Using AI and Neuroscience in Immersive 3D Flight Simulation Device to Accelerate Pilot Training

Jean-François Delisle
Hamza Nabil
CAE Inc.
Montréal, Québec, Canada
jean-francois.delisle@cae.com
hamza.nabil@cae.com

Theophile Demazure,
Pierre-Majorique Léger
HEC-Tech3lab
Montréal, Québec, Canada
theophile.demazure@hec.ca
pierre-majorique.leger@hec.ca

ABSTRACT

To improve and accelerate pilot training, this paper explores the capture of cognitive and psychophysiological states using biometric sensors and flight telemetry to drive an intelligent human performance assessment system for adaptive learning. Specifically, we explore the neuroscience capabilities that could enable real-time adaptive flight training using a variety of data collected from a flight training session. Assessing pilot performance during a training session is a capability that can be partially performed by an AI-based algorithm. With technical data gathered during a flight manoeuvre, such assessment can provide objectivity during flight training, can be a predictor of future pilot performance, and adapt simulation training using a combination of flight telemetry (technical skills) and biometric/behavioural data (non-technical skills). Evaluation of non-technical skills remains difficult without the support of data analytics and proper visualization tools. Additionally, soft skills are inherently more difficult to grade compared to technical performance. An AI engine can provide cues on behaviours and cognitive/psychophysiological states that cannot be easily observed by the instructor. We conducted an experiment with 16 novice pilots in a fast-jet flight simulator with an e-Series Medallion visual. During the simulated flight, we recorded a wide range of neurophysiological including electroencephalogram (EEG), eye-tracking device, and flight telemetry. N-Back, BART & IGT cognitive paradigms were used prior to the simulation as a baseline. We present a concept using artificial intelligence to improve the cognitive load computing with the constraint of non-intrusiveness of biometric sensors. We also present the design of the experiment protocol as a training scenario during initial training. With an increasing difficulty scenario, we assessed performance based on criteria and thresholds and we provide results of a performance analysis.

ABOUT THE AUTHORS

Jean-François Delisle, AI Innovation Lead, Advanced Technology & Innovation , CAE Inc., Jean-François has over 25 years of experience in software engineering, artificial intelligence and data solutions. Focusing his research on AI solution that helps optimize training solutions based on human behaviour and performance. He joined CAE in 2010 and his current mandate is to define flight training solutions using data analysis and artificial intelligence capabilities in advance air mobility and eVTOL engineering solutions. He has a PhD in artificial intelligence for adaptive flight training.

Theophile Demazure, Théophile Demazure is a Ph.D. candidate in Information Technology at HEC Montréal under the NSERC-Prompt Industrial Research Chair in User Experience. Théophile published in multidisciplinary outlets such as Frontiers in Human Neuroscience, Business & Information Systems Engineering, or NeuroIS. His research explores the human factor and neurophysiological side of human-computer interaction in performance-oriented environments.

Hamza Nabil, Hamza Nabil is a Data Analyst professional, with a Bachelor's degree specialized in aerospace engineering. He joined CAE in 2019 within a leadership rotation program to pursue a career in data. His mandate was to research and analyze specific patterns between the correlation of biometric data and flight telemetry of pilots using various statistical models and to also automate the performance of each pilot using an objective score. He has a Master's degree in Mechanical Engineering specialized in Computational Fluid Dynamics.

Pierre-Majorique Léger, Professor Pierre-Majorique Léger is the senior chairholder of the NSERC-Prompt Industrial Research Chair in User Experience (UX). He is a full professor of IT at HEC Montreal. He holds a Ph.D. in industrial engineering from École Polytechnique de Montréal and has done post-doctoral studies in information technologies at HEC Montréal and NYU Stern School of Business. He is the co-director of the Tech3Lab, a research laboratory and devoted to the investigation of human factors in information technologies.

Using AI and Neuroscience in Immersive 3D Flight Simulation Device to Accelerate Pilot Training

Jean-François Delisle
Hamza Nabil
CAE Inc.
Montréal, Québec, Canada
Hamza.nabil.cae.com
jean-francois.delisle@cae.com

Theophile Demazure,
Pierre-Majorique Léger
HEC-Tech3lab
Montréal, Québec, Canada
theophile.demazure@hec.ca
pierre-majorique.leger@hec.ca

INTRODUCTION

The objective of this paper is to explore artificial intelligence capability & human factors in the context of pilot training in a flight simulation environment by using Electroencephalography (EEG), eye trackers, Facial Emotion recognition, and Flight Telemetry data with N-Back, BART & IGT paradigms. Biometrics analytics and data science methodologies are used to determine pilot performance & human factors associated with cognitive workload and decisions involving risk in a mission rehearsal using a 3D immersive simulation environment. In other words, our objective is to explore autonomous artificial intelligence capability that could enable adaptive flight training using a variety of data collected from a full-flight simulator and biometric sensors data. An intelligent adaptive flight-training system shall consider the human mental state in the learning process of the pilot. A better understanding of cognitive workload, risk-taking behaviour, and immersion levels are essential aspect to succeed in the real-time adaptation.

The risk-taking behaviour of the pilot is a flight safety issue considering that the reward effect can incite pilots to make riskier decisions. This behaviour is very hard to reproduce during flight simulation and pilots know that they are in a simulation, so their risk tolerance is significantly higher. By increasing the immersion level of a training device, we hypothesize that we can improve the normalization of the risk-taking behaviour and better train on this aspect during mission rehearsal. The visual system of a training device is essential for the immersion level. Consequently, the technology level in the human-machine interface can influence the learning effectiveness. The introduction of a new immersive device brings the need for understanding key factors that impact human-technology interaction such as decision-making, reasoning, memorization, and perception.

One of the research goals is to identify a method that can optimize the cognitive load calculation using non-intrusive biometric sensors. Can machine learning be leveraged for model transferability from engineering where intrusive sensors can be used, to live operation where intrusive sensors cannot be used because of privacy & logistics reasons? Another goal is to explore if flight performance & pilot behaviour are correlated during initial training. In a sequence of initial training manoeuvres, we aimed to answer questions such as: What is the variability of cognitive load between manoeuvres and pilot profiles?

To ensure that the specific areas of focus and the experiment protocol is influenced by previous research studies, a focused literature review is first carried out. We also present the design of the experiment protocol as a training scenario during initial training. With an increasing difficulty scenario, we assessed performance based on criteria and thresholds. This is followed by low flight altitude manoeuvres involving risk-taking decisions as part of a contest between participants. The experiment protocol ends in a 2D/3D AB/Testing during an Air-to-air refuelling manoeuvre to observe if there are significant differences in performance with this new technology addition. In section 4, we provide results with a performance analysis. In conclusion, we identified potential future research using biometric sensors in flight training operations and present the concept of using artificial intelligence to improve the cognitive load computing with the constraint of non-intrusiveness of biometric sensors. More in-depth analysis of the risk-taking behaviour and a comparative evaluation of the perceived experience of the 3D visual are left for another publication.

LITERATURE REVIEW

EEG and eye tracking data are tools used to study and understand the mental state of pilots in aviation as well as the processing of visual information. The applicability to the aviation domain as well as the methodology of data collection and analysis will be the primary focus of this literature review.

Eye Tracking for Cognitive Workload Estimation

Pupil size and eye blinking can be used as an index of cognitive workload where a lower eye blink rate is thought to indicate a higher workload, a higher eye blink rate may indicate fatigue, and larger pupils may also indicate greater cognitive effort or more pleasurable stimuli. Cognitive workload evaluation based on eye tracker data (Marshall, 2002), and a patented method and apparatus (Marshall, 2000) are used in the evaluation of cognitive activity in aviation. (Cabestrero, Crespo, & Quiros, 2009) also analyzed how pupil diameter can be used to reflect mental effort and processing resource allocation when performing a recall task under multiple cognitive load conditions. The pupillary diameter increased systematically until the appearance of the small plateau. No reduction in pupil diameter was observed when exceeding processing resource limits besides the appearance of the plateau during the last tasks. This indicates that participants can continue to actively process even if resources are exceeded.

Scan Pattern With Eye-Tracking Data

Other important eye tracking metrics include blinks, fixation duration and location, and saccades, the rapid eye movements occurring between fixations. A higher number of saccades indicate seeking behaviour. (Škvareková, Iveta 2020) uses an eye tracker to record eye movements. The article confirms that experienced pilots were able to receive information in less time and had higher saccades per minute than inexperienced pilots. (Stephanie Brams et al. 2018) examined differences in gaze and visual scanning behaviour between high-performing and low-performing pilots. They also provide insight into the underlying processes that may explain perceptual-cognitive expertise under the theory of long-term working memory, the information reduction hypothesis, and the holistic model of perception of knowledge. The number of downtime and the number of transitions between AOIs differed between high-performing and low-performing pilots. Poorly performing pilots perform a more exhaustive search and make more transitions between extreme areas of interest. Pilots are better able to shift their attention between global and local information processing.

Understanding pilot's reaction in flight operation using neuroscience

(Lan, Sourina, Wang, Scherer, & Muller-Putz, 2019) recorded EEG data with 15 subjects using 32 EEG channels. In their paper, the authors adopted Differential Entropy (DE) as features for emotion recognition. DE features have been extensively used in cited literature studying the application of transfer learning techniques in EEG-based emotion recognition. Extending our data sample with emotion will be an addition that will complement well in an intelligent adaptive flight training system.

(Binias, Myszor, Palus, & Cyran, 2020) attempted to predict the reaction time to an unexpected event based on the brain activity recorded before the event using EEG data. They measured the time lag in the participant's reaction time to a visual cue using regression in a flight simulation experiment with autopilot enabled. The prediction algorithms used are the least absolute shrinkage operator, Kernel Ridge regression and Radial Basis Support Vector Machine. Automated systems placed the pilot in a passive role which introduced an additional challenge should any issues arise as the pilot must move into an active role and resolve complex issues. (Binias, Myszor, & Cyran, 2018) dealt with the problem of discrimination between brain activity states related to anticipation and reaction to a visual signal and the selection of an appropriate classification algorithm for such problems. In this work, an EEG signal processing and classifier tuning methodology was proposed with the aim of analyzing data containing brain activity states related to an inactive but focused anticipation of a visual signal and a reaction to this signal. Experimental electroencephalographic data were acquired using an Emotiv EPOC+ headset. The methodology involved the use of different classification algorithms, such as Linear Discriminant Analysis, *k*-nearest neighbours, Support Vector Machines, Random Forests, and Artificial Neural Networks. The results suggested that the performance of a neural network could be significantly better than that of other algorithms and validated by an analysis of variance (ANOVA).

In an investigative article by (Monteiro, Skourup, & Zhang, 2019), an in-depth review of techniques for using EEG to assess MF mental fatigue was performed and supported by an overview of the principles of acquisition, interpretation, algorithms, and trends. There are subjective ratings based on self-report to assess MF states and include the NASA Task Load Index, Karolinska Sleepiness Scale, Epworth Sleepiness Scale, and Chalder Fatigue Scale, but they are subject to bias. When evaluating MF sensing, EEG signals are composed of five main frequency bands: delta, theta,

alpha, beta, and gamma. The theta (θ) band can be found during drowsiness and sleep in adults. The alpha (α) band can be found in adults who are awake, relaxed, or mentally inactive. Frontal θ and occipital α and parietal activity are likely to increase as a person becomes fatigued. The beta (β) band signifies tension and anticipation and can be found in alert and anxious subjects. The most used preprocessing methods for MF detection using EEG include digital filtering, independent component analysis (ICA), and discrete wavelet transform (DWT). Commonly used feature extraction methods include power spectral density (PSD), statistics, and entropy measurements. When a person is fatigued, a decrease in the level of entropy of their EEG signals can be expected, indicating a decrease and weakening of brain synapses. The most used measures of entropy are Sample Entropy (SampEn), Fuzzy Entropy (FuzEn), Approximate Entropy (AppEn), and Spectral Entropy (SpecEn). Since the MF state is a constructed process where fatigue accumulates over time, a dynamic approach considering the temporal aspect becomes possible with the development of models such as LSTM. The article concluded by suggesting a model based on kernel partial least squares discrete output linear regression as a good overall option for an FM evaluation system.

Neuroscience for Pilot Workload

With ECG, Eye Tracker, and EEG complemented by NASA-TLX questionnaires, (Thomas C. Hankins, 1998) measured the mental workload of pilots collected during flight scenarios. Combining multiple measures such as psychophysiological states and subjective measures can provide a broader picture of the mental state of the pilot. Heart rate is useful to measure the flight demand but not on the mental workload. Eye tracker was more powerful for the diagnostic task while EEG theta band increased during mental calculation.

Augmented cognition is a form of human-computer interaction in which sensing a user's cognitive state is used to invoke system automation on demand. The study by (Nicholas Wilson, 2021) monitored the pilot's in-flight physiological state to determine the optimal combination of EEG cues to predict changes in cognitive workload. Data collection was executed in a real-world flight environment with scenarios that varied in workload with a group of undergraduate aviation students using a single-engine trainer equipped with Garmin G1000 avionics. Some of the higher workload flight manoeuvres were executing a missed approach to minima and performing consecutive steep turns. Conversely, manoeuvres categorized as low workload included straight and level flight and taxiing in an airport. Power spectral density values were calculated and subjected to machine learning methods to distinguish periods of high and low workload. The feature extraction step was performed using power spectral analysis. Fast Fourier transform (FFT) was used to transform EEG into power spectral density (PSD). The Lasso cross-validation algorithm was used to select the most important features. The support vector machine (SVM) algorithm was used as a binary classifier for its robust approach to complex pattern recognition, good generalization performance, and efficient computational cost.

Neuroscience in flight training

(Zhang, Chen, & Wu, 2020) compared the pilot's EEG signal at different phases of flight, different weather conditions, and different levels of training. The results showed that EEG entropy could be used to assess the pilot training effect. The entropy value in windy and rainy conditions was more dispersed, which means that the frontal workload is greater than in sunny conditions. According to the cerebral plane, the load on the occipital lobe, part of the parietal lobe and the right temporal lobe increased. The increase in the occipital load during the take-off phase comes from the change in the exterior view of the cabin. This change will bring higher mental load to the pilots and the difficulty in processing information has caused the load on the frontal lobe as the whole processing centre of the brain to fluctuate considerably. The student will adapt to this environmental change in the later stages of flight, reducing the load on the temporal lobe. Trained pilots demonstrate that regular training increases excitation of the frontal and occipital lobes and as training time increases; the average level of entropy approach a fixed value.

The ability to identify the learner's workload is crucial for their implementation of an adaptive training system. The study by (Baldwin & Penaranda, 2012) used an artificial neural network (ANN)-based classification algorithm using neurophysiological measures requiring effective real-time mental workload classification based on electroencephalographic (EEG) activity during the performance of short-period tasks. Classifiers determined the workload and the cognitive-emotional response of a learner during training were essential for the implementation of adaptive training. With fifteen participants, signals from the EEG and EOG electrooculogram were recorded using a 40-channel NeuroScan NuAmps amplifier, and a 40-channel QuikCap Ag/AgCl electrode cap. Three different working memory tasks, the Reading Span task, the Visuospatial n-back (n-back) task, and the Sternberg Memory Scanning task were used. The ANN could distinguish between low and high difficulty levels quite reliably.

In their article, (Liu et al., 2019) proposed the use of emotion, workload, and stress recognition algorithms based on the use of an EEG, in addition to questionnaires and traditional feedback to study the optimal duration of the training of air traffic control officers (ATCOs). A 14-channel Emotiv EEG device was used to monitor the brain states of ATCOs as they learn to use a new 3D interface in performing aircraft trajectory operations in different weather and terrain conditions in addition to the traditional 2D display. Emotion and workload recognition algorithms from EEG signals and a stress recognition algorithm are proposed. It was observed that while emotions did not have a definite impact on training duration, workload and stress levels were significantly different between training duration and optimal training duration. Correlation analysis indicates that if ATOs have more confidence in a new system, their emotion is more positive, and stress and workload are lower when they learn to use this new interface.

METHODOLOGY

Problem Statement

Observing students and operating the simulator can be complex for the instructor; they may miss some behaviour in fast pilot operations during the evaluation of complex manoeuvres. Moreover, trust in the instructors' evaluation could be challenged. The use of an expert is not scalable or cost-effective when assessing non-technical skills. Soft skills are not as easily assessed by another human comparatively to technical skills. Without the support of data analytics and visualization tools, it will be impossible to identify levels of correlation across the entire rich collection of data and parameters available. The understanding of the cognitive load at the microlevel task requires high temporal resolution data that electroencephalogram can provide. However, the intrusiveness of the EEG in flight operation is preventing the collection of this data.

Experimentation Device

In our research, we used a fastjet flight simulation device with an immersive visual and flight controls. The experimentation occurred in an F16 - Flight Simulator with CAE Medallion MR e-Series visual system that provided natural hi-fidelity visual immersion with the objective of reducing eyestrain and fatigue. The prototype simulation consisted of a 200° partial sphere screen with a radius of 1m and a height of 1.5m giving a 9,42m viewable surface area as shown at Figure 1. The aim of such prototype was to provide Smearing/Motion-Blur reduction from unequalled dynamic 120Hz resolution and active eyewear was used for head movement compensation to virtually eliminate parallax error and 3D depth perception (**Erreur ! Source du renvoi introuvable.**). The flight simulator will not have motion enable during the experimentation.

Figure 1 - Flight Simulation Visual System, Head tracker and 3D glasses

The flight simulator did not use a motion system during the experimentation and the key recorded flight parameters were: Latitude (LAT), Longitude (LONG), Mean Sea Level (MSL), Altitude Above Ground Level (AGL), Calibrated Airspeed (CAS), Ground Speed (GSPD), G-Force (GTRK), Heading/Yaw (HDG), Pitch (PITCH), Roll/Banking (ROLL), Angle of Attack (AOA), and Engine Thrust (ENG_THRUST).

Sensor's selection

The first step of the research was to select the biometric sensors to be used during the execution of the experiment. This selection must be able to consider the specific nature of a fastjet flight simulator offering the pilot a 200-degree field of view while maintaining data quality for the purpose of calculating cognitive and emotional state, as well as the path of gaze on the instruments or Area of Interest (AOI)

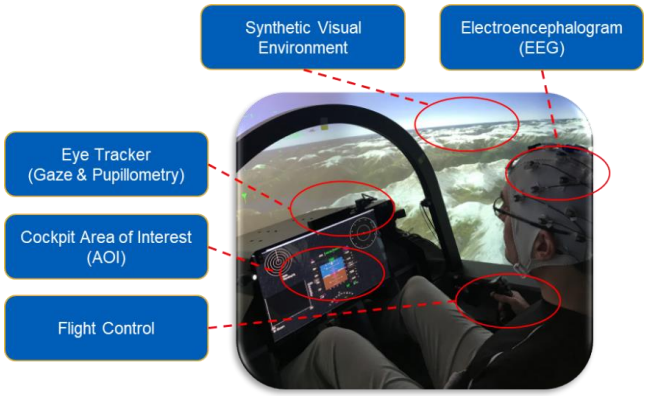


Figure 2 - Flight Simulation Cockpit and Biometric Sensors

The EEG device is a tool for measuring brain electrical activity and infer mental states such as mood, anxiety, or stress as well as cognitive state. We recorded EEG with 32 electrodes at 1000 Hz with BrainAmp amplifiers from Brain Product. The EEG signal was recorded using the standard 10–20 montage. The signal was first referenced into Fz and then filtered using a 2nd order Butterworth 1-40 Hz bandpass IIR filter and a 60 Hz notch filter. Muscle and eye movement related artifacts were removed using blind source separation by independent component analysis. Before analysis, EEG signal was downsampled to 500 Hz. The Eye Tracker system is using 5-cameras and a Bar Tracker model 5-CAM DX 2.0 MP Smart Eye Pro with IR mini flashes 60Hz. Data Collected are Fixation/Dwell time, Blinks, Pupil diameter, Saccades and Intersection names with a 3D world model composed of Area of Interest (AOI) of the instrument panel.

Figure 3 Smart Eye Pro dx camera + IR and Smart Eye Aurora system (Bar tracker)

The bar tracker has 2 built-in cameras and 2 infrared illuminators. The system can obtain more precise metrics for instrument checks with the 5-camera system and was suitable for the context of a wide range FOV. The definition of the world is composed of Area of Interests (AOIs) corresponding to the specific cockpit instrument area, which is done using a laser and a laser chessboard to calculate the world coordinates using various geometric shapes to design the 3D world. The device is able to obtain the metric to about 150 cm around the subject. However, if the subject is seeing a point beyond, the system does not detect properly the world intersection, or it might have interference. The gaze calibration is performed with every subject using the device. If a gaze calibration is not performed, the results may have a variation between 3 and 7 degrees with respect to the point to which the subject is looking at.

Experimentation Protocol

The experimental protocol was designed in a mirror fashion composed of two artificial and controlled tasks and then, valid ecological flying tasks. The first experimental task was a synthetic n-back task (Susanne M Jaeggi, 2010), which is known to gradually manipulate mental workload. The corresponding simulator task was a sequence of maneuvers designed to incrementally increase the mental workload of the pilot through maneuver difficulty. The second experimental task was a BART, created to manipulate risk-taking behaviors. The corresponding task was a risk-taking free-flying task during which the participant had to fly near mountains and valleys. The artificial tasks were implemented using the software package E-Prime 3.0 from Psychology Software Tools and performed before the two ecologically valid flying task. Sixteen participants recruited from a Flight Training Company with a beginner profile or first step in piloting were favoured in order to promote the collection of data in cognitive overload. All human subjects have signed a consent form to participate in this experiment. As shown in Figure 4, the simulation tasks are divided in three categories, the initial training manoeuvres, the low flight altitude task and the Air-to-Air refueling with a MRTT Tanker.

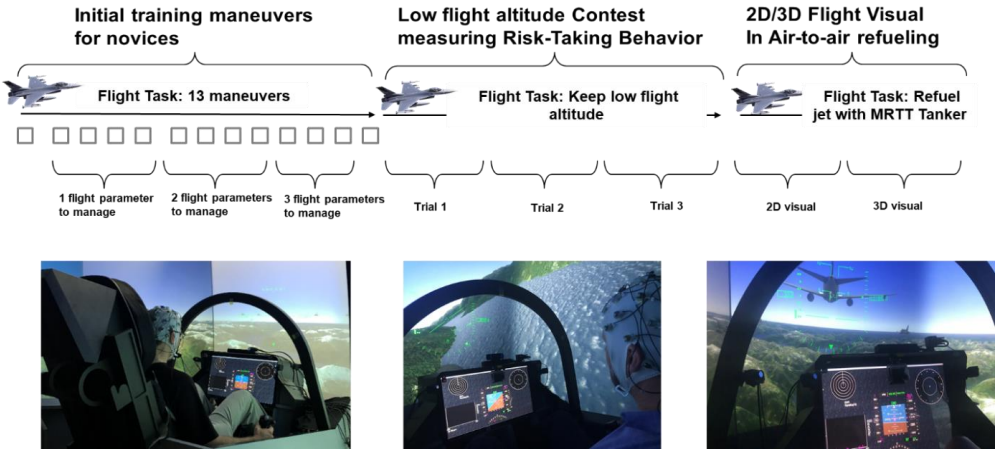


Figure 4 - Flight Simulation Tasks Sequence

The human testing activities will consist of simulated real-world scenarios in various controlled setting environments.

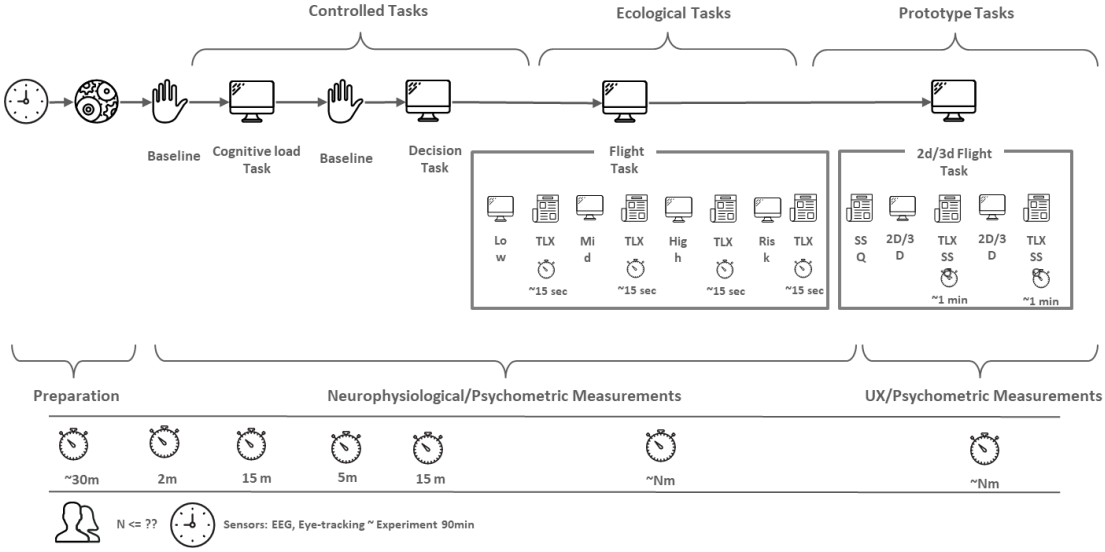


Figure 5 - Experimentation Protocol Diagram

Initial manoeuvre with increasing difficulty

The first phase of the experiment consists of flying “twisler”. It consists of a few little simple manoeuvres, which requires looking at various flight parameters such as banking, pitch, altitude, and speed. This phase is inline with the technical test done on the A310 simulator where a pilot was coached by an instructor to execute a few simple manoeuvres in an initial training context. Manoeuvres corresponded to a variation of 4 parameters: speed, altitude, heading, and banking. After a free flight to familiarize with flight control and the aircraft reaction, we started the exercise with an aircraft stabilization manoeuvre. The following 4 manoeuvres consisted of changing one of the flight parameters in isolation, one at a time. The following 4 manoeuvres consisted of the same principle but varying two parameters simultaneously. The following two manoeuvres varied 3 and then 4 parameters simultaneously to complete this exercise phase with a vertical loop before stabilizing the aircraft. There are three blocks of flight manoeuvres shown as low, moderate, and high mental load. Each manoeuvre was associated with appropriate flight actions and led by a flight instructor, who recorded each participant's performance and signalled the end of each manoeuvre as a synchronization marker between the data. The assessment was made by an instructor that evaluated the participants after the execution of each maneuver. The factor for assessment was time, fluid stabilization, slope of the change, and compliancy to the threshold. We used a tolerance for the various flight parameters as followed: Altitude: 100 ft, Knots: 5 ft, Deg. Heading: 5 deg, Deg. and Banking: 5 deg

Low flight altitude for Risk-Taking Behaviour

To assess risky behaviours, we performed a computerized decision-making and risk-taking activity followed by a simulated flight activity involving a risk-taking task. The first task, two experimental computerized risk-reward risk-taking tasks were performed, an Iowa game task (Antoine Bechara, 1994) and a Balloon Analog Risk Task (BART) (Lejuez, 2002). The second phase involved risk taking. Where a participant was asked to sustain an altitude above ground at various thresholds, where each threshold paid points to a leaderboard. The objective was to incite pilots to manage risk and their strategy in order to gain points. In a simulation through the British Columbia Rockies Mountains, we asked the participate to choose a mountain or a valley that may create a multiplier factor of: Valley = x1, Small Mountain = 2x, Big Mountain= x3 with the above point distribution per altitude level 0-250 ft = 100 pts, 251-500 ft = 50 pts, 501 – 1000ft = 25 pts, Crash or >1000 ft = 0 pts. Participants got 3 trials to accumulate points and we gave a TLX questionnaire after the third trial.

2D/3D AB Testing during Air-to-Air Refuelling Manoeuvre

In the third phase, half of the participants started with 2D visual display then switched to 3D, another half started with 3D then finished with 2D. We were inspired by (Wen-Chin Li, 2014) where the scenario was an air-to-air task in a jet fighter simulator studying eye movement. We asked participants to execute a Mission rehearsal of an air-to-air refuelling. We assessed the performance based on time to reach the in-flight refueling pole (boom) of the MRTT Tanker and the stability of the flight while keeping the boom in range. We gave a TLX questionnaire after each trial.

Data Analysis Methodology

In this section, we will present the methodology of the analysis of the performance data for this experiment. Our goal was to develop an autonomous method to score or grade each pilot's performance through various flight manoeuvres based on their telemetry data. Our prediction would be as the manoeuvre's difficulty increases; the performance of the pilots would decrease. In relationship to the telemetry data, we also want to analyze and correlate the results of the eye-tracking data using an ANOVA test. It is important to note that all participants were anonymous for the experiment and analysis to be objective. Successful training of supervised machine-learning approaches for classification required objective ground truth to provide annotated examples of the classification target. In the case of the research, we used two deep learning models, a fully convolutional neural network (FCN and a residual network (ResNet) to estimate mental workload.

Data Description

There were three data sources taken from this experiment, Smarteye's eye Tracker, Brain Vision's EEG, and Objective Assessment. Smarteye log files are generated from the videos captured by the Smarteye camera system. There is one log file per participant. The log files contain tab-separated entries. The number of rows in each log file is the frame numbers captured by the Smarteye camera. Each log file has 485 columns. The log files can be loaded as a data frame in pandas (a 2D data structure), which could be very useful for various analytics tasks. Each time stamp is 16.7 ms with a frequency of 60 HZ. Objective Assessment was a measure of the telemetry data from the flight manoeuvres performed. The measurement of the objective assessment was on different parameters that include Altitude, Banking, Heading, Speed, and Pitch. The objective assessment of the experiment was based on the exceedance occurrences, standard deviation, and time. Each time stamp is 240ms. We used two deep learning models, a fully convolutional neural network (FCN and a residual network (ResNet) to estimate mental workload. EEG data was exported as CSV files that include FCN & RESNET methods.

Performance assessment method

Objective assessment was based on the telemetry data of each pilot and the following factors: Exceedance occurrences, Standard Deviation and Time of flight. The objective score was scored from 0-4 and was averaged based on the 3 factors above. The Standard Deviation factor is scored based on the standard deviation of each parameter: Altitude, Banking, Heading, Speed and Pitch. Each manoeuvre differed from one another, since some of them had one, two, or all parameters. The pilot's standard deviation for a given manoeuvre was in comparison to the preset tolerance each manoeuvre had for each parameter (Altitude, Banking, Heading, Speed & Pitch). Time was to be scored based on the difference the pilot took on each manoeuvre in comparisons to the time of tolerance for the specific manoeuvre. The time tolerance was preset based on the mean time it took for all participants to finish the given manoeuvre. Time score was also scored from 0-4. The Exceedance occurrence factor is based on how consistent the pilot would stay within the tolerance of each manoeuvre and its parameters. The factor was measured on how much time of the entire manoeuvre did the pilot spend outside of the tolerance. It was also scored from 0-4.

Hypothesis test method

One-way ANOVA test was performed to compare the means of multiple grouped sets of data.

Null hypothesis: Groups means are equal (no variation in means of groups)

H_0: $\mu_1=\mu_2=\ldots=\mu_p$

p-value< 0.05

Alternative hypothesis: At least, one group mean is different from other groups

H_1: All μ are not equal

p-value> 0.05

After finding the p-value, a post hoc comparison was made using a Tukey honestly significantly differenced (HSD) test to know which group was significantly different from each other. The ANOVA test assumption was primarily checked by using the Shapiro-Wilk test that analyzed the normal distribution of the residuals. Depending on whether the results were drawn from normal distribution or not, a Bartlett's test was used to check the homogeneity of the variances to see that it was normally distributed and Levene's test when not normally distributed.

Null hypothesis: Group variances are equal (no difference in variance of groups)

H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$

p-value< 0.05

Alternative hypothesis: At least, one group variance is different from other groups

H_1: All σ are not equal

p-value> 0.05

AOI and Workload Correlation method Using the Gaze Tracking method

A one-way ANOVA test was performed to compare the means of 3 AOI grouped sets (Not looking, slightly looking, Looking a good amount) of data in comparison to each flight parameter (Altitude, Banking, Heading, Speed & Pitch). ANOVA test was done 5 times in total, split up by manoeuvres that flight parameter mattered. A one-way ANOVA test was to be performed twice for the workload (fixation & saccade). The means of 3 different groups for the objective and subjective scores were compared to the fixation and saccade. The objective and subjective scores were split into: Bad (Score under 2), Average (Score between 2-3), Good (Score between 3-4)

RESULTS

Flight Profiles

With the 5 parameters studied as time series per task (Altitude, Speed, Heading, Pitch, Banking). Figure 6 is an example of a flight profile for the altitude change task. We can see different profiles with good/bad stabilization, some short/long manoeuvre time, and multiple peaks prior to the targeted altitude prescribed by the training scenario.

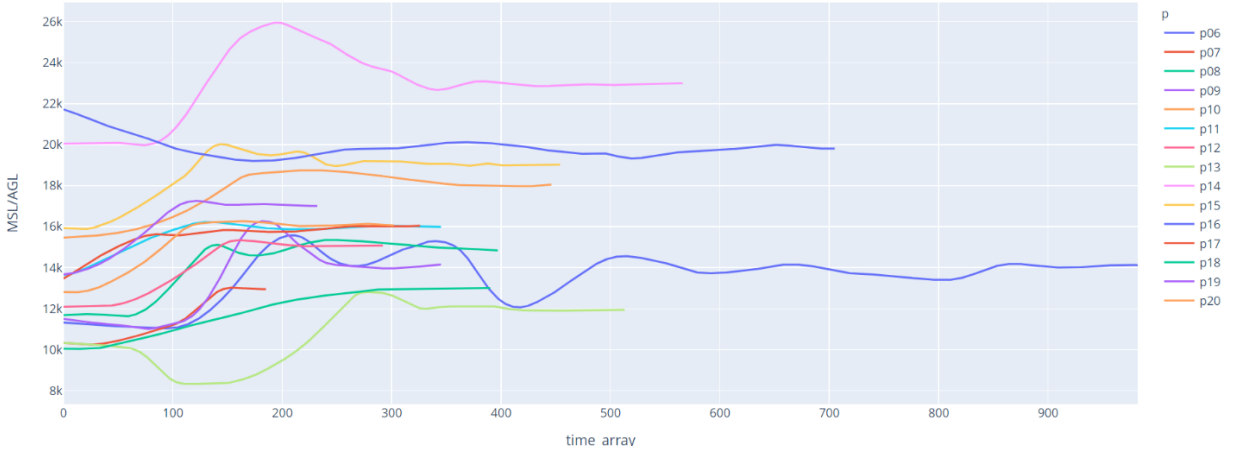


Figure 6 - Flight profile - Altitude change task

The Table 1 presents the result of tested accuracy and device specifications using a variety of candidates prior to the experimentation.

Table 1 - Eye Tracker Accuracy Assessment and Specifications of Smart Eye's eye tracker

	Gaze Tracking (Able to track gaze) (1)		Gaze tracking - Calibration results (Deviation/Accuracy)		Head tracking, (Able to track head) (3)(4)(5)		Head Tracking, Field of vision (Range) (2)	
	5-Cam	Bar Tracker	5-Cam	Bar Tracker	5-Cam	Bar Tracker	5-Cam	Bar Tracker
No glasses	99%	90%	1.5° deviation, 0.8° accuracy	0.2° deviation, 0.6° accuracy	> 97%	> 97%	> 180°	90° – 130°
Wearing glasses	80%	70%	3.5° deviation, 2.5° accuracy	2.7° deviation, 3.2° accuracy	> 97%	> 97%	> 180°	90° – 110°
Long hair	99%	90%	1.3° deviation, 0.9° accuracy	0.7° deviation, 1.3° accuracy	> 90%	> 90%	> 180°	90° – 130°
Helmet	99%	90%	1.1° deviation, 0.2° accuracy	0.6° deviation, 0.2° accuracy	> 90%	> 90%	90° – 150°	90° – 110°
Communication device	90%	90%	1.9° deviation, 2.7° accuracy	0.5° deviation, 0.8° accuracy	> 90%	> 90%	90° – 150°	90° – 110°
Covering ears	90%	90%	1.5° deviation, 2.4° accuracy	0.3° deviation, 0.8° accuracy	> 90%	> 90%	90° – 110°	90° – 110°
Facial hair	90%	90%	1.8° deviation, 2.6° accuracy	0.4° deviation, 0.7° accuracy	> 97%	> 97%	90° – 150°	90° – 130°
Make-up	99%	90%	N/A	N/A	> 97%	> 97%	> 180°	90° – 130°
Covering mouth	90%	90%	N/A	N/A	> 90%	> 90%	> 180°	90° – 130°

(1) The use of multiple cameras compensates both eyes and a unified gaze direction is streamed, (2) The metric of the 5-Cam according to the position of the cameras, (3) Affected by the field of vision and the angle between the subject and the camera, (4) Affected by covering facial features, (5) This result can improve creating a manual profile for the subject

Performance Assessment Analysis

The Figure 7 represents the average of objective and subjective scores for the entire experiment for each participant. There is a discrepancy between the objective and subjective scoring. The subjective scoring can be influenced by the emotional aspect of the person scoring the participants. However, with the objective scoring, everything is rational, and feelings are not a factor when giving a score. Figure 7 also presents each manoeuvre graphed with the comparison of the objective and subjective scores. As the manoeuvres increase in difficulty, the objective scores start to decrease which goes along with the initial hypothesis.

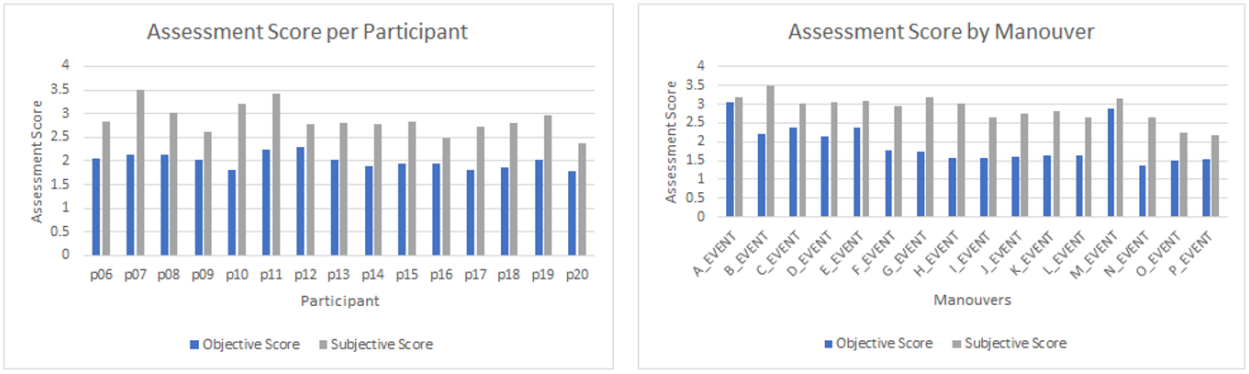
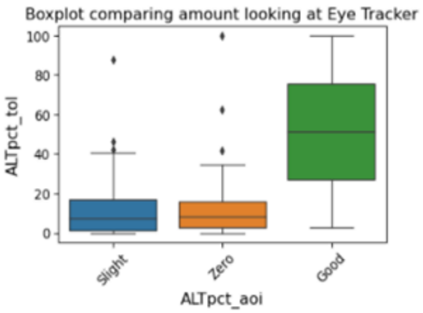


Figure 7 - Grades per participants and per manoeuvres

AOI Correlation results

An ANOVA was done to analyze the correlation between the eye tracking of the participants and their performance with the objective score. From the results below, it was concluded that the hypothesis of manoeuvres that altitude, pitch, and speed factored in, the more time the participants/pilot looked at the gauges for those parameters, the better their objective score would be. The results, however, differed for the speed and heading parameter where a p-value above 0.05 was seen.

Test	P-Value	Result
Altitude	0.003	Reject, means are different
Pitch	0.046	Reject, means are different
Banking	0.049	Reject, means are different
Speed	0.56	Accept, means are same
Heading	0.22	Accept, means are same



Workload Analysis

Two end-to-end deep learning models were trained to learn features and mental workload levels based on the EEG signal recorded during the n-back tasks. Two models, an FCN (91% accuracy) and a ResNet (92% accuracy) algorithms trained on the task data provided us a good mental workload level estimation. The models were trained to discriminate mental workload level from the cleaned signal data. Sanitary checks of the models were performed for their physiological plausibility. The results matched expectations: an increase of manoeuvre complexity led to an increase of high mental workload classifications, and a decrease of maneuver performance. There is also a close link between the pilot's mental workload and eye movements, if a pilot does not have optimum conditions, eye movement changes. We define the two parameters as saccades and fixation. The saccade is defined as the rapid movement of the eyes between two fixations. Fixation is defined as a condition in which an individual visually collects and interprets information available in the range of the eye over a period of time.

In an article from the 9th International Conference on Air Transport "Number of Saccade & Fixation Durations as Indicators of Pilot Workload". They measured the effects of mental stress caused by a lack of training and flight experience. They've concluded that with a higher saccade per minute, experienced pilots were able to receive information in less time. A shorter stay on flight instruments allowed more experienced pilots to scan other areas of interest. They also had more time to detect any errors and then start the correction. To compare our results with the research paper, we've used an ANOVA test to validate the performance assessment in relationship to the Saccade & Fixation of each pilot. The hypothesis was done four times, twice each for saccade and fixation using both objective and subjective scores. The scores were divided into 3 categories of bad, average, and good.

	Bad	Average	Good
count	179.000000	38.000000	23.000000
mean	17.110838	30.928947	141.886087
std	19.348152	30.881550	434.929001

Test	Objective Score P-Value	Subjective Score P-Value	Result
Saccade	0.0002	0.48	Reject, means are different. Major Correlation between "Good" and other scores.
Fixation	4.9E-9	0.1	

The results of the ANOVA test further validates the research previously done. Both the Saccade and Fixation p-values were under 0.05 by a significant amount showing that the means of the different groups of the objective score have a major correlation between each other. With a higher saccade per time frame, the pilots achieved a higher objective score. However, the same cannot be said using the subjective score, further proving the objective scoring scheme is more accurate and should be utilized more than the subjective score. For further validation, we have also graphed the ratio of fixation over saccades in comparison to the assessment score per manoeuvre to give a more robust level of attention to the pilots. Pilots who achieve a higher score tend to have a lower ratio of fixation per saccade. Incorrect scanning patterns from the pilots could lead to information overload and staying longer on a flight instrument. The difference between these numbers of fixation per saccade for pilots is mainly because successful pilots were able to receive information in a shorter time and continued with the scanning technique of instruments.

CONCLUSION

The objective of this study was to explore artificial intelligence capability & human factors during initial flight training session. Those insights contribute to explore how we can enhance the instructor's awareness of the cognitive workload, and scan student's pattern. As an eventual capacity of an intelligent adaptive flight training system, the session can be tailored to maximize the students in real time. We also aimed to develop a conceptual method by using artificial intelligence to keep the strength of an EEG device in the engineering phase and prevent the addition of intrusive sensors during real flight training operations. For this we used biometric sensors in a simulator training session and explored how we could use data science on a number of training manoeuvres that could be used to assess pilot performance during initial training tasks, risk-taking behaviour tasks and air-to-air refueling tasks within a new human interfaces machine technology to improve training programs and immersive systems. As expected, we found that as manoeuvre complexity increased, workload and perceived mental workload measured by NASA-TLX also increases. We also found that correlation exists between the scanning pattern and the flight performance. This indicates to us that biometry sensors can bring a new kind of insight towards objective measurement in the assessment of the human performance in the flight operation.

For future work, we consider going further in the cognitive workload estimation methodology. We would like to identify and compare multiple algorithms that are able to classify and predict cognitive load, flight performance, and risk-taking behaviour. We would like to address questions such as: How can we predict the outcome (technical/non-technical) of a manoeuvre on biometric & flight telemetry data? Does cognitive load risk-taking behavior is affected by the cognitive load? What is the correlation between flight performance with telemetry and psychophysiological state? How can we detect and predict flight performance based on the cognitive load index? The dataset we accumulate can be used in future analysis around risk-taking behaviour at low altitude manoeuvre. The results of formal data analysis using anticipated statistical methods would provide insight into participants' risk behaviours and the level of cognitive workload required when taking risks. This study will aim to analyze the human factors associated with risky behaviours using the characteristics of central and autonomic nervous system activity and answer questions such as: What is the neurophysiological state involves risk-taking behaviour. What is the performance impact for the various risk behaviors? We will also propose to use machine-learning methods to build a risk classification model using other flight simulation models to valid portability of machine learning algorithms for pilot performance & behaviour. Analysis using biometric sensors to assess initial training and risk-taking behaviour of novice pilots. Which machine learning algorithms are more suitable to apply risk-taking behaviour classification and prediction?

To contribute to flight safety, those capabilities can be matured up with an R&D project into a business-centric initiative. Using a real Crew Resource Management (CRM) training session on a large number of experienced pilots, we will explore how we can augment the technical readiness level of neuroscience capability. We will consider operating emergency manoeuvres in a flight-training session commercial aircraft simulator in a flight-training centre using non-intrusive biometric sensors and certified flight instructor. We will show a method to evaluate perceived experience using self-reported data with the 2D vs. 3D visual system as an A/B testing results. We will also test the usage of artificial intelligence to provide an optimization of the cognitive load index measured by eye tracker and pupillometry. By using supervised machine learning with pupillometry as a feature and EEG cognitive load index as the target label, we can provide a machine-learning model that is deployable without the intrusion of an EEG in flight operation, as presented by the Figure 8

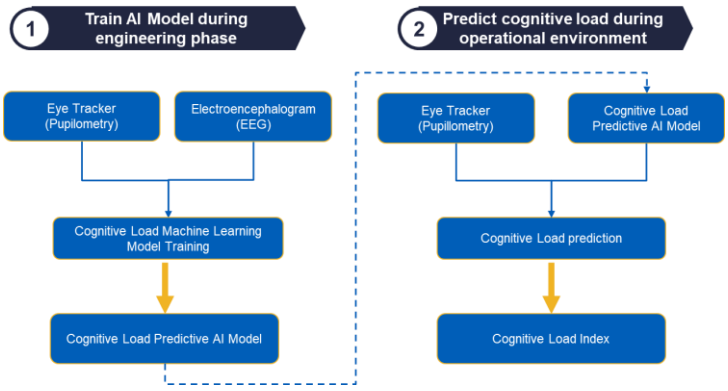


Figure 8 - Cognitive Load Optimization with EEG label in machine learning process

ACKNOWLEDGEMENTS

Acknowledgements to Andrea Lodi, Research Director from Polytechnique de Montréal, Patricia Gilbert, Marc St-Hilaire, Andrew Fernie, Ginet Andreina Calderon Perez Philippe Perey from CAE Inc. and Alexander Karan, Nicolay Nonchev from HEC-Tech3lab

REFERENCES

Antoine Bechara, A. R. D., Hanna Damasio, Steven W.Anderson. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition,, 50*(1-3).

Baldwin, C. L., & Penaranda, B. N. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *NeuroImage, 59*(1), 48-56. doi:10.1016/j.neuroimage.2011.07.047

Binias, B., Myszor, D., & Cyran, K. A. (2018). A Machine Learning Approach to the Detection of Pilot's Reaction to Unexpected Events Based on EEG Signals. *Comput Intell Neurosci, 2018*, 2703513. doi:10.1155/2018/2703513

Binias, B., Myszor, D., Palus, H., & Cyran, K. A. (2020). Prediction of Pilot's Reaction Time Based on EEG Signals. *Front Neuroinform, 14*, 6. doi:10.3389/fninf.2020.00006

Cabestrero, R., Crespo, A., & Quiros, P. (2009). Pupillary dilation as an index of task demands. *Percept Mot Skills, 109*(3), 664-678. doi:10.2466/pms.109.3.664-678

Lan, Z., Sourina, O., Wang, L., Scherer, R., & Muller-Putz, G. R. (2019). Domain Adaptation Techniques for EEG-Based Emotion Recognition: A Comparative Study on Two Public Datasets. *IEEE Transactions on Cognitive and Developmental Systems, 11*(1), 85-94. doi:10.1109/tcds.2018.2826840

Lejuez, C. W., Read, J. P., Kahler, C. W., Richards, J. B., Ramsey, S. E., Stuart, G. L., ... & Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: The Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied, 8*(2).

Liu, Y., Lan, Z., Traspsilawati, F., Sourina, O., Chen, C.-H., & Muller-Wittig, W. (2019). *EEG-Based Human Factors Evaluation of Air Traffic Control Operators (ATCOs) for Optimal Training*. Paper presented at the 2019 International Conference on Cyberworlds (CW).

Marshall, S. P. (2000). 6090051. U. S. Patent.

Marshall, S. P. (2002). *The Index of Cognitive Activity Measuring Cognitive Workload*. Paper presented at the IEEE 7' Human Factors Meeting, Scottsdale Arizona.

Monteiro, T. G., Skourup, C., & Zhang, H. (2019). Using EEG for Mental Fatigue Assessment: A Comprehensive Look Into the Current State of the Art. *IEEE Transactions on Human-Machine Systems, 49*(6), 599-610. doi:10.1109/thms.2019.2938156

Nicholas Wilson, H. T. G., Jessica VanBree, Bradley Hoffman, Kouhyar Tavakolian. (2021). *Identifying Opportunities for Augmented Cognition During Live Flight Scenario An Analysis of Pilot Mental Workload Using EEG*. Paper presented at the International Symposium on Aviation Psychology.

Susanne M Jaeggi, M. B., Walter J Perrig, Beat Meier. (2010). The concurrent validity of the N-back task as a working memory measure. *PubMed, 18*(4).

Thomas C. Hankins, G. F. W. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation Space and Environmental Medicine*.

Wen-Chin Li, C.-s. Y., Lon-Wen Li, Matthew Greaves. (2014). *Pilots Eye Movement Patterns during Performing Air-to-Air Mission*. Paper presented at the Proceedings of 31st European Association for Aviation Psychology Conference.

Zhang, F., Chen, D., & Wu, D. (2020). *Analysis of Pilot's Training Effect Based on EEG Signal*. Paper presented at the 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT.