Multimodal assessment of pilots' affective states using psychophysiological sensor signals and facial recognition analysis

Agata Lawrynczyk,
CAE Canada; Carleton University
Ottawa, Ontario
Agata.Lawrynczyk@cae.com

Maher Chaouachi
CAE Canada, McGill University
Montreal, Quebec
Maher.Chaouachi@cae.com

Susanne P. Lajoie
McGill University
Montreal, Quebec
Susanne.Lajoie@mcgill.ca

ABSTRACT

Human error in aviation can lead to catastrophic results. Since 1959, the majority of fatal airplane accidents worldwide occurred during the takeoff and landing phases of flight and upwards of 60% are attributed to pilot error. Psychologists Robert Yerkes and John Dodson proposed an empirical relationship (Yerkes-Dodson law) that performance increases with physiological or mental arousal, but only up to a point, after which it decreases. Understanding a pilot's arousal during flight and its relationship with performance can ultimately contribute to the design of improved flight safety and flight training systems.

The objective of this research is to provide guidelines for the multimodal assessment of pilots' arousal level and affective states using noninvasive biosensors to answer the following three questions: i) Which data processing approach must be adopted to track pilots' arousal; ii) What affective/cognitive model can be used to interpret these measurements; and iii) How can these measurements be used as training assessment criteria?

This paper reports on an experimental study we conducted in a full-flight simulator to explore the answers to the above questions. Pilots' heart rate and galvanic skin response (GSR) were non-invasively and continuously recorded using a wristwatch biosensors during a 45 min flight of varying levels of complexity. Their facial expressions were recorded using a cockpit webcam to enable facial recognition analysis. Results revealed that physiological patterns may be related to the complexity level of the flight phase and to the pilots' performance and experience. Specifically, the mean amplitude sum of the GSR phasic component revealed the pilot arousal level while the valence of the pilot's affective state was captured through facial emotion recognition. These results can be mapped to the Circumplex Model of Affect as a framework to assess both individual and group performance.

ABOUT THE AUTHORS

Agata Lawrynczyk is a Human Factors Consultant at CAE Canada, where she conducts operational research in various domains including aviation, medical, and navy. She is pursuing her Master's degree in Human Computer Interaction at Carleton University where she is interested in pilot biometric responses in various flight training environments. She earned a Bachelor's of Applied Science in Mechanical Biomedical Engineering at the University of Ottawa as well as a Bachelor's of Arts in Kinesiology at Western University.

Dr. Maher Chaouachi is a postdoctoral research fellow at McGill University. He obtained his Ph.D. in computer science at the University of Montreal in Artificial Intelligence (AI). His research examines the application of affective computing principles to enhance learning outcomes and experience. Dr. Chaouachi research interests include using AI techniques to infer, visualize and understand humans' affective and cognitive states from multimodal data. Since September 2015, he has been working with CAE in projects involving data mining and modeling to improve pilots' training.

Dr. Susanne Lajoie is a Canadian Research Chair in Advanced Technologies for Learning in Authentic Settings at McGill University. She explores how theories of learning and affect can be used to guide the design of advanced technology rich learning environments in different domains, i.e. medicine, mathematics, etc. These environments serve as research platforms to study student engagement and problem solving in authentic settings. She uses a cognitive approach to identify learning trajectories that help novice learners become more skilled in specific areas and designs computer tools to enhance self-regulation, memory, and domain-specific learning.

Multimodal assessment of pilots' affective states using psychophysiological sensor signals and facial recognition analysis

Agata Lawrynczyk,
CAE Canada; Carleton University
Ottawa, Ontario
Agata.Lawrynczyk@cae.com

Maher Chaouachi
CAE Canada, McGill University
Montreal, Quebec
Maher.Chaouachi@mcgill.com

Susanne P. Lajoie
McGill University
Montreal, Quebec
Susanne.lajoie@mcgill.ca

INTRODUCTION

Aviation, as an industry, requires high reliability since human error can lead to catastrophic results. Between 1959 and 2015, the majority of fatal airplane accidents worldwide occurred during critical phases of flight: takeoff/initial climb (12%), and final approach/landing (49%) (Boeing, 2015). Upwards of 60% of preventable incidents are attributed to the pilot. Pilot error is linked to fluctuations in mental and/or physical demands that impact cognitive resources. Understanding a pilot's arousal during flight and its relationship to workload and performance can ultimately contribute to the design of improved flight safety and flight training systems.

Cognitive Load in Pilots

Cognitive load refers to the fact that there are inherent limitations of working memory while problem solving (Sweller, 1988) and that these limitations must be considered with respect to the task demands (O'Donnel & Eggemeier, 1986). Cognitive load is due to an interaction that emerges from the circumstances under which the task is being performed; the skills, behaviours, and perceptions of the pilot; and the requirements of the task (Hart & Staveland, 1988) as opposed to a property inherent in a pilot's brain. As the demands increase, so does the cognitive load of the pilot. While it may be obvious that a pilot is likely to experience high cognitive load in complex situations, it should not be overlooked that monotonous situations (e.g., long, trans-Atlantic flights in Autopilot) can lead to a vigilance decrement, as performing the task requires a high effort to remain awake and alert (Paxion, Galy, & Berthelon, 2014).

Cognitive load is more than just performance on a task, and therefore cannot be determined solely by monitoring a pilot's performance metrics (e.g., flight path control and navigation accuracy). Performance can remain constant even though pilots are experiencing increased workload conditions, until a point is reached where the performance level drops drastically (Skinner & Simpson, 2002). Task shedding of lower priority tasks may suggest that the pilot is approaching an overloaded condition, but often times this is only clear once failure occurs. This suggests that monitoring a pilot's psychophysiological arousal has critical safety implications.

Today, a pilot's role is increasingly that of monitoring, information management, and decision making (Skinner & Simpson, 2002) while manual tasks (which may have been good indicators of workload based on primary task performance) are now mostly automated. Monitoring cognitive load using measures beyond performance is thus important since performance alone may not be sensitive and timely enough. For example, Pilot 1 may have plenty of attentional resources remaining while Pilot 2 used all his/her attentional resources and has no resources available should unexpected circumstances occur (e.g., answering a radio call) implying that the cognitive load is higher for Pilot 2 than for Pilot 1, even though their performance is identical (Yeh & Wickens, 1988).

Suitably, psychophysiological measures are leveraged to measure cognitive load demands objectively to account for differences that cannot otherwise be observed. This approach is based on the evidence that varying task difficulty influences psychophysiological signals (Haapalainen, Kim, Forlizzi, & Dey, 2010). Psychophysiological workload measurement techniques are based on the fact that the autonomic nervous system unconsciously regulates the bodily functions as workload changes (Mandrick, Peysakhovich, Rémy, & Lepron, 2016). Previous studies examining psychophysiological signals (e.g., heart rate (HR), galvanic skin response (GSR), etc.) used invasive equipment that either required the proper placement of electrodes with gel, or a chest strap that pilots found obtrusive and uncomfortable (Roscoe, 1992). Although these methods captured the required data, they can be problematic in operational settings due to the time-intensive preparatory work involved, the unnatural-like feeling experienced by pilots wearing these devices, and the technical expertise required for set-up. As such, the study of automatic arousal

detection of pilots by minimally invasive computerized systems is a promising avenue of research in human-computer interaction and is a focus of this study. As such, we will consider the combination of heart rate, galvanic skin response, and affective state recognition from facial expression analysis in this study.

Measuring Pilot's Heart Rate to Infer Cognitive Load

Heart rate has been the most widely studied measure of cognitive load in flight research, dating back to 1917, when Gemelli measured blood pressure, breathing rate, and pulse rate in experienced pilots during flight using an electrocardiogram (ECG) recording method (Roscoe, 1992). He found an association between increased stress and increased HR. This research has steadily continued in the flight environment (e.g., White, 1940; Lewis, 1967; Howitt, 1969; Sekiguchi, et al., 1977) and while they all found various HR changes depending on flight phases, little or no effort was put into explaining the results. The studies typically agreed that the landing phase caused the largest increase in HR for the pilot, followed by the takeoff phase of flight. As this research used real flight (as opposed to simulators), the transition phase between takeoff and landing was typically stable without any unforeseen events that could otherwise increase the complexity of the flight (e.g., mechanical failures, adverse weather conditions, etc.).

In the 1970s, researchers began to measure pilots' HR in flight simulators, which permitted increased task difficulty scenarios (e.g., low visibility, mechanical failures, adverse weather, etc.) that were previously not possible in real aircraft due to safety concerns (e.g., Roscoe & Goodman, 1973; Lindholm & Cheatham, 1983). Experienced pilots' HRs are typically lower during a simulated flight than during the same task in a real flight (Roscoe, 1992). Pilots without much experience in a particular flight simulator however, respond with HR increases with changing task demands. The popularity of this technique is unsurprising as it does not compromise flight safety and is readily accepted by pilots. ECG remained the most common method to measure HR in the flight simulator, however interest in less obtrusive, less expensive, and less complex HR monitors is on the rise (Wang & Fu, 2016).

The most common devices used to accurately measure heart rate beyond ECG recordings are chest straps which closely emulate a real ECG machine by measuring electrical pulse and sending that information to a wrist watch (Terbizan, Dolezal, & Albano, 2002). While these devices have been shown to be valid, they are obtrusive, uncomfortable, take considerable time to set up, and therefore are not ideal for a flight training environment. The recent ubiquity of wearable biosensors, such as smart watches and wristbands, enables the collection of HR data in a socially acceptable and non-invasive manner, ensuring it does not detract from the pilots' performance (Wang & Fu, 2016). These devices use a photoplethysmography (PPG) sensor which consists of LEDs (located at the back of the watch) that shine light onto the skin to detect changes in blood volume. The light that is scattered is sensed with a photodetector that is run through an algorithm to calculate HR.

Wang and Fu (2016) evaluated the accuracy of a strapless heart rate watch, the Mio Alpha (Mio Alpha; Physical Enterprises Inc., Vancouver, B.C.) against ECG-derived HR. Ten pilots performed three flight tasks (i.e., wind shear go-around, takeoff and climb, and hydraulic failure) in a full flight simulator with both the strapless and more invasive HR measurement devices. They found a high correlation between the HR watch measurements and the ECG-derived HR for all tasks, suggesting that the strapless HR watch provides valid measurements of HR during simulated flight tasks, and may be a useful tool for pilot workload evaluation. The Mio Alpha, however, transmits HR data over Bluetooth, and therefore would be prohibited in a military environment where wireless transmission is strictly banned. The use of a non-transmitting HR smart watch in a flight environment has not been explored, to our knowledge, and will therefore be explored in this study.

Measuring Pilot's Galvanic Skin Response to Infer Cognitive Load

GSR, also referred to as electrodermal activity (EDA), is a sensitive measure of changes in sympathetic arousal, which can be measured to assess cognitive load (Critchley, 2002). A person's GSR is directly related to their stress level so as stress increases, so does their GSR (Shi, Ruiz, Taib, Choi, & Chen, 2007; Brunken, Plass, & Leutner, 2003). Testing GSR to determine cognitive load in a flight environment began only in 2002, when Wilson tested ten pilots in a 90-minute visual flight rules (VFR) and instrument flight rules (IFR) real-flying scenario (2002). Each pilot repeated the same flight twice (up to several weeks apart) to test the reliability of the psychophysiological measures. GSR was collected from electrodes placed on the arch of the right foot. No statistically significant differences were found in GSR response between the two flights indicating reliability of the method. Within each flight, GSR activity was greatest during takeoff and landing, indicating the high cognitive load involved in these flight segments, however

subjective workload ratings did not indicate the same results. Subjective workload ratings suggested that pilots found the two IFR tracking segments as the most difficult. This may be due to the unfamiliarity of those segments, versus takeoff and landing which are familiar, even though cognitive load may be high. This same study found a high correlation between GSR and HR, with HR more sensitive to varying flight demands, indicating that perhaps it is sufficient to measure only HR. However, it is possible that GSR data might reveal unique information through a different sort of analysis. Overall, this study (Wilson, 2002) suggests that the combination of various workload measures provides a more accurate picture of the effect of piloting. However, both HR and GSR data do not reveal whether the arousal was a result of positive or negative stimuli, thereby indicating the case for measuring emotional valence, or the quality of the emotions.

Affective State Recognition from Facial Expression Analysis

Facial expression analysis software can computationally measure emotional valence by extracting and classifying, in real-time, specific key points in facial muscles of the target face (Ekman, 1992). In particular, the automatic detection and analysis of spontaneous micro-expressions could be used to discriminate between positively and negatively valenced affective states. For example, tension in the forehead muscles, elevated eyebrows or tightened lips are generally associated with a negative affective experience while raised cheeks, pulled lip corner, or slight (subtle) smiles signal a positive one (Leppanen & Hietanen, 2005; Pantic & Bartlett, 2007). These features are used to infer the affective state, where negative affective states such as fear, confusion or stress could be associated with cognitive overload, while positive or neutral affective states could signal suitable cognitive load (Gluck & Gunzelmann, 2013).

Previous research on affective states and their impact on performance in a flight simulator has been mostly constrained to self-report measures, both verbal and paper-based, which both require psychological disengagement from the simulation (Tichon, Mavin, Wallis, Visser, & Riek, 2014). Pupillometry and electromyography (EMG) was recently explored in this environment and Tichon et al. (2014) found a high correlation between saccade rate, muscle activation, and self-report measures of anxiety. While these results were promising, the obtrusiveness and high setup time of the equipment (i.e., large goggles/headset for pupillometry, and several electrodes for EMG) makes it impractical for a flight training environment, where both flight instructors and student pilots do not have the time nor expertise required for the set-up. Facial expression recognition, conversely, is noninvasive and minimally obtrusive and does not require any timely setup for each pilot (i.e., a camera is installed in the cockpit once).

While affect recognition using facial expression recognition analysis has not been studied in the cockpit of an airplane, Bullington (2005) identified it as the most likely environment to employ this technology. He identified benefits such as detecting an overloaded pilot thereby allowing automated systems to engage to improve flight safety. As such, the present research employs this technology in a full flight simulator for the first time. Additionally, this dataset (i.e., valence) is combined with psychophysiological data (e.g., GSR, HR) representing arousal, and mapped onto the Circumplex Model of Affect with the student's performance to assess the training experience (Chaouachi & Frasson, 2012; Chaouachi, Chalfoun, Jraidi, & Frasson, 2010).

Circumplex Model of Affect

The Circumplex Model of Affect (Russell, 1980) holds that all affective states can be described as linear combinations of two underlying, independent neurophysiological systems, valence on x-axis and arousal on y-axis (Posner, Russell, & Peterson, 2005). HR and GSR are widely used to assess the level of arousal (Lang, Greenwald, Bradley, & Hamm, 1993) while facial recognition analysis provides the valence score (Leppanen & Hietanen, 2005). The combination of facial expressions, HR and GSR could be used to situate pilots in one of the four quadrants (areas) of the Circumplex Model as a way to classify the student's workload (arousal) and the impact that workload is having on their performance (valence).

Classifying workload in real time in this manner can improve flight safety and pilot training systems. In a real flight environment, automated systems can engage to decrease pilot workload if necessary, or vice-versa (i.e., disengage to prevent errors due to boredom). During pilot training, the flight instructor can be aware of the student's cognitive state and can adjust the level of difficulty of the training to reap maximal training benefit by keeping the student engaged. Moreover, a flight instructor can assess the level of difficulty of the whole training program by mapping the participants' states in the Circumplex Model to assess the average induced effect.

This paper will argue that a multimodal approach to measuring cognitive load in pilots, in a non-obtrusive and minimally invasive manner is a worthwhile avenue of research. The objectives of this work are to provide guidelines for the multimodal assessment of pilots' arousal level using noninvasive biosensors to answer the following three questions: i) Which data processing approach must be adopted to track pilots' arousal; ii) Can the Circumplex Model of Affect be used to interpret these measurements; and iii) How can these measurements be used as training assessment criteria? The hypothesis of this work is that a combination of HR, GSR, and facial recognition data can be mapped onto the Circumplex Model of Affect to predict the pilot's affective state, indicating whether or not he/she is at an optimal training level, overloaded, or under-stimulated.

METHODOLOGY

We conducted this study at the CAE flight training facility in Montreal, Canada. We recorded HR, GSR, and facial expressions while participants completed a 45-minute flight scenario with tasks of varying difficulty levels. Each participant filled out a demographics questionnaire before the flight, and a cognitive load questionnaire where they rated their perceived cognitive load after each of the flight tasks.

Pilot Study

We first ran a pilot study with four male pilots, to test the methodology and compare several psychophysiological sensors. We used the same flight path (described below) but a different flight simulator due to availability (a C295 twin-turboprop tactical military transport full flight aircraft simulator). Based on the pilot study, we eliminated voice recognition from the biometrics as the student pilots do not speak often enough during the training flight to allow for a substantial analysis. We also determined to use the Empatica E4 sensor instead of the Scosche Rhythm+ (Scosche Industries, Oxnard, CA) for HR collection. The Scosche Rhythm+ used Bluetooth to stream data, which resulted in a lot of noise in the data, likely due to the interference within the aircraft. The Empatica E4 can store data locally on the device, and has been shown to be reliable in previous research (e.g., Gjoreski, Gjoreski, Lustrek, & Gams, 2016; Hoover, MacAllister, Holub, Gilbert, & Winer, 2016).

Participants

We recruited twelve participants (eleven males) from within CAE in Montreal, Canada who had a pilot's license and prior flying experience. Participants' previous flight hours ranged from 70 to 225 hours with a mean of 145.83 (*SD = 48.94*) hours, while simulator flying experience ranged from 0 to 15 hours, with a mean of 3.92 (*SD = 7.04*) hours. None of the pilots were experienced in the simulator platform used for the experiment. Pilots ranged in age from 20 to 54 years of age with a mean age of 33.42 (*SD = 12.50*) years. All participants gave their informed consent to participate in the study.

Simulator Apparatus

Participants conducted the flight using a CAE Airbus A310-221 full flight simulator that has been in operation since 1987. It has a six-degrees-of-freedom motion base platform, and belongs to the CAE 600 Series. This simulator has

Figure 1. Airbus A310-221 full flight simulator (CAE, 2017)

the highest Federal Aviation Administration (FAA) rating of Level D, indicating that the arrangement of the cockpit and aerodynamic models are identical to the actual aircraft. The cockpit contains realistic sounds, and special motion and visual effects. The visual system is a 5-channel CAE Tropos IG, using Canon projectors with a Windows 6 system display. There are two adjustable seats in the cockpit, where the participant occupied the left, pilot side, and the flight instructor/co-pilot occupied the right side. Figure 1 shows both the exterior and the cockpit.

Psychophysiological Data Collection

Participants wore the Empatica E4 sensor (Figure 2) that collected their real-time HR and GSR at 1 and 4 Hz respectively. This medical grade device is portable, wearable on the wrist, and does not require a chest strap or electrodes. It uses a PPG sensor to measure HR using two green LEDs located on the back of the watch, and a GSR sensor to measure the electrical conductance of the skin. The data was recorded locally on the device then transferred to a computer where the rest of the processing was performed.

Figure 2. Empatica E4 sensor

A pilot-facing, cockpit mounted GoPro HERO4 Black camera recorded the participant's facial expressions during the flight task. The camera recorded at a resolution of 1080p. The videos were processed using Noldus FaceReader software. Noldus FaceReader analyzes the pilot's facial expressions and provides the probabilities of their emotional states, with six values every second. The algorithm automatically extracts the emotional reactions from each frame and recognizes Ekman's six basic emotions, namely, *happy*, *sad*, *angry*, *surprised*, *scared* and *disgusted 2* (Ekman, 1992), in addition to the *neutral state.* The detection process is performed using an Active Appearance Model which models participants' facial expressions and a Neural Network, which classifies in real-time 500 key points in facial muscles of the target face. In addition to the probability of the presence of these six emotions, the software output also contains the probability of the *valence* of the emotional state. Information about the emotional valance defines the nature of the emotion and ranges from −1 to +1. A positive valence value refers to pleasant emotions, whereas a negative valence value characterizes unpleasant emotions. The accuracy of Noldus FaceReader has been validated through studies comparing its results with human coders and self-reports of emotional states (erburg, Wiering, & Uyl, 2005; Terzis, Moridis, & Economides, 2010).

Simulator Procedure

Participants flew an approximately 45 min simulated flight in the full flight simulator with a flight instructor who also acted as the co-pilot. All participants received the same instructions from the flight instructor, which consisted of conducting the following seven critical events:

1. Maximum effort takeoff
2. Right 360 degree turn followed by a left 360 degree turn at 45-degree angle of bank (AOB)
3. Left 360 degree turn followed by a right 360 degree turn at 60-degree AOB
4. Maximum effort landing attempt
5. Unplanned go-around due to an obstacle on the runway, and co-pilot incapacitated on go-around
6. Circuit back to landing attempt
7. Second maximum effort landing

After events 1, 2, and 3, participants were instructed to maintain 2000 feet straight and level, for 2 minutes. The flight instructor rated their performance on a scale of 1 to 5 (1 = needs improvement, 5 = excellent) based on deviations of the flight status (e.g., airspeed, heading, AOB, altitude, etc.) for all seven flight tasks. During the flight, the participants' HR and GSR were recorded using the Empatica E4 biosensor wristwatch, and facial expressions were recorded using a cockpit mounted GoPro camera. Participants completed a subjective workload questionnaire post-flight where they rated their perceived cognitive load on each of the seven critical events.

RESULTS

Post Processing of raw GSR Signals

GSR data, which is used to infer the skin conductance (SC), consists of a superposition of a tonic and a phasic component, called respectively skin conductance level (SCL) and skin conductance response (SCR). The SCL varies

slowly and changes slightly within an individual participant, depending on their skin dryness, autonomic regulation, and hydration levels, and can also vary markedly across individuals (iMotions, 2017). As such, it is often not informative on its own. The SCR consists of peaks on top of the tonic response, called GSR peaks. The SCR is sensitive to emotional stimuli, and occur between 1-5 seconds after the onset of a stimulus. SCL is generally associated with general states of arousal and alertness while SCR is related with attention processes (i.e. how the person perceives the relevance, novelty and intensity of the stimulus) (Henriques, Paiva, & Antunes, 2013). However before extracting both components, noise in GSR signal caused by sudden movements must be filtered out because it might be mistaken for spontaneous shifts in the stress level. To this end, we used a *median filter technique*, which removes the local disturbances while maintaining the shape of the signal and the typical peaks. Then we split the GSR signal into its tonic and phasic components using *a deconvolution method* (Benedek & Kaernbach, 2010a; Benedek & Kaernbach, 2010b). In this study we focus on the analysis of the phasic GSR component. Thus, we extracted the z-scores sum of the amplitude (i.e. peak height) of the detected SCRs (i.e., SCR Amplitude-Sum) for all the participants during each flight event. We performed z-score normalization to allow comparison of these components between all pilots.

GSR and Performance Results

Group Performance. Our first objective was to assess whether there was a difference between the pilots' cognitive load and their performance. More precisely we aimed to investigate if for example, a pilot who had difficulty following the instructor's directives or maintaining control of the aircraft trajectory during the 45 degree AOB turn would experience more intense reactions and thus have a higher SCR amplitude. Three types (groups) of performance were considered for each flight event: high performance (pilots who perfectly executed the manoeuver and had a grade of 4 or higher), medium performance (pilots who had an average to good performance and obtained a grade between 3 and 4 in the flight event) and low performance (pilots who missed their manoeuver and had a rating lower than 3). A one-way between subjects ANOVA was conducted to compare the level of SCRs Amplitude-Sum for the three performance levels across all the flight events. Results showed an overall significant impact of the pilots' performance on their GSR ($F(2, 65) = 4$, $p<0.05$). Figure 3 depicts the mean normalized SCRs Amplitude-Sum for all three performance groups. Post hoc comparisons using Tukey HSD tests showed that low performance in a flight event had significantly higher SCRs Amplitude-Sum (*M=0.18, SD=0.52*) than medium performance (*M=0.47, SD=0.28*) and almost significantly SCRs Amplitude-Sum than high performance (*M=0.22, SD=0.52, p=0.06*). However, no significant difference was obtained between high performance and medium performance. This result suggests an interplay between cognitive load and performance which is not necessarily linear as lowest and highest performers showed high levels of SCR amplitude.

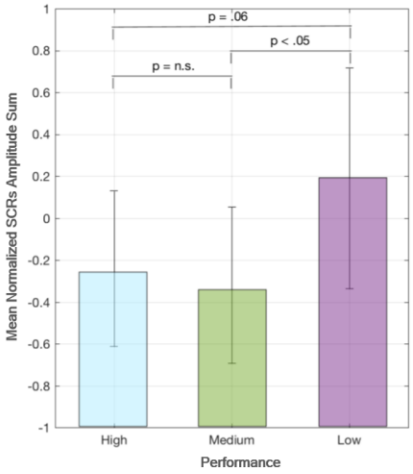

Figure 3. Mean SCR amplitudes

Individual performance. The correlational analysis of the SCRs Amplitude-Sum with respect to each pilot's individual performance showed a positive correlation for six pilots (with *r* ranging from *0.25* to *0.77* with a mean correlation r=*0.49*). A negative correlation between performance measures and SCRs Amplitude-Sum was obtained for two pilots (*r=-0.35* and *r=-0.67*) and no correlation was found for the remaining pilots (*r* around zero). These findings suggest that the cognitive load and performance relationship is not linear, and pilots may employ adaptive strategies which may be related to their experience.

Training assessment. The second objective of this experiment was to evaluate the evolution of the GSR data – particularly the SCRs Amplitude-Sum- across the different flight events. We aimed to investigate whether the different events of the flight training induced different levels of cognitive load on the pilots and which training episode was the most demanding. A one-way repeated measure ANOVA showed a significant effect of the flight task performed by the pilot on their SCRs Amplitude-Sum ($F(6, 60) = 4.142$, $p < .001$). A series of 45 paired samples t-tests were undertaken to make post hoc comparisons between the different flight events, 15 among them were statistically significant. More precisely, there are significant differences in the normalized SCRs Amplitude-Sum of the participants starting from the second baseline of the experiment to the first right turn. Normalized SCRs Amplitude-Sums were significantly higher in the second baseline of the session (*M=-0.16, SD=0.45*) when compared to the 45

degree AOB turn (*M=-0.77, SD=0.32*) and the first baseline (*M=-0.75, SD=0.57*) with *t(10)=-6.9*, *p<0.05* and *t(10)=-6.9*, *p<0.05* respectively. The 60 degree AOB turn (*M=0.36, SD=0.86*); second baseline (*M=0.33, SD=0.81*); the first landing attempt (*M=0.34, SD=0.87*) and the go around (*M=0.73, SD=1.51*) events were also significantly higher than the 45 degree AOB turn (*t(10)=-4.63*, *p<0.01; t(10)= -3.94*, *p=0.02; t(10)=-3.44*, *p<0.01; t(10)= -3.94*, *p<0.01*) and the first baseline (*t(10)= -3.89*, *p<0.01; t(10)= -3.22*, *p<0.01; t(10)= -2.79*, *p<0.01; t(10)= -3.33*, *p<0.01*). The t-test also showed that the normalized SCRs Amplitude-Sum during the approach (*M=0.43, SD=0.9*) and the landing (*M=-0.34, SD=0.76*) were significantly lower compared to the go around with *t(10)= -2.61*, *p<0.05* and *t(10)= -3.64*, *p<0.05)* respectively. The approach was also significantly higher than the first baseline and the 45 degree AOB turn with *t(10)=* -3.47, *p<0.01* and *t(10)= -2.87*, *p<0.01)* respectively. To sum up, as depicted in Figure 4, three main phases can be identified in the flight training path with respect to the variation of SCRs Amplitude-Sum. The first (first green zone of Figure 4) starts from the takeoff and ends after the 45 degree AOB turn. During this phase, pilots expressed a low to moderate level of SCR. Then, starting from the second straight line (i.e., baseline), the participants' SCRs Amplitude-Sum physiological signal started to significantly increase compared to the first phase across all the participants (red zone in Figure 4). Finally, a third phase (second green zone in Figure 4) starting from the approach to the end of the landing showed lower SCRs Amplitude-Sum compared to the previous phase but still significantly higher than the first phase.

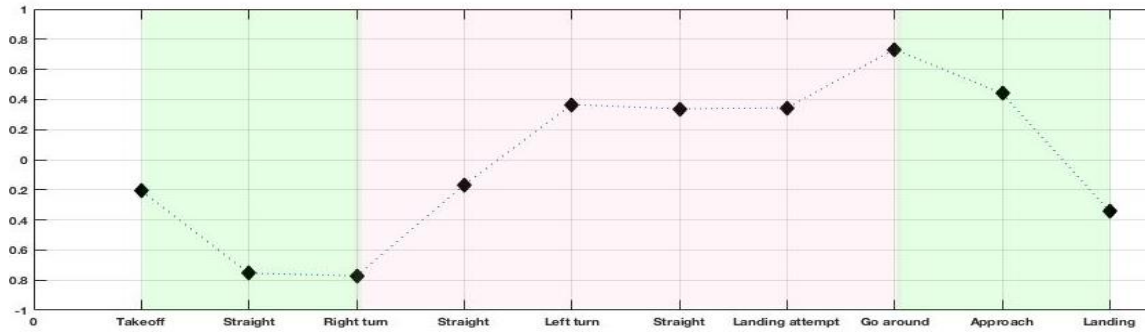

Figure 4. Average SCR for all participants with the three identified zones

HR Results

We normalized the HR data for all the participants in order to allow between-subject comparison for all the flight events. The normalization was performed by computing a z-score of the signal (i.e., subtracting the mean HR and dividing by the standard deviation). The analysis of HR showed a positive correlation with GSR across all the events (*r=0.43*), however no significant results were found relating HR to the pilots' performance nor to the impact of the different flight events.

Figure 5 presents the mean normalized SCRs Amplitude Sum for each event, normalized HR, and normalized subjective workload ratings across all participants, for each of the flight events. There is no subjective workload rating for the 3 baseline measures, as participants were not asked to rate those events, however GSR and HR data are collected continuously. Subjective workload ratings do not show a significant correlation with the GSR or HR data. Figure 5 suggests that the Go-Around event invoked the highest mean HR and SCRs Amplitude Sum, however this result was not significant for the HR.

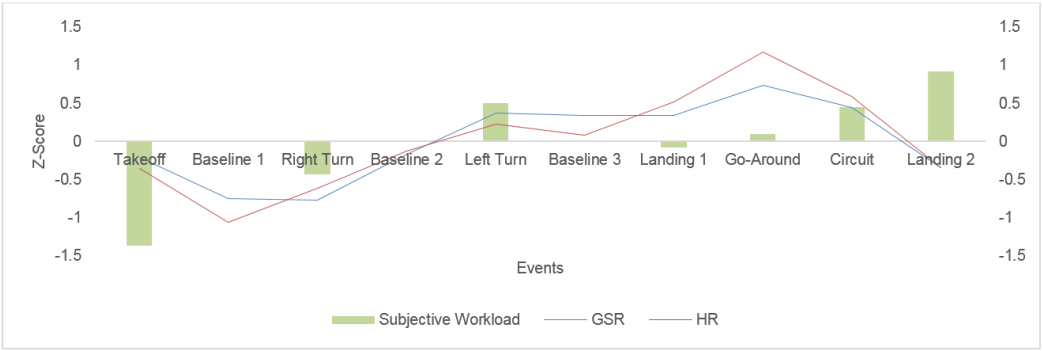

Figure 5. Average normalized Subjective Workload score, amplitude sum of SCR's, and HR for all participants

Facial Recognition

We segmented the recorded video for each participant into seven sequences corresponding to the main flight events. As FaceReader provides a score between 0 and 1 for each frame of each participant's video for the six basic emotions and the neutral state, we first computed a list of emotions, whose scores exceeded a minimal threshold value of 0.1 for more than 2 seconds while not completely vanishing for more than 1 second. Table 1 shows the most dominant emotion in terms of frequency, for each of the twelve participants for each flight task. Any missing data (denoted by a dash) is due to either camera issues while recording (e.g., insufficient battery) or processing issues due to insufficient/abundant light in the frames. The mode emotion for each task is bolded, where 'Surprise' is the most dominant emotion between all tasks/participants.

Table 1. Mode Emotional Response

Takeoff	Right Turn	Left Turn	Landing #1	Go-Around	Circuit	Landing #2
Joy	**Surprise**	Fear	**Fear**	Fear	Fear	**Surprise**
Fear	**Fear**	**Surprise**	-	-	-	-
Joy	**Fear**	-	-	-	-	-
-	-	-	-	-	-	-
Fear	**Fear**	**Surprise**	**Fear**	**Contempt**	**Sadness**	Contempt
Sadness	Sadness	**Sadness**	**Sadness**	**Contempt**	**Sadness**	**Sadness**
Fear	**Surprise**	**Surprise**	-	-	-	-
Surprise	**Surprise**	**Surprise**	**Surprise**	**Surprise**	Surprise	Fear
Sadness	Sadness	**Surprise**	**Sadness**	Sadness	**Sadness**	**Sadness**
Disgust	Joy	Joy	**Joy**	Joy	Fear	Joy
Sadness	Sadness	Disgust	**Joy**	Disgust	Disgust	Disgust
Surprise	Disgust	Sadness	**Surprise**	**Surprise**	Surprise	**Surprise**

Circumplex Model of Affect – Framework for Pilot Performance

The first landing attempt and the ensuing go-around were the two tasks that elicited the most significant arousal response from the participants (see Figure 4 and Figure 5), thus we selected them for a case study to map onto the Circumplex Model of Affect. Figure 6 maps the normalized GSR Amplitude-Sum along the y-axis (i.e., arousal), the valence value from facial recognition analysis along the x-axis, and the dominant emotion for each of those data points. Each data points represents one participant, where red data points indicate a low performance score (i.e., 1-2), orange a moderate performance score (i.e., 3), and green a high performance score (i.e., 4-5). Only data from seven participants was used due to incomplete emotion recognition data from four participants, and noisy GSR data from one participant. If we consider the Circumplex Model of Affect as a coordinate plane, quadrants three (i.e., lower left)

Landing # 1 **Go-Around**

Figure 6. Performance mapped with Valence, Arousal (from GSR), and Dominant Emotion

and four (lower right) in Figure 6 contain only high performing individuals. Meanwhile, low performance is constrained to the second quadrant (top left), along with one high performer, who is exhibiting '*disgust'* as his dominant emotion. The first quadrant (top right) contains all the average performance score, as well as a lot of '*surprise'* as the dominant emotion. While there are clear trends in this framework, more participants are required to validate the model.

DISCUSSION

We conducted a full flight simulator task with varying levels of flight difficulty where we measured student pilots' arousal and valence using HR, GSR, and facial emotion recognition analysis. We discuss our hypothesis that we can map this data onto the Circumplex Model of Affect to predict the pilot's affective state, which has significant training implications. This discussion is guided by the three objectives of this research.

Objective 1: Which data channel was the most informative to assess pilots' arousal?

GSR provides a clearer indication of the pilot's cognitive load/arousal during a flight task compared to HR. By filtering (using a median filter) and processing the GSR data (using a deconvolution method), we removed the tonic component and focused our analysis on the phasic, specifically the SCR-amplitude sum, which is not impacted by individual differences, hydration levels, skin dryness, and autonomic regulation. We found a significant interplay between GSR and group performance, where the low performing group (i.e., rating of 1-2) had the highest GSR, indicating that their cognitive load was also the highest. By assessing individual performance and GSR, six pilots showed a positive correlation between GSR amplitude-sum and individual performance (more than 50% of the pilots), two pilots showed a negative correlation while three pilots showed no correlation. The results suggest there is a complex non-linear relationship between the performance and GSR, and that many factors such as pilots' experience or personality caracteristics may play an important role to define how they modulate their response to performance/cognitive load demands. For example, in this experiment the pilot who showed the *highest negative* correlation between performance and GSR amplitude (*r=-0.67*) was the most highly experienced pilot (i.e. greatest number of flight hours). But at the same time, the second highest experienced pilot showed the *highest positive* correlation between performance and GSR amplitude (*r=0.77*). Thus, even though these two pilots had almost the same experience level and both had a high overall performance, their coping responses to the different performance/cognitive load demands was opposite. The first pilots tended to have low (respectively high) GSR Amplitude-sum with higher (respectively low) performance and thus seemed to have better performance when less aroused. Conversely, the second pilot showed better performance when he experienced less arousal and vice versa. Overall, GSR is a possible indicator of arousal, however the relationship with performance is not linear, and other characteristics (e.g., experience, motivation, personality, goals, etc.) of the pilots should be considered.

We found no significant differences between the HR during the different flight tasks, although the level of difficulty varied (according to the flight instructor) and the GSR response varied according to the level of difficulty. This finding is not consistent with several full flight simulator studies that found that the average HR was directly related to task difficulty (Roscoe, 1992; Wang & Fu, 2016; Dussault, Jouanin, Philippe, & Guezennec, 2005), as well as studies that found GSR and HR to have a high correlation (e.g., Wilson, 2002). It is unclear whether the previous studies also used a simulator with a six-degrees-of-freedom motion base platform or a stationary one, which may have an impact on the HR since it is highly related to movement. Suitably, we concluded that the GSR, specifically the SCR Amplitude Sum, is a better indicator for arousal ratings in pilots for the purpose of the Circumplex Model of Affect.

Objective 2: Can the Circumplex Model of Affect be used to interpret these measurements?

The Circumplex Model of Affect provides a framework where we mapped the pilot's arousal level using GSR, and valence using facial emotion recognition analysis for the two most difficult tasks: Landing Attempt 1 and the subsequent Go-Around. We found a trend based on performance, where quadrants one and two (in Figure 6) contained low (red dot) and medium (orange dot) performing individuals respectively, whereas quadrants three and four contained only high performing individuals (green dot).

These results are consistent with the learning cycle proposed by Kort and Reilly (2001) who suggest that students begin to learn in quadrant one or two, where they are typically curious and fascinated about a topic. The red and orange dots indicate that these students are still learning, and are not experts at the particular task. Additionally, in quadrant one, where there is a positive valence, they experience more positive emotions, which is consistent with our findings during these two tasks (i.e., surprise and joy). Students in quadrant two, on the negative valence side, experience more negative emotions, where we see fear and disgust during the two flight tasks. The performance scores (red and orange dots) are lower in these two quadrants as the students are still in the earlier learning phases for the assigned tasks, as described by Kort and Reilly (2001). Conversely, the performance scores are higher (represented by green dots in Figure 6) in the lower quadrants where students are not learning new tasks, as they are already somewhat more

experienced. On the left side of the valence scale, there are more negative emotions, such as sadness, fear, and contempt, while on the right side, only joy.

Although this case study had limited participants, it is promising to see that phases of learning (suggested by performance score) can be related to emotions, indicated by GSR (arousal) and facial recognition analysis (valence). The implications for assessing performance and adapting the training program based on this framework are considered next.

Objective 3: How could these measurements be used as training assessment criteria?

The Circumplex Model of Affect, adapted to pilot training performance using GSR (arousal) and facial emotion recognition analysis (valence) has potential to be used to assess both individual pilot training, and full training programs. This framework allows instructors to monitor which quadrant a student is performing in to determine, based on learning theories (Kort & Reilly, 2001) how the training needs to be adapted (if at all). For example, a student is in an optimal learning zone when they are on the positive valence and arousal side (i.e., quadrant one; Feidakis & Daradoumis, 2012). However, a student should cycle through all four quadrants as they learn and acquire new skills therefore the instructor can progress onto more difficult tasks if the performance is flawless and on the positive valence side. This has the possibility of optimizing performance, and also ensuring the student does not get bored. While we were not able to find statistical evidence for this model, the trend is very encouraging for future research.

On a training program level, this framework provides the instructor with a tool to map how all students are performing to determine if the training program is at an appropriate level of difficulty. For example, if most students are scoring in a negative valence and negative arousal, and performance is low, then it suggests that the training is too difficult and requires adjustments (Tichon, Mavin, Wallis, Visser, & Riek, 2014). The opposite is also true if the training is too easy, which may be seen as a poor use of resources. The analysis of each physiological sensor across all the participants could also be informative to assess the difficulty of a training program and to identify the phases eliciting the highest arousal level. In this research, for example, the analysis of GSR showed that three different zones could be clearly identified in the training program (see Figure 5). The first zone elicited a decreasing level of GSR across all the participants; the second zone a significantly increasing GSR and finally, a third zone showed a significantly decreasing GSR. We also found that the Go-Around event was the event that evoked the highest GSR and HR for all the participants which was not expected since in theory, the takeoff and the landing were supposedly the most difficult events. Such a deconstruction of the different training sessions could be highly beneficial for the training program designers as well as for the instructor to help them fine-tuning the training elements and to anticipate, grade and understand trainees' problems.

While full flight simulators are much less expensive to operate than an actual aircraft, their costs and availability are still sufficiently prohibitive that the time spent using them should be optimized. As such, the use of physiological sensors can provide the instructor with a tool to monitor student's arousal and valence during training, indicating when he or she should make real time adjustments to optimize the training session. The sensors required for this analysis (i.e., a wristwatch GSR sensor and a cockpit mounted camera) are non-invasive, non-obtrusive, and do not require any specialized technical training or laborious setup time. It follows that they are ideal for this domain, where both instructors and students do not have the time or expertise to set-up complicated biosensors.

Despite the relatively small sample size, we obtained promising results, with significant trends between GSR and performance. We presented, for the first time, how facial emotion recognition can be employed in a full flight simulator, and then mapped to the Circumplex Model of Affect to provide a framework for training optimization. Future studies with larger sample sizes can build a more reliable model, perhaps in quasi or real time, to provide the instructor with objective data to optimize the training, ultimately improving flight safety for all involved.

ACKNOWLEDGEMENTS

We would like to thank Katherine Horvath, Lucas Pollok, Kassidy Schmitz, Matt Spadafora from George Washington University for their help with the pilot study. We would also like to thank Houssam Alaouie, Alain Bourgon, David Bowness, Gene Colabatistto, Aurelian Constantinescu, Graham Estey, Shelley Kelsey, Paula Mazzaferro, and Jack Russ who provided the necessary resources and logistical support to run this study. And finally, we'd like to thank the CAE Montreal Flight Training Facility and their staff, and all the pilots who volunteered to participate.

REFERENCES

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of neuroscience methods*, 80-91.

Benedek, M., & Kaernbach, C. (2010). Decomposition of skin conductance data by means of nonnegative deconvolution. *Psychophysiology*, 647-658.

Boeing. (2015). *Statistical Summary of Commercial Jet Airplane Accidents - Worldwide Operations 1959-2015.* Boeing.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 53-61.

Bullington, J. (2005). 'Affective' computing and emotion recognition systems: the future of biometric surveillance? *Information Security Curriculum Development (InfoSecCD) Conference '05* (pp. 95-99). Kennesaw, GA: ACM.

CAE. (2017). *Simulator Specs & Information.* Montreal: CAE.

Chaouachi, M., & Frasson, C. (2012). Mental workload, engagement and emotions: an exploratory study for intelligent tutoring systems. *Intelligent Tutoring Systems*, 65-71.

Chaouachi, M., Chalfoun, P., Jraidi, I., & Frasson, C. (2010). Affect and mental engagement: towards adaptability for intelligent systems. *In Proceedings of the 23rd International FLAIRS Conference.* Florida: AAAI Press.

Critchley, E. (2002). Electrodermal responses: what happens in the brain. *Neuroscientist*, 132-142.

Dussault, C., Jouanin, J. C., Philippe, M., & Guezennec, C. Y. (2005). EEG and ECG changes during simulator operation reflect mental workload and vigilance. *Aviation, Space, and Environmental Medicine*, 344-351.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 169-200.

Feidakis, M., & Daradoumis, T. (2012). Design of an emotion aware e-learning system. *International Journal of Knowledge and Learning*, 219-238.

Gjoreski, M., Gjoreski, H., Lustrek, M., & Gams, M. (2016). Continuous stress detection using a wrist device - in laboratory and real life. *UbiComp/ISWC* (pp. 1185-1193). Heidelberg, Germany: ACM.

Gluck, K. A., & Gunzelmann, G. (2013). Computational process modeling and cognitive stressors: Background and prospects for applications in cognitive engineering. In J. D. Lee, & A. Kirlik, *The Oxford handbook of cognitive engineering* (pp. 424-432). New York: Oxford University Press.

Haapalainen, E., Kim, S. J., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-Physiological Measures for Assessing Cognitive Load. *UbiComp '10*, (pp. 301-310). Copenhagen.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, 139-183.

Henriques, R., Paiva, A., & Antunes, C. (2013). Accessing emotion patterns from affective interactions using electrodermal activity. *2013 Humaine Assocation Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 43-48). IEEE.

Hoover, M., MacAllister, A., Holub, J., Gilbert, S., & Winer, E. (2016). Assembly training using commodity physiological sensors. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, (pp. 1-12). Orlando.

Howitt, J. (1969). Flight-deck workload studies in civil air transport aircraft. *Conference Proceedings (No. 56) Measurement of Aircrew Performance.* Paris: AGARD.

iMotions. (2017). *GSR Pocket Guide.* iMotions.

Kort, B., & Reilly, R. (2001). *Analytical models of emotion, learning and relationships: towards an affect-sensitive cognitive machine.* MIT Media Lab Tech Report No. 548.

Kuilenburg, H. V., Wiering, M., & Uyl, M. D. (2005). A Model Based Method for Automatic Facial Expression Recognition . In J. Gama, P. Brazdil, A. Jorge, & L. Torgo, *Machine Learning: European Conference on Machine Learning* (pp. 194-205). Berlin: Springer.

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 261-273.

Leppanen, J. M., & Hietanen, J. K. (2005). Positive facial expressions are recognized faster than negative facial expressions, but why? *Psychological Research*, 22-29.

Lewis, C. (1967). Flight research programme IX: Medical monitoring of carrier pilots in combat II. *Aerospace Medicine*, 581-592.

Lindholm, E., & Cheatham, C. (1983). Autonomic activity and workload during learning of a simulated aircraft carrier landing task. *Aviation Space Environmental Medicine*, 435-439.

Mandrick, K., Peysakhovich, V., Rémy, F., & Lepron, E. (2016). Neural and psychophysiological correlates of human performance under stress and high mental workload. *Biological Psychology*, 62-73.

O'Donnel, R., & Eggemeier, F. (1986). Workload assessment methodology. In B. Kaufman, & T. Wiley, *Handbook of Perception and Human Performance, Volume II, Cognitive Processes and Performance* (pp. 41-42). New Work.

Pantic, M., & Bartlett, M. (2007). Machine analysis of facial expressions. In K. Delac, & M. Grgic, *Face Recognition* (pp. 377-416). Vienna, Austria: I-Tech Education and Publishing.

Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. *Frontiers in Psychology*, 1-11.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Developmental Psychopathology*, 715-734.

Roscoe, A. (1992). Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology*, 259-287.

Roscoe, A., & Goodman, E. (1973). *An investigation of heart rate changes during a flight simulator approach and landing task.* RAE Technical Memorandum Avionics.

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 1161-1178.

Sekiguchi, C., Hunda, Y., Gotoh, M., Kurihara, Y., Nagasawa, Y., & Kuroda, I. (1977). Continuous ECG monitoring on civil air crews during flight operations. *Aviation Space Environmental Medicine*, 872-876.

Shi, Y., Ruiz, N., Taib, R., Choi, E., & Chen, F. (2007). Galvanic Skin Response (GSR) as an index of cognitive load. *CHI* (pp. 2651-2656). San Jose: ACM.

Skinner, M. J., & Simpson, P. A. (2002). Workload Issues in Military Tactical Airlift. *The International Journal of Aviation Psychology*, 79-93.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 257-285.

Terbizan, D. J., Dolezal, B. A., & Albano, C. (2002). Validity of Secen Commercially Available Heart Rate Monitors. *Measurement in Physical Education and Exercise Science*, 243-247.

Terzis, V., Moridis, C., & Economides, A. (2010). Measuring instant emotions during a self-assessment test: the use of FaceReader. *Proceedings of the 7th International Conference on Methods and Techniques in Behavioral Research* (pp. 181-184). New York City: ACM.

Tichon, J. G., Mavin, T., Wallis, G., Visser, T. A., & Riek, S. (2014). Using papillometry and electromyography to track positive and negative affect during flight simulation. *Aviation Psychology and Applied Human Factors*, 23-32.

Wang, Z., & Fu, S. (2016). Evaluation of a strapless heart rate monitor during simulated flight tasks. *Journal of Occupational and Environmental Hygiene*, 185-192.

White, M. (1940). The effect of anoxia in high altitude flight on the electrocardiogram. *Journal of Aviation Medicine*, 166-180.

Wilson, G. F. (2002). An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures. *The International Journal of Aviation Psychology*, 3-18.

Yeh, Y., & Wickens, C. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, 111-120.

Yerkes, R. M., & Dodson, J. (1908). The relationship of strength of stimulus to rapidity of habitat formation. *Journal of Comparative Neurology*, 459-482.