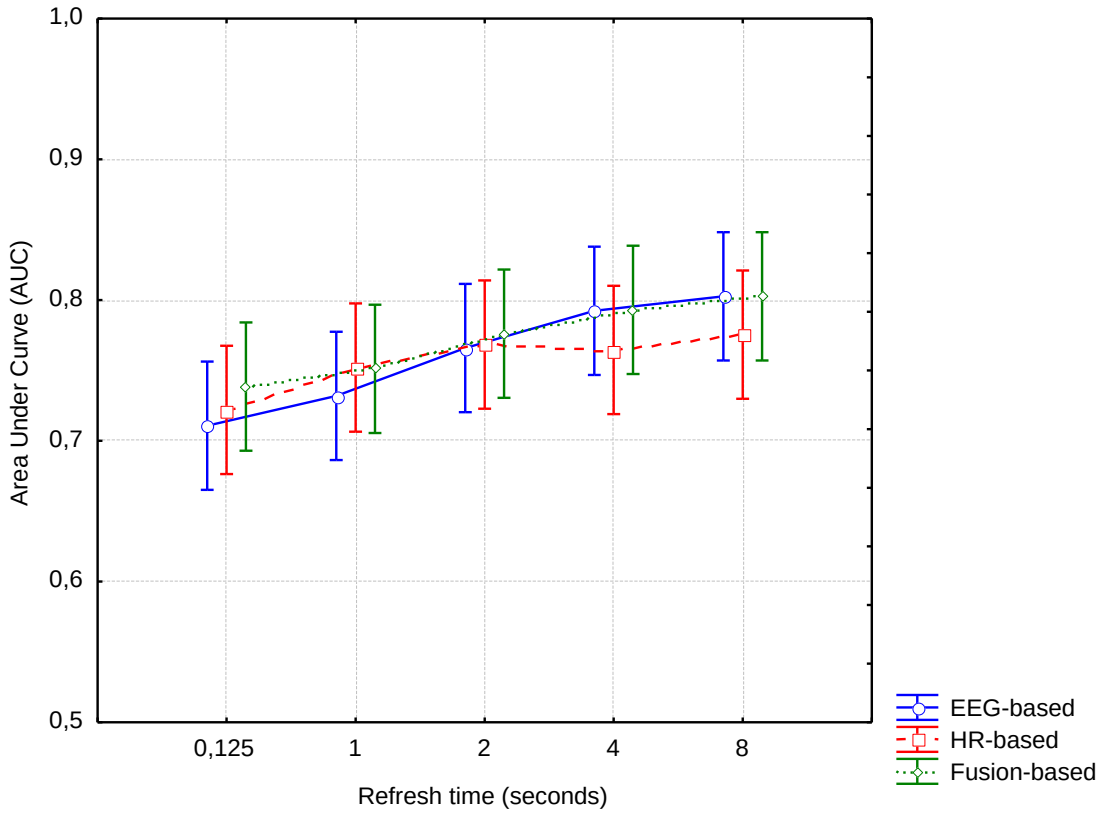


Figure 9.

Workload score distribution: The ANOVA analyses (Figure 10) revealed that the score distributions related to the different difficulty conditions (Easy, Medium and Hard) for all the three classifiers were significantly separated (EEG-based: $F(2,10)=8.90$, $p=.006$, $\eta^2_p=.64$; HR-based: $F(2,10)=4.93$, $p=.032$, $\eta^2_p=.50$, Fusion-based: $F(2,10)=10.05$, $p=.004$, $\eta^2_p=.67$). Furthermore, no significant differences were found between the workload scores related to the INTER and the INTRA cross-validations (EEG-based: $F(1,5)=.22$, $p=.66$, $\eta^2_p=.04$; HR-based: $F(1,5)=1.40$, $p=.29$, $\eta^2_p=.22$, Fusion-based: $F(1,5)=1.78$, $p=.24$, $\eta^2_p=.26$).

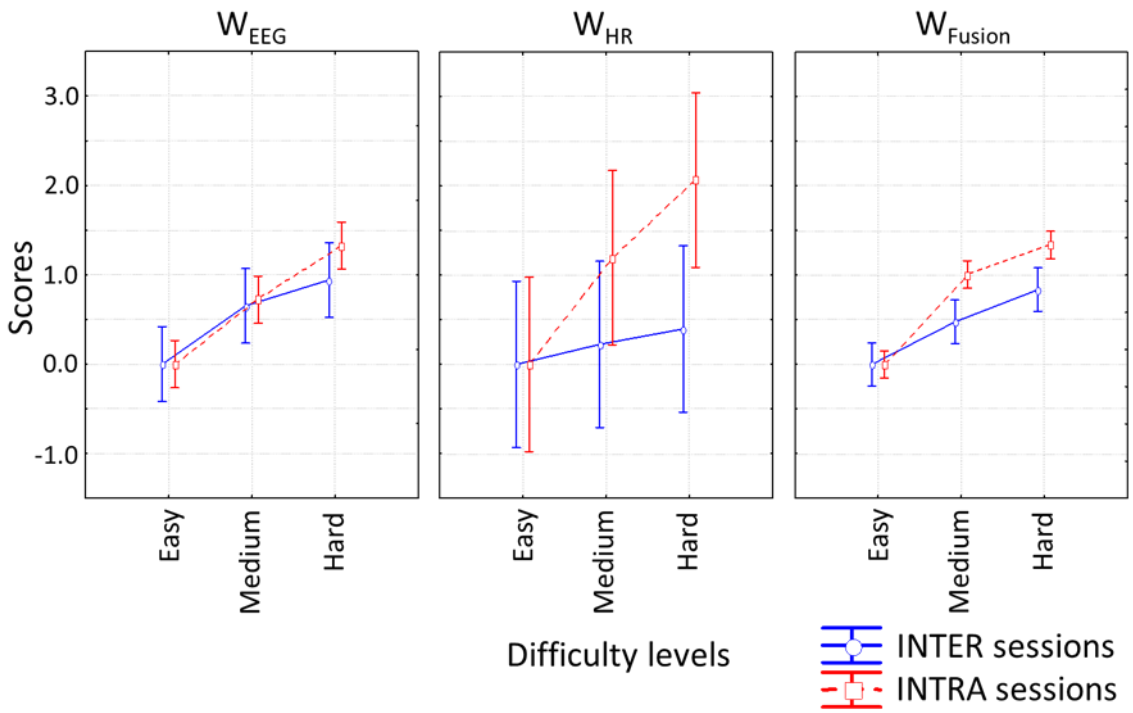


Figure 10.

NASA-TLX analyses: The ANOVA results revealed a main effect of the difficulty levels ($F(1,10)=8.38$, $p=.016$, $\eta^2_p=.46$). Subjective perception of the workload increased as the difficulty of the task increased. This result is consistent with the score distribution analyses (Figure 10), showing a high reliability of the estimated mental workload index.

DISCUSSION

The neurophysiological signals, the task performance scores and the experienced workload describe a story in which the subjects found their own strategies for the correct execution of the proposed task (LABY) and then got completely confident with the execution of it. At the central part of the training period (T3) the cognitive and emotive engagement reached the highest levels, as the frontal PSD theta and the HR showed the highest increment. The trends of the parietal PSD alpha and of the EBR showed how the subjects kept to pay attention to the execution of the task. In fact, the parietal PSD alpha and the EBR continued to decrease and to increase, respectively, up to the last training session of the first week (T5). From a perception point of view, the NASA-TLX scores demonstrated that the subjects experienced less workload throughout the sessions, especially at the end of the training period (T5) respect to the beginning of it (T1).

Once the subjects became trained with the LABY task, an algorithm able to estimate in real-time the mental workload of the user has been tested on the next two weeks, by using the combination of EEG and ECG signals. There are several studies in literature in which these biosignals are used for assessing the mental workload, but normally they are used separately. In this study, it has been demonstrated that the combination of EEG and HR allows to differentiate significantly the workload level over three different difficulty levels, showing a high discrimination accuracy (AUC > .7). Furthermore, the fusion of the EEG and HR information allows to significantly increase the reliability of the algorithm with respect of using only EEG or HR alone. Also, the subjective features used for the evaluation of the mental workload remained stable over a week, between the "week 2" and "week 3" of the experimental protocol., so, even after a week, it might not be necessary to recalibrate the algorithm with new EEG data. The aspects related to the classifier stability and

accuracy are highly important for the usability of the system. In fact, to use such system in real environments, it could be enough to calibrate the algorithm with the specific parameters of the operator only once and then just use them without further adjustments maintaining a high reliability and stability over, at least, a period of one week. Finally, the calculated workload index showed the same trend of the NASA-TLX workload assessment.

CONCLUSIONS

Two protocols have been presented in this study, the training and the workload evaluation of ATCOs by means of neurophysiological signals. The integration of information derived by the brain activity, through the EEG, and the physiological signals of ECG and of EOG with the supervision of Experts could be used as possible innovative "cognitive metric" for evaluating the degree of the learning and the training progress throughout their periods of formation. Also, this method could be applied when the comparison between subjects is required in terms of required cognitive resources for the execution of specific tasks. In fact, after a fixed period of training it could be possible i) to quantify how well the subjects can complete a task, in terms of cognitive resources necessary to the correct execution, and ii) to compare subject's cognitive performances by estimating the neuro-physiological EEG, HR and EBR parameters presented in this study. In addition, an algorithm able to estimate the mental workload of an operator by using the combination of EEG rhythms and HR signals has been proposed. It has been demonstrated that i) the system is able to significantly differentiate three workload levels related to the three difficulty level of the task employed; ii) the subjective features used for the evaluation of the mental workload remain stable over a week; iii) the combination between the information derived from the EEG and the HR signals allows to significantly increasing the reliability of the system. Finally, iv) the subjective evaluation of the workload shows the same trend of the physiological workload indexes (W_{EEG}, W_{HR}, W_{Fusion}). The evaluation of the mental workload by using the information derived by biosignals, allows to have an objective and more reliable measure than using the subjective questionnaires, such as the NASA-TLX. Also, another advantage with respect to the subjective measurements (e.g. NASA-TLX) is the assessment of the real-time variations of the workload within the same task, e.g. each 125 msec. Further experiments will be performed to even further test and extend the long term use of the algorithm. The present study, carried out in a laboratory environment, should be replicated on a larger sample size (more than six subjects) and in a more realistic scenario involving professional operators.

REFERENCES

1. Gronlund SD, Ohrt DD, P R, Perry JL, Manning CA. Role of memory in air traffic control. Journal of Experimental Psychology: Applied. 1998;4(3):263–80.
2. Rodgers MD, Drechsler GK. Conversion of the CTA, Inc., en route operations concept database into a formal sentence outline job task taxonomy. Wash-ington, DC: Federal Aviation Administration Office of Aviation Medicine.; 1993.
3. Christien R, Benkouar A, Chaboud T, Loubieres P. Air traffic complexity indicators ATC sectors classification. Digital Avionics Systems Conference, 2002 Proceedings The 21st. 2002. p. 2D3–1–2D3–7 vol.1.
4. Majumdar A, Ochieng WY, McAuley G, Michel Lenzi J, Lepadatu C. The Factors Affecting Airspace Capacity in Europe: A Cross-Sectional Time-Series Analysis Using Simulated Controller Workload Data. The Journal of Navigation. 2004;57(03):385–405.
5. Grossberg M. Relation of sector complexity to operational errors (Quarterly Rep. of the FAA Office of Air Traffic Evaluations and Analysis). Washington, DC: Federal Aviation Administration. (Quarterly Rep. of the FAA Office of Air Traffic Evaluations and Analysis). Washington, DC: Federal Aviation Administration; 1989.
6. Kirwan B, Scaife R, Kennedy R. Investigating complexity factors in UK air traffic management. Human Factors and Aerospace Safety. 2001 Jul.
7. Manning CA, Mills SH, Fox C, Pfleider E, Mogilka HJ. Investigating the Validity of Performance and Objective Workload Evaluation Research (POWER). 2001 Jul.
8. Athènes S, Averty P, Puechmorel S, Delahaye D, Collet C. Complexity and controller workload: Trying to bridge the gap. International Conference on Human-Computer Interaction in Aeronautics. Cambridge: Massachusetts Institute of Technology; 2002. p. 56–60.
9. Chatterji G, Sridhar B. Measures for air traffic controller workload prediction. 1st AIAA, Aircraft, Technology Integration, and Operations Forum. American Institute of Aeronautics and Astronautics. 2001.

10. Sperandio JC. Variation of Operator's Strategies and Regulating Effects on Workload. Ergonomics. 1971;14(5):571–7.
11. Averty P, Athenes S, Collet C, Dittmar A. Evaluating A New Index Of Mental Workload In Real ATC Situation Using Psychophysiological Measures. Proceedings of the 21 st Digital Avionics Conference. Forthcoming; 2003. p. 1–13.
12. Cullen L. Validation of a methodology for predicting perfor-mance and workload (European Experimental Centre Note No. 7/99). Brussels, Belgium: EUROCONTROL. 1999.
13. Hilburn B. Cognitive complexity in air traffic control: A litera-ture review (EUROCONTROL Experimental Centre Note No. 04/04). Brussels, Belgium: EUROCONTROL.
14. Histon J. The Impact of Structure on Cognitive Complexity in Air Traffic Control. 2002 Jun.
15. Borghini G, Astolfi L, Vecchiato G, Mattia D, Babiloni F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neurosci Biobehav Rev. 2012 Oct 30;
16. Brehmer B, Dörner D. Experiments with computer-simulated microworlds: Escaping both the narrow straits of the laboratory and the deep blue sea of the field study. Computers in Human Behavior. 1993;9(2–3):171–84.
17. Gonzalez C, Vanyukov P, Martin M. The Use of Microworlds to Study Dynamic Decision Making. Department of Social and Decision Sciences [Internet]. 2005 Jan 1; Available from: http://repository.cmu.edu/sds/35
18. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. Human Mental Workload. North-Holland; 1988. p. 139–83.
19. Gratton G, Coles MG, Donchin E. A new method for off-line removal of ocular artifact. Electroencephalogr Clin Neurophysiol. 1983 Apr;55(4):468–84.
20. Klimesch W. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. Brain Res Brain Res Rev. 1999 Apr;29(2-3):169–95.

21. Aricò P, Borghini G, Graziani I, Bianchini F, Cincotti F, Babiloni F. A brain computer interface system for the online evaluation of ATCs' workload. Italian Journal Of Aerospace Medicine. 10th ed. Rome, Italy; 2013 Jun.
22. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology. 1975 Nov;12(4):387–415.

FIGURE LEGENDS

Figure 1. The LABY is a dynamic environment whereby an ATC must issue directional commands to guide *N* airplane(s) around a predetermined route, indicated by a green path, in order to avoid any conflicts or obstacles which may occur during the flight-route. LABY has been developed and tested with a professional ATCO in collaboration with ENAC in Tolouse.

Figure 2. Fusion based workload index assessment (W_{Fusion}). The Fusion workload index (W_{Fusion}) has been calculated as a linear combination of the EEG and the HR based workload indices. The two classifiers outputs were synchronized before the computation of the fusion-based index.

Figure 3. The trend of the global LABY score across the five different training sessions (T1-T5). The figure reports the mean performance value and the standard deviations for the sessions. A statistical significant increase of the performance was obtained at the end of the period when compared to the first day of training

Figure 4. Mean EEG PSD (r-square) in theta band over the frontal EEG channels AF3, AF4, F3, Fz and F4 across the training sessions T1, T3 and T5. At T3, the frontal PSD theta reached the highest increment ($p < 10^{-5}$).

Figure 5. Parietal EEG PSD in alpha frequency band during the training period. The graph reports the signed r-square values estimated in the training sessions (T1, T3 and T5). The continuous decrement of the parietal PSD alpha is significant across all the training sessions ($p<10^{-5}$).

Figure 6. Heart Rate (z-score) values across the training sessions. The trend shows how in the central part of the training period (T3) the subjects showed an high emotive engagement, as the HR got the highest value.

Figure 7. Eyesblink rate (z-score) values across the training sessions. The values are all negative because the subjects paid attention to the task for the whole training period and it shows how the subjects got more confident with task session after session.

Figure 8. Average NASA – TLX scores of the training sessions. After each training session the subjects perceived the difficulty of the experimental task easier than the previous one.

Figure 9. Mean values and related standard errors (CI = .95) of the AUC values achieved using the different classifiers (EEG, HR and Fusion-based) for each refresh time value.

Figure 10. Mean values and related standard errors (CI = .95) of the distributions of the workload indices (W_{EEG}, W_{HR} and W_{Fusion}) evaluated by the three classifier (EEG, HR and Fusion based).