This is authors version. The published paper is available online at Springer: https://doi.org/10.1007/978-3-032-29459-3_18

Author-accepted manuscript

Psychophysiological Response, User Experience and Performance in Routine and Emergency Flight Simulation Training Scenarios Using Multimodal Wearable Sensing
AF Shafi, JE Muñoz, S Cao
International Conference on Human-Computer Interaction, 256-274, 2026

Psychophysiological Response, User Experience and Performance in Routine and Emergency Flight Simulation Training Scenarios Using Multimodal Wearable Sensing

Anika Fairooz Shafi[1], John Edison Muñoz[1,2] and Shi Cao[1]

[1] Systems Design Engineering, University of Waterloo, Waterloo ON N2L 3G1, Canada
[2] User Experience Design, Wilfred Laurier University, Brantford ON N3T 2Y3, Canada
{afshafi,john.munoz.hci,shi.cao}@uwaterloo.ca

Corresponding author: afshafi@uwaterloo.ca

Abstract. Wearable sensors are continually extending the possibilities for physiological monitoring and physio-adaptive (biocybernetic) systems, with particularly relevant applications in pilot training. This study explored psychophysiological responses, performance and user experience of pilot trainees (n = 16) as they interacted with routine and high-stress flight simulation scenarios. Outcome measures include electrodermal activity (EDA), heart rate (HR) and heart rate variability (HRV) parameters, competency-based performance metrics and perceived user experience and stress. Sympathetic activation was observed, characterized by elevated HR and reduced HRV time-domain measures (Mean NN, RMSSD, and pNN50) alongside decreases in non-linear measures (SD1 and the SD1/SD2 ratio) across progressively more stressful conditions. Self-reported stress significantly differed across conditions, while EDA data showed notable individual variability. Performance, measured by mean application of procedure scores, was significantly affected. Results indicate initial evidence that the multimodal wearable physiological sensors employed in the study sensitively and consistently gauge changes in stress inducements. These findings have important implications for stress inoculation training, personalized instructor feedback and development of future biocybernetic training systems.

Keywords: Physiological Monitoring, Multimodal Psychophysiology, Wearable Sensors, Adaptive Simulation Training

1 Introduction

Physiological monitoring using wearable sensors allows for non-invasive, near real-time, objective and continuous quantification of psychophysiological states such as stress. Given how affordable and accessible wearables are becoming and advances in sensing technologies, we can now easily integrate psychophysiological responses into simulations. In the context of aviation, wearables have been utilized to investigate

various factors such as arousal, gaze behavior, performance [1-8], while a smaller number of studies focussed on assessing stress [4-10]. There is a need for flight simulator studies to validate the use of unobtrusive sensing technologies for assessing pilot states [3] and reliably inducing stress during training and testing [11]. Moreover, systematic coupling of stress training with flight skill acquisition has shown to enhance performance in future stressful situations [12]. The few studies that have examined stress responses in pilot trainees using wearables have predominantly relied on a single modality [6-9]. Cacioppo and Tassinary [13] proposed a construct to describe the relationship between physiological measures and psychological states, such that if all psychological states constitute one domain and all physiological measures another, five general categories may represent the possible relationships: one-to-one, one-to-many, many-to-one, many-to-many and null. For instance, in a one-to-one relationship, a single psychological state corresponds uniquely to a physiological measure, while in one-to-many, one psychological state is associated with multiple physiological measures. Considering the non-exclusive and complex nature of psychophysiological inference and validation, a multimodal approach is essential for characterizing stress responses and in designing and implementing adaptive training systems.

Early work by Kelley [14] explored the potential of adaptive flight simulation training "in which the problem, the stimulus, or the task is varied as a function of how well the trainee performs" (p. 547). They proposed five key elements of an adaptive trainer: the adaptive variable, measuring performance, adaptive logic, error standard and difficulty level and knowledge-of-results displays (informing the trainee of their progress). In the same year, Caro [15] described an army-wide system of aircraft simulators, Synthetic Flight Training System (SFTS), aimed at automating the flight instructor's function of problem selection to effectively challenge trainees. Moving from performance as the adaptive variable to psychophysiological metrics, physio-adaptive (biocybernetic) systems leverage psychophysiology to sense, analyze and respond to a user in real-time [16]. Palsson and Pope [17] proposed stress counter-response training based on the concepts of instrument functionality feedback and biocybernetic modulation, where a flight simulator changes its behavior according to a trainee's stress management skills. Other scholars have proposed systems that use a trainee's state to modify flight simulation exercise [18], training content and presentation [19]; and context-aware learning such that flight instructors and trainees know when mastery of a skill has been achieved [20].

There is a need for physio-adaptive systems that integrate states beyond mental workload or engagement [21]. Before a trainee's stress levels can be used to adapt training, research needs to explore which combinations of wearable sensors and metrics reliably and sensitively capture such psychophysiological state change, effectiveness of stress inducement scenarios and best ways and practices to collect and synchronize multimodal physiological data while maintaining signal quality. These are gaps that this current study seeks to address by studying psychophysiological response, performance and user experience in simulated routine and high-stress flight scenarios using multimodal wearable sensing.

2 Related Work

Psychophysiological measures have been utilized in a wide range of aviation environments. Electroencephalography (EEG) measures electrical activity of the brain and is a reliable tool to study and predict psychophysiological states in different task conditions [22]. Spontaneous brain activity can be characterized by oscillatory rhythms, which can be quantified using power spectral density (PSD) into different frequency bands, including delta (δ, 1–4 Hz), theta (θ, 4–8 Hz), alpha (α, 8–12 Hz), beta (β, 12–30 Hz), and gamma (γ, 30–100 Hz) and show certain relationships with stress. Another widely supported measure is the P300 event-related potential (ERP) (also referred to as 'evoked potential' and is the brain's response to an event or stimuli) [23]. Additionally, central to understanding stress responses are two branches of the autonomic nervous system: sympathetic and parasympathetic nervous systems. The sympathetic nervous system is responsible for mobilizing our body's response to threat or stressors, known as the "fight-or-flight" response. In contrast, the parasympathetic nervous system promotes relaxation and recovery or "the-rest-and-digest" response [24]. Of the autonomic measures, heart rate variability (HRV) is the variation in time interval between successive heartbeats (measured as RR intervals in an ECG), while Electrodermal Activity (EDA) is the electrical conductivity of the skin. HRV, a measure of autonomic balance [25] and EDA, a reliable quantifier of sympathetic arousal [26].

In commercial airline pilots, HRV has been used to explore the impact of maneuver difficulty on performance [4]. Liang et al. [27] tested stress induction with Tier Social Stress Test (TSST) in a sample of 42 pilots, all holding commercial pilot licenses. The authors utilized heart rate, State Anxiety Inventory and salivary cortisol concentrations. Additional research in drone, general aviation, military pilots and non-pilots leveraged HRV and EEG to explore stress responses. Wojciechowski et al. [10] sought to elucidate the impact of stress on a single drone pilot's sympathetic and parasympathetic activity in familiar and unfamiliar simulation environments. Wang et al. [28] reported that different flight simulation tasks successfully induced varying stress responses in a group of 20 university engineering students. In another study, researchers evaluated 11 fighter pilots' stress levels as they conducted high performance maneuvers in long-term training [5]. Bustamante-Sánchez et al. [29] analyzed the effect of night and instrument flights on cortical arousal, autonomic modulation, and stress perception. Causse and Dehais [30] demonstrated that compared to other flight segments, take-off and landing involve higher stress. Bruna et al. [31] assessed stress inducement in ultra light aircraft simulators using HRV and respiration measures. Lastly, Villafaina et al. [32] used both HRV and EEG band powers to study the psychophysiological responses of Spanish Air Force pilots to different simulated combat maneuvers.

Among pilot trainees, most of the studies used HRV and rarely employed EDA [8]. With a small sample size of 5 pilot trainees and 5 instructors, Lee [7] studied the role of flight experience, approach area and glide angle on stress and performance. Peng et al [6] explored how urgency and familiarity impact stress and performance in pilot trainees during unexpected events. In another study, 9 aviator cadets and 6 instructors performed a series of tasks mirroring ground-based flight training allowing the authors to explore the utility of HRV time-domain, frequency-domain and non-linear

parameters in stress response monitoring [33]. Regula et al. [9] in their study with beginner pilots with no flying experience examined how transitioning from analog to digital presentation of avionic data elicits enhanced stress responses in trainees. Overall, almost all prior studies relied on a single modality, leaving multimodal assessment in pilot trainees largely unexplored.

3 Method

The setup (see Fig. 1). was utilized to create a controlled environment where pilot trainees interacted with the ALSIM AL 250 flight simulator [34], while data on their physiological responses, user experience, performance, and self-reported levels of stress were collected.

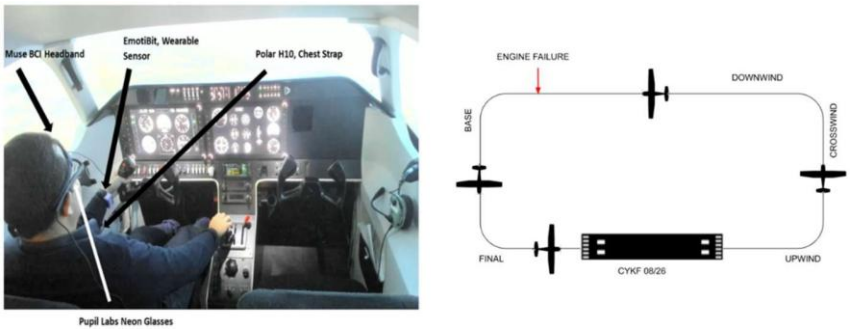


Fig. 1. The various physiological sensors employed in the study (left) and the standard left-hand circuit flown by the pilot trainees (right).

3.1 Participants

A total of sixteen pilot trainees (male: 11, female: 5; age range: 18-34 years, mean: 21.53 years) participated in the study. All participants had prior experience with simulation technology, and most belonged to a single university aviation program. Additionally, the majority already obtained their Private Pilot License (PPL) or Commercial Pilot License (CPL), along with various ratings, including, night, multi-engine and instrument flight ratings. The study was approved by the Research Ethics Board at the University of Waterloo (45409) and conducted in alignment with the principles outlined in the Declaration of Helsinki. All participants provided written informed consent prior to the start of the study. De-identified data was stored in a password-protected OneDrive folder.

3.2 Scenario

ALSIM AL 250 flight simulator was configured to replicate the aircraft dynamics of a single-engine Cessna 172. The simulation scenario employed weather settings

reflecting Ceiling and Visibility OK (CAVOK) conditions suitable for Visual Flight Rules (VFR) flight. Visibility was plus 6 statute miles with 110 at 10 kts wind, no clouds, altimeter setting: 29.92 ″Hg and a surface temperature of 15°C. Pilot trainees were tasked with performing one taxi and three standard left hand circuits at the Region of Waterloo International Airport (CYKF) (Figure 1, right). One of the three circuits simulated an emergency in the form of an engine failure. Order of circuit 2 and emergency were counterbalanced.

3.3 Procedure

Upon arrival, the participants filled out informed consent forms and completed a demographic questionnaire. Following this, they were equipped with wearable sensors, and a three-minute physiological baseline was recorded. Afterwards, a briefing took place where participants received information about the scenario and guidance on how to perform the tasks. Each participant's session was organized into five conditions: baseline, taxi, circuit 1, circuit 2, and emergency. The conditions were designed to reflect increasing task difficulty, ranging from baseline (lowest), taxi (second lowest) to emergency (highest). Circuit 1 and circuit 2 were anticipated to impose similar task difficulty. For each condition, a flight instructor seated next assessed performance. At the end of each condition, participants self-reported their stress. After the final condition, they completed a user experience questionnaire. Following which, sensors were removed and a brief debriefing held.

3.4 Outcome Measures

Data from 16 participants is reported here. While additional data types (electroencephalography, eye tracking, simulator-based performance metrics, and self-report of workload and situational awareness) were collected from a larger sample, analysis is ongoing.

Physiological Metrics.

Cardiovascular. Cardiovascular responses were recorded using a Polar H10 chest strap. The sensor contains built-in proprietary algorithms specifically designed to filter and calculate ECG parameters essential for Heart Rate Variability (HRV) analysis. The polar streams data at 1 Hz and calculates RR Intervals with a unit of 1/1024s [35]. A python toolbox for neurophysiological signal processing, NeuroKit2 [36], was used to analyze the HRV data. Extracted HRV time-domain measures include standard deviation of NN intervals (SDNN), root mean square of successive RR interval differences (RMSSD) and percentage of successive RR intervals that differ by more than 50 ms (pNN50). Moreover, non-linear analysis was conducted to capture the complexity and fractal properties of variability in the cardiac signal that may be missed by linear time- and frequency-domain analysis [37].

Electrodermal Activity. Electrodermal activity (EDA) was recorded using the Emotibit wristband sensor. The sensor streams data wirelessly from over 16 bio signals over any

Adafruit Feather-enabled protocol (Wi-Fi, Bluetooth, LoRA, etc.) at 15 Hz to a local computer connected to the same network [38]. EDA data preprocessing included noise removal using a low-pass Butterworth filter with a 0.3 Hz cutoff frequency, outlier removal and normalization. EDA data was decomposed into Skin conductance level (SCL) or tonic reflecting the overall arousal level during the session using a low-pass Butterworth filter with 0.05 Hz cutoff frequency.

Physiological Data Acquisition and Synchronization. Collection of multimodal physiological data and synchronizing them to conduct consolidated analysis is one of the biggest challenges in the field of psychophysiology and aviation training. A combination of third party and custom-built software clients was used to interface the various physiological sensors and stream the data using the lab streaming layer (LSL) communication protocol [39]. The Excite-O-Meter Devices[1] android application compatible with the chest strap sensor Polar H10 was used to send Heart Rate (HR) and Heart Rate Variability (HRV) data over LSL (see Fig. 2).

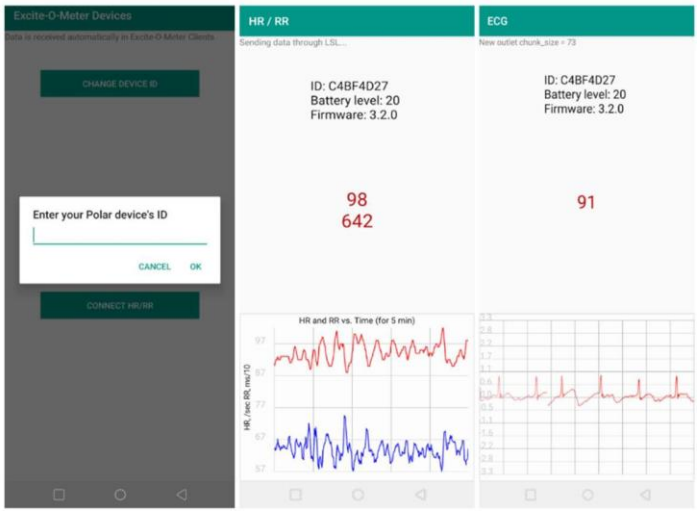


Fig. 2. Excite-O-Meter Devices android application.

The Emotibit Oscilloscope[2] (v1.11.4) was used to visualize and stream the various physiological metrics collected from the Emotibit device in real time via LSL (see Fig. 3).

[1]https://github.com/luisqtr/exciteometer/blob/main/docs/2_SetupDevices.md

[2] https://github.com/EmotiBit/ofxEmotiBit/releases



Fig. 3. Emotibit Oscilloscope Interface.

The Pupil Labs Companion LSL Relay [3] was used to stream gaze and event data from the Neon Companion Device to LSL (see Fig. 4).

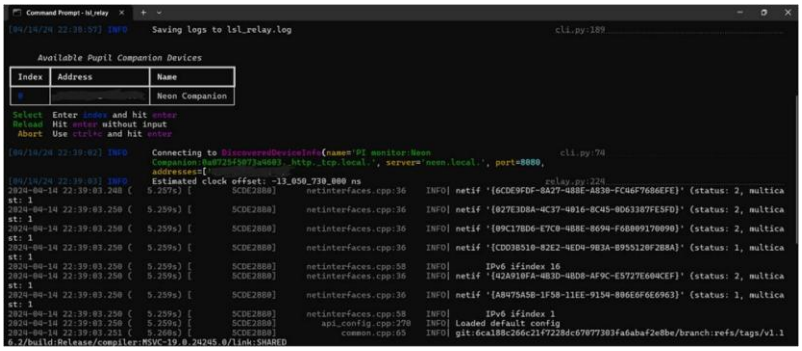


Fig. 4. LSL Relay for Neon.

The Petal Metrics desktop application [4] provides a Graphical User Interface (GUI) to connect to a Muse BCI headband and was used create an LSL outlet to stream the EEG data (see Fig. 5).

[3] https://pupil-invisible-lsl-relay.readthedocs.io/en/stable/index.html
[4] https://petal.tech/

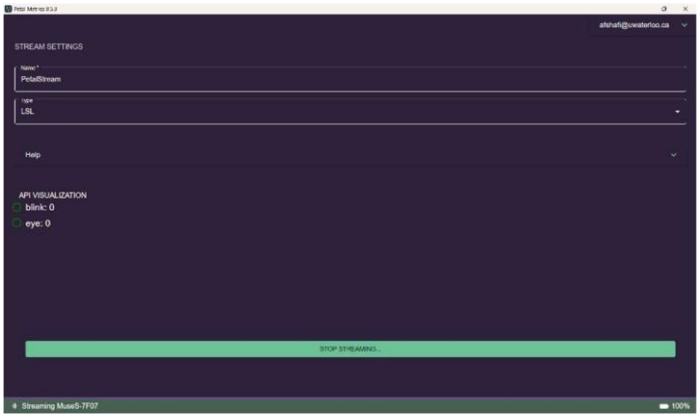


Fig. 5. Petal Metrics GUI.

Lab Recorder[5], an LSL recording application, was used to subscribe and synchronize all the data samples being pushed to LSL and store the data in a single file in the Extensible Data Format (XDF).

Flight Instructor Assessment. A flight instructor seated next to the pilot trainees assessed their performance using a modified version of an assessment tool created by Moncion et al. [40]. Based on Transport Canada's 4-point marking scale and assessment methodology used by their examiners during Private Pilot License - Aeroplane (PPL - A) flight test as well as ICAO Doc 9995, this tool allows for performance ratings across eight competencies. For this study, performance ratings across five competencies were calculated: Application of Procedures, Aircraft Flight Path Management – Manual, Problem Solving and Decision Making, Situational Awareness, and Workload Management. For each participant, the total score for each competency in each condition was first calculated and then converted to a mean score by dividing the total score by the number of items. These mean scores were then averaged across all participants to obtain a final mean for each competency in each condition.

User Experience and Stress. The Virtual Reality Neuroscience Questionnaire (VRNQ) was utilized to evaluate the quality of the flight simulation in two main aspects: user experience (UX) and VR-induced symptoms and effects (VRISE). Participants responded to five questions for each category using a 1 to 7 Likert scale with 1 representing extremely low immersion/absent VRISE and 7 extremely high immersion/intense VRISE [41]. We reverse-coded the VRISE portion of the VRNQ so 7 would be intense VRISE instead of 1. The VRNQ specifies cutoffs for sub-scores to determine adequate simulation quality. A minimum ≥ 25 and parsimonious ≥ 30 cutoff scores were applied to UX. We modified the cutoff for VRISE scores to maximum ≤ 15

[5] https://github.com/labstreaminglayer/App-LabRecorder

and parsimonious ≤ 10. Participants also reported their perceived stress levels on a 0 to 10 Likert scale with 0 representing minimal and 10 maximum stress.

4 Data Analysis

All analyses were conducted in Python. Data normality was assessed using the Shapiro-Wilk test. Friedman tests were applied for non-normally distributed data followed by Wilcoxon-signed rank test with Bonferroni correction. Results were significant at a corrected α level $p < .005$ for physiological and $p < .008$ for performance and self-report data, respectively.

5 Results

5.1 Physiological Responses

Cardiovascular Responses: Stress Regulation. Cardiac regulation responses were quantified using HR and HRV parameters. Fig. 6. shows a box plot illustrating the distribution of HR across the conditions. Friedman test revealed a significant effect of condition on HR, $\chi^2(4) = 39.09$, $p < .001$. Post-hoc Wilcoxon signed-rank tests showed HR increased significantly from baseline to all conditions: taxi ($Z = -3.41$, $p < .001$), circuit 1 ($Z = -3.35$, $p < .001$), circuit 2 ($Z = -3.35$, $p < .001$), and emergency ($Z = -3.41$, $p < .001$). Additionally, HR was significantly higher in emergency compared to taxi ($Z = -3.07$, $p < .001$), circuit 1 ($Z = -3.35$, $p < .001$) and circuit 2 ($Z = -3.24$, $p < .001$).

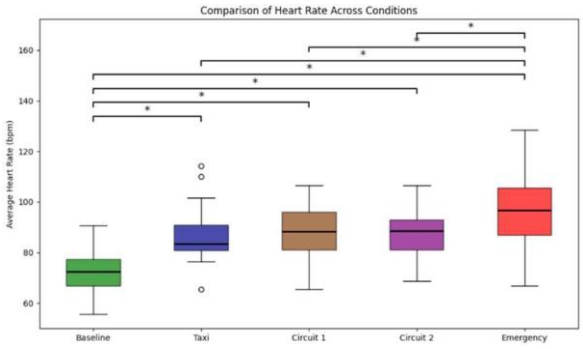


Fig. 6. Heart Rate boxplot across conditions. Asterisk (∗) denotes significant results following Wilcoxon tests.

Further HRV analysis involved computing time-domain metrics. The time-domain metrics were Mean NN, SDNN, RMSSD and pNN50. Freidman test revealed a significant effect of condition for all four metrics. Mean NN, $\chi^2(4) = 37.49$, $p < .001$; SDNN, $\chi^2(4) = 23.57$, $p < .001$; RMSSD, $\chi^2(4) = 33.97$, $p < .001$; and pNN50, $\chi^2(4) = 30.33$, $p < .001$. Post-hoc Wilcoxon signed-rank tests showed significant pairwise comparisons for each

of the metrics. Significant reductions occurred in Mean NN (Fig. 7., top-left) from baseline to other conditions: taxi (Z = -3.41, p < .001), circuit 1 (Z = -3.41, p < .001), circuit 2 (Z = -3.41, p < .001), and emergency (Z = -3.41, p < .001). Additionally, differences between emergency and other conditions: taxi (Z = -2.78, p = .003), circuit 1(Z = -3.01, p = .001), and circuit 2 (Z = -3.18, p < .001) were also significant. RMSSD values (Fig. 7., bottom-left) significantly declined from baseline to all conditions: taxi (Z = -3.35, p < .001), circuit 1 (Z = -3.35, p < .001), circuit 2 (Z = -3.41, p < .001), emergency (Z = -3.41, p < .001). A similar trend emerged for pNN50 (Fig. 7., bottom-right) from baseline to other conditions: taxi (Z = -3.41, p < .001), circuit 1 (Z = -3.35, p < .001), circuit 2 (Z = -3.41,p < .001) and emergency (Z = -3.35, p < .001). In contrast, differences in SDNN (Fig. 7., top-right) were significant only between baseline to circuit 1 (Z = -3.41, p < .001) and circuit 1 to emergency (Z = -3.24, p < .001).

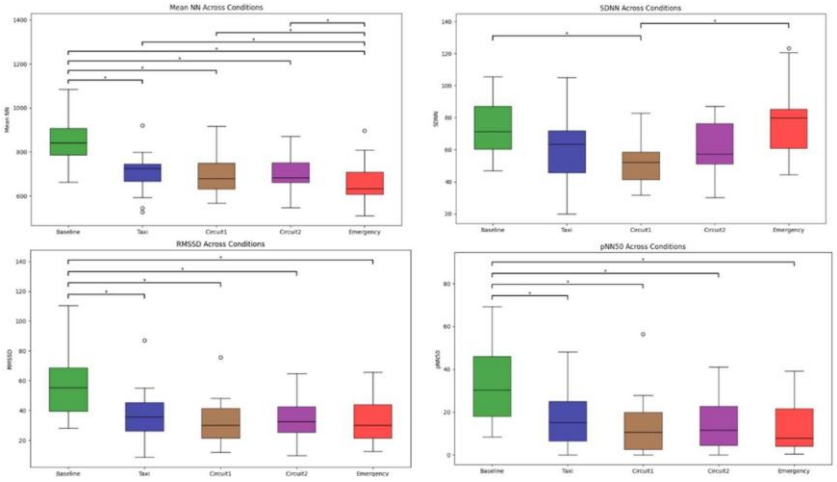


Fig. 7. Time-domain features of HRV across conditions: (Top-left) Mean NN, (top-right) SDNN, (bottom-left) RMSSD and (bottom-right) pNN50. Asterisk (∗) denotes significant results following Wilcoxon tests.

A Poincaré plot is a graphical means of visualizing and quantifying the evolution of dynamic systems, most represented on a cartesian coordinate space with x and y axes. In a Poincaré plot, consecutive RR intervals are plotted in pairs forming a time series (R-R1,R-R2) (R-R2,R-R3) (R-R3,R-R4)…..(R-Rn-1,R-Rn). Successive points with equal values fall along the 45° diagonal, known as the line of identity, where y=x. SD1 is the standard deviation of points perpendicular to the line of identity while SD2 is the standard deviation of points along the line of identity. SD1 and SD2 can be used to draw an ellipse which summarizes the shape of the Poincaré plot. SD1 reflects short term heart rate variability dominated mainly by parasympathetic activity and SD2 reflects long term heart rate variability or both sympathetic and parasympathetic activity. The SD1/SD2 ratio is a measure of autonomic balance [37]. Fig. 8. depicts visual patterns of the Poincaré plot that vary with stress. As seen in the Poincaré plot for

Participant 003, the shape of the ellipse gets increasingly narrower and points cluster near the identity line from baseline to emergency. The length of the SD1 (blue vector) also decreases with increasing stress and is shortest for the emergency condition consistent with parasympathetic or vagal withdrawal and sympathetic dominance.

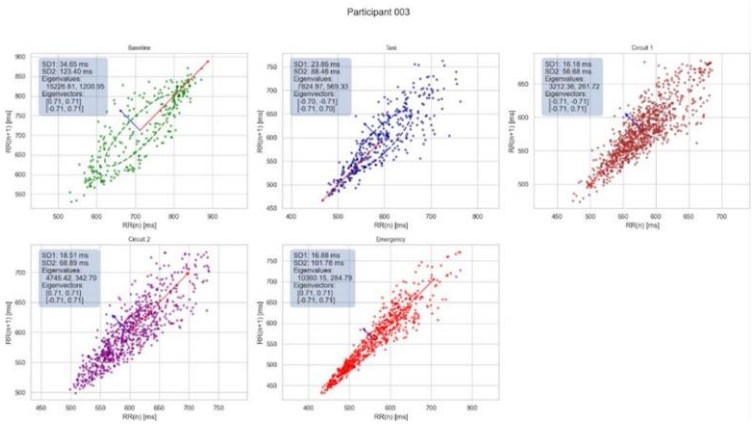


Fig. 8. Poincaré plot across conditions for Participant 003.

Statistical analysis of non-linear parameters revealed significant effect of condition: SD1($\chi^2(4) = 33.97$, $p < .001$), SD2 ($\chi^2(4) = 17.23$, $p = .002$) and SD1/SD2 ($\chi^2(4) = 26.29$, $p < .001$). SD1(Fig. 9., top-left) showed a general downward trend from baseline to emergency with post-hoc Wilcoxon-signed rank test difference being significant from baseline to all conditions: taxi ($Z = -3.35$, $p < .001$), circuit 1 ($Z = -3.35$, $p < .001$), circuit 2 ($Z = -3.41$, $p < .001$) and emergency ($Z = -3.41$, $p < .001$. SD2 baseline to circuit 1 ($Z = -3.12$, $p < .001$) and circuit 1 and emergency ($Z = -3.24$, $p < .001$) differences were significant (Fig. 9., top-right). SD1/SD2 ratio (Fig. 9., bottom-left) pairwise comparisons were significant for baseline to conditions: circuit 1 ($Z = -2.71$, $p = .004$), circuit 2 ($Z = -3.07$, $p = .001$), emergency ($Z = -3.41$, $p < .001$); taxi to emergency ($Z = -2.90$, $p = .002$), circuit 1 to emergency ($Z = -2.72$, $p = .004$) and circuit 2 to emergency ($Z = -2.84$, $p = .003$).

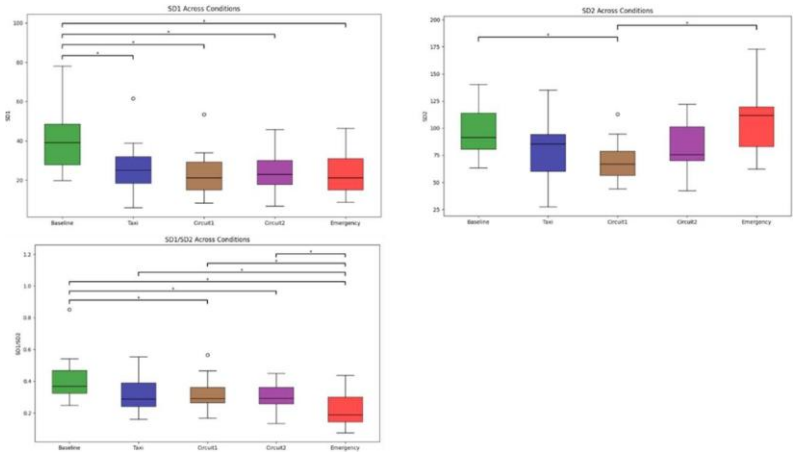


Fig. 9. Poincaré plot parameters analysis across conditions: SD1 (top-left), SD2 (top-right) and SD1/SD2 (bottom-left). Asterisk (∗) denotes significant results following Wilcoxon tests.

Electrodermal Activity: Physiological Arousal. EDA data was analyzed to infer physiological arousal levels. Due to technical challenges with the EmotiBit device, data from only eight participants was used for analysis. The line plot (see Fig. 10.) displays EDA patterns for Participant 004 across conditions showing the raw signal (top-left), clean signal (top-right), and tonic decomposition (bottom). Higher arousal responses were observed during conditions other than baseline, with taxi responses higher than circuit 1 and emergency. Additionally, the emergency triggered a visible increase in arousal after the emergency onset.

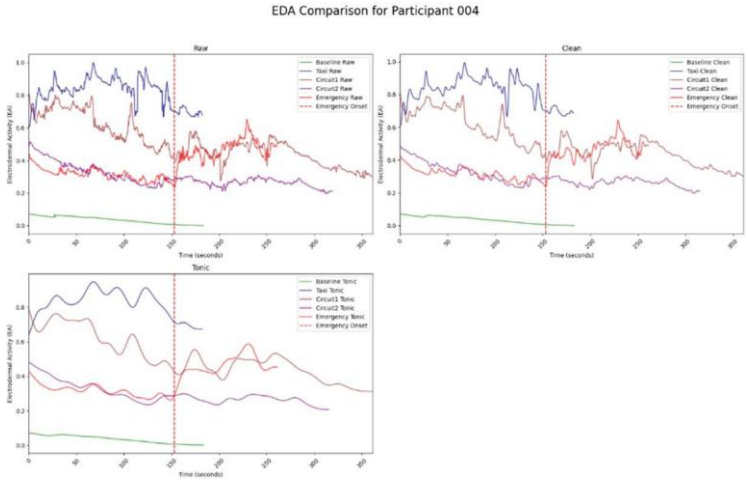


Fig. 10. EDA patterns for Participant 004 depicting baseline behavior and arousal responses to the scenario training in flight simulator.

The clean EDA signal was used for all analysis and visualization. The EDA heat map and radar plot below illustrate individual differences in arousal across conditions. Compared to the rest of the group, Participants 004 and 016 exhibited elevated responses in taxi, Participant 002 during circuit 2, and Participant 009 in emergency (Fig. 11., left). Taxi in general evoked moderate (e.g., Participant 001) to high responses in participants (e.g., Participants 004 and 016) (Fig. 11., right).

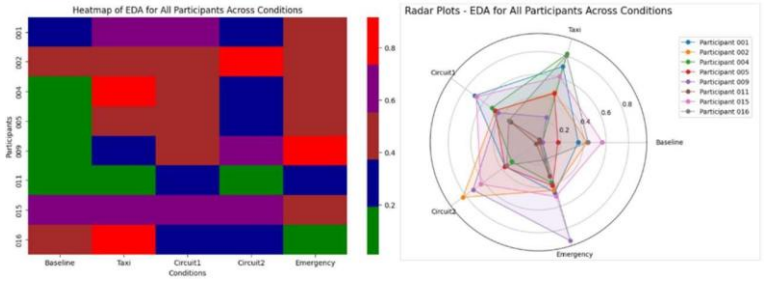


Fig. 11. EDA heat map (left) and radar plot (right) created to compare the arousal levels of participants across all conditions.

The emergency conditions were segmented into pre- and post-emergency for further comparison to see how arousal levels change during high-stress conditions (see Fig. 12.). The upward trend in EDA from baseline to post-emergency indicates an increase in physiological arousal, reflecting enhanced sympathetic activation occurring during the post-emergency phase.

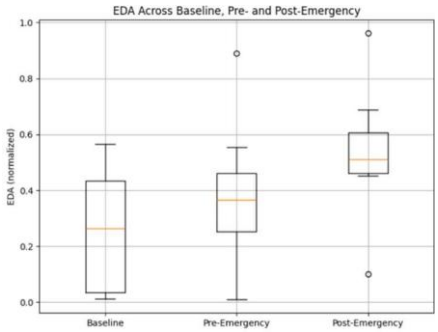


Fig. 12. Box plot comparing EDA across conditions (baseline, pre- and post-emergency)

5.2 Flight Instructor Assessment

The visualization on left (see Fig. 13.) compares the mean performance scores for all five competencies. Scores for four of the five competencies did not significantly differ across conditions. However, statistical analysis revealed a significant effect of condition on mean application of procedure scores (Fig. 13, right), $\chi^2(3) = 23.38$, $p < .001$. Post-hoc Wilcoxon-signed rank found significant differences for taxi and circuit 1 ($Z = -2.73$, $p = 0.006$), taxi and circuit 2 ($Z = -3.12$, $p = 0.002$), circuit 1 and emergency ($Z = -2.98$, $p = 0.003$) and circuit 2 and emergency ($Z = -3.18$, $p = 0.001$) (see Fig. 13., right). The average application of procedure scores were taxi (M = 3.44, SD = 0.35), circuit 1 (M = 3.84, SD = 0.24), circuit 2 (M = 3.91, SD = 0.20) and emergency (M = 3.11, SD = 0.71) out of total possible score of 4. The lowest competency scores were for situational awareness with average score being taxi (M = 2.88, SD = 0.72), circuit 1 (M = 3.25, SD = 0.68), circuit 2 (M = 3.56, SD = 0.63) and emergency (M = 3.00, SD = 0.55).

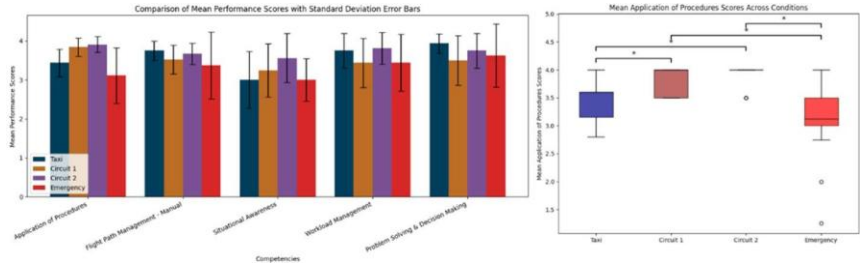


Fig. 13. Comparison of Mean Performance Scores with Standard Deviation Error Bars (left) and Mean Application of Procedures Scores Across Conditions (right). Asterisk (∗) denotes significant results following Wilcoxon tests.

5.3 Perceived User Experience and Stress

The visualization (Fig. 14., left) shows that the simulation met the minimum UX cutoff, with a median score of 27, whereas VRISE met the more stringent cutoff, with a median score of 6. Additionally, statistical analysis revealed a significant effect of condition on self-report of stress (see Fig. 14., right), $\chi^2(3) = 40.56$, $p < .001$. Post-hoc Wilcoxon-signed rank found significant difference for taxi and circuit 1 ($Z = -3.41$, $p = 0.001$), taxi and circuit 2 ($Z = -3.08$, $p = 0.002$), taxi and emergency ($Z = -3.52$, $p < .001$), circuit 1 and emergency ($Z = -3.52$,$p < .001$) and circuit 2 and emergency ($Z = -3.41$, $p = 0.001$).

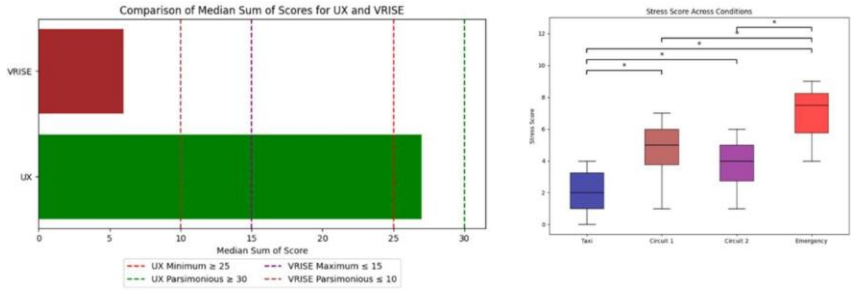


Fig. 14. Median Total Score for UX and VRISE (left) and Stress Score Across Conditions (right). Asterisk (∗) denotes significant results following Wilcoxon tests.

6 Discussion

6.1 Psychophysiological Responses and Autonomic Balance

Findings for HR and HRV time-domain measures (RMSSD, Mean NN and pNN50) are consistent with studies demonstrating that HR increases during high stress situations while time-domain measures generally decrease with increasing stress [4, 6, 10, 28, 29]. Non-linear analysis suggests an overall reduction in SD1 and a lower SD1/SD2 across conditions from baseline to emergency. Taken together, the findings offer support for the notion that participants experienced dynamic shifts in their "fight-and-flight" (sympathetic) and "rest-and-digest" (parasympathetic) activity [24] from baseline to emergency, with emergency eliciting the most pronounced autonomic imbalance. Additionally, contrary to our expectation, taxi and circuit 1 and taxi and circuit 2 HR differences were not significant. This likely reflects anticipatory stress related to upcoming task demand and performance evaluation. In terms of EDA, we could only use 56.25% of the data. With small sample sizes outliers can skew findings so we refrained from computing any statistical analysis. From the visualizations, however, the individual variability is evident. For some taxi had the highest EDA values, while for others emergency or circuit 1. Reinforcing these individual differences, studies have reported that not all participants demonstrate higher EDA values in stressful or emergency situations [1].

6.2 Performance

Consistent with findings of Cao et al. [4], Shao et al. [5], Peng et al. [6] and Lee [7], the results suggest that stress negatively impacts performance. For Emergency, flight instructors noted memory items not being completed, glide speed not maintained and radio calls not made given sufficient time. Failure to check instruments, brakes and aileron not being positioned into the wind during taxi. Additionally, taxi resulted in the second lowest mean application of procedure scores, consistent with HR findings. Furthermore, stress appears to impact certain competencies more than others. For instance, emergency application of procedure scores were significantly lower than other

conditions while situational awareness had the lowest performance scores of all competencies across conditions, with worst performance in taxi followed by emergency. One explanation could be that competencies such as flight path management largely rely on procedural skills. On the other hand, stress has been shown to impair situational awareness through its effect on higher-order cognition processes in the form of working memory decline, attentional narrowing, among others [42].

6.3 User Experience and Perceived Stress

Exposure to virtual environments, including flight simulators, can result in specific side effects, known as simulator sickness, which can encompass physiological symptoms reflecting changes in cardiovascular, respiratory or temperature regulation functions [43]. UX and VRISE can modulate stress responses making it imperative to rule out their confounding effects. UX and VRISE scores meeting minimum and parsimonious cutoffs, respectively, suggest that the simulator induced quality immersion and realism without any significant VRISE impacting the experience. Additionally, perceived stress differed significantly across conditions except for circuit 1 and circuit 2, while physiological measures did not consistently align with this finding, suggesting a partial dissociation between subjective and psychophysiological stress responses.

6.4 Sensors and Signal Quality

The scenario and the sensor employed in the study are certainly challenging for data quality. Because of the long duration of recordings (often 30+ minutes), we encountered short circuits in the EDA for the final circuit in several participants. Given within-subject design, we had to exclude any participants who had missing data in one of the conditions. Other researchers have also commented on challenges with EDA. In one study using Emotibit, only 2 out of 10 participants' EDA was usable [44], and in another study using the Empatica E4, only 18 out of 30 participants' EDA could be included in the analysis [3]. Heart rate and heart rate variability data from the Polar H10 turned out to be extremely stable, with no data loss or quality issues. The eye tracker and aviation headset worn over the Muse BCI produced noisy but manageable signals.

6.5 Future Work

While still preliminary, this study lays the groundwork for integrating multimodal wearable physiological sensing into flight training. Understanding stress responses can help flight instructors assess which maneuvers or phases of a flight may be more stressful or challenging. This can support stress inoculation training and personalized instructor assessment and feedback. Physiological insights interfaced with biofeedback technologies may also aid in acquisition of important self-regulation and stress-management skills. A secondary objective of the study is to create a database with the collected physiological, system metrics and user experience data for the development of computational models to characterize, simulate and predict pilot trainees' psychophysiological states, addressing the current paucity of such datasets [45-46]. This includes rule-based or machine-learning-based models to inform the design of adaptive training systems

that modulate flight simulation difficulty (for instance, by injecting instrument failures or environmental variable modifications through the instructor station interface) using interpreted psychophysiological indices or state.

7 Conclusion

This preliminary, incremental work, is one of the few studies using multimodal, fully wearable sensors to quantify and characterize pilot trainees' stress responses across a wide variety of training conditions. Results indicate initial evidence that the multimodal wearable physiological sensors employed in the study sensitively and consistently gauges changes in stress inducements. HR and HRV time-domain and non-linear metrics varied significantly across conditions, whereas EDA showed a non-significant trend of increased stress post-emergency compared to pre-emergency. Performance and self-report of stress in general corroborated psychophysiological insights. User experience scores further confirm this and the fidelity of the simulator and minimum VRISE indicate stress responses were highly likely to be task-induced. Envisioning future implications, we anticipate psychophysiological responses associated with skills training in interactive simulator programs beneficially complementing conventional training and yielding valuable insights to designing biocybernetic training systems.

Acknowledgement. This study was partially funded by a Natural Sciences and Engineering Research Council of Canada Discovery Grant (RGPIN-2024-04808 to S.C.). The authors express gratitude to the pilot trainees, flight instructors, and Waterloo Institute for Sustainable Aeronautics Research Coordinator Raphael Repato.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Guerrero, I., Vallès-Català, T.: Virtual reality flight simulation for pilot training: studying arousal levels during an emergency landing. In: INTED2023 Proceedings, pp. 2594–2599 (2023). https://doi.org/10.21125/inted.2023.0729
2. Vallès-Català, T., Guerrero, I.: Comparing arousal and workload during an emergency landing in a virtual reality and a conventional flight simulator. Int. J. Hum.-Comput. Interact. (2025). https://doi.org/10.1080/10447318.2025.2474464
3. Lutnyk, L., Rudi, D., Schinazi, V.R., Kiefer, P., Raubal, M.: The effect of flight phase on electrodermal activity and gaze behavior: a simulator study. Appl. Ergonom. 109, 1 (2023). https://doi.org/10.1016/j.apergo.2023.103989
4. Cao, X., et al.: Heart rate variability and performance of commercial airline pilots during flight simulations. Int. J. Environ. Res. Public Health 16(2), 237 (2019). https://doi.org/10.3390/ijerph16020237
5. Shao, S., Zhou, Q., Liu, Z.: A new assessment method of the pilot stress using ECG signals during complex special flight operation. IEEE Access 7, 185360–185368 (2019). https://doi.org/10.1109/ACCESS.2019.2959626

6. Peng, X., Niu, Q., Liang, Y., Luo, Y., Lu, N., Li, X.: Effects of unexpected event urgency and flight scenario familiarity on pilot trainees performance and stress responses. Front. Physiol. 16, 1599122 (2025). https://doi.org/10.3389/fphys.2025.1599122
7. Lee, K.: Effects of flight factors on pilot performance, workload, and stress at final approach to landing phase of flight. Electronic Theses and Dissertations, 2004–2019, 1628 (2010). https://stars.library.ucf.edu/etd/1628
8. Vallès-Català, T., Pedret, A., Ribes, D., Medina, D., Traveria, M.: Effects of stress on per formance during highly demanding tasks in student pilots. Int. J. Aerospace Psychol. 31(1), 43–55 (2021). https://doi.org/10.1080/24721840.2020.1841564
9. Regula, M., et al.: Study of heart rate as the main stress indicator in aircraft pilots. In: Pro ceedings of the 16th International Conference on Mechatronics – Mechatronika 2014, Brno, Czech Republic, pp. 639–643 (2014). https://doi.org/10.1109/MECHATRONIKA.2014.7018334
10. Wojciechowski, P., et al.: Evaluation of drone pilots' sympathetic and parasympathetic ner vous system responses during simulated flight in familiar and unfamiliar environments. In: 2024 IEEE International Workshop on Technologies for Defense and Security (TechDe- fense), Naples, Italy, pp. 7–12 (2024). https://doi.org/10.1109/TechDefense63521.2024.10863445
11. Bruna, O., Levora, T., Holub, J.: Stress measurement and inducement in experiments with low-cost flight simulator for testing of General Aviation pilots. Commun. Comput. Inf. Sci., 218–223 (2017). https://doi.org/10.1007/978-3-319-58750-9_31
12. McClernon, C.K., McCauley, M.E., O'Connor, P.E., Warm, J.S.: Stress training improves performance during a stressful flight. Hum. Factors 53(3), 207–218 (2011). https://doi.org/10.1177/0018720811405317
13. Cacioppo, J.T., Tassinary, L.G.: Psychophysiology and psychophysiological inference. In: Cacioppo, J.T., Tassinary, L.G. (eds.) Principles of Psychophysiology: Physical, Social, and Inferential Elements, pp. 3–33. Cambridge University Press, Cambridge (1990)
14. Kelley, C.R.: What is adaptive training? Hum. Factors 11(6), 547-556 (1969). https://doi.org/10.1177/001872086901100602
15. Caro, P.W.: Adaptive training—an application to flight simulation. Hum. Factors 11(6), 569– 576 (1969). https://doi.org/10.1177/001872086901100605
16. Fairclough, S.H.: Fundamentals of physiological computing. Interact. Comput. 21(1–2), 133– 145 (2009). https://doi.org/10.1016/j.intcom.2008.10.011
17. Palsson, O.S., Pope, A.T.: Thermal biofeedback: clinical applications and potential for pilot stress counter-response training. In: AIAA Paper 99-3501 (1999)
18. Schmorrow, D.D.: Aviation training: a future avenue. Avionics Mag., October 2005
19. Vartak, A.A., Fidopiastis, C.M., Nicholson, D.M., Mikhael, W.B., Schmorrow, D.: Cogni- tive state estimation for adaptive learning systems using wearable physiological sensors. In: Biosignals 2, pp. 147–152 (2008)
20. Wilson, J.C., Nair, S., Scielzo, S., Larson, E.C.: Objective measures of cognitive load using deep multi-modal learning: a use-case in aviation. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 5(1), 1–35 (2021)
21. Rissler, R., Nadj, M., Li, M.X., Knierim, M.T., Maedche, A.: "Got Flow?". In: Extended abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, R. Man- dryk, M. Hancock (eds.), pp. 1–6. ACM, New York (2018)
22. Guo, Z., Pan, Y., Zhao, G., Cao, S., Zhang, J.: Detection of driver vigilance level using EEG signals and driving contexts. IEEE Trans. Reliab. 67(1), 370–380 (2017). https://doi.org/10.1109/TR.2017.2778754

23. Sanei, S., Chambers, J.A.: EEG Signal Processing. John Wiley & Sons Ltd., Chichester (2007). https://doi.org/10.1002/9780470511923
24. Cacioppo, J.T., Tassinary, L.G., Berntson, G.G. (eds.): Handbook of Psychophysiology, 3rd, ed. Cambridge University Press, Cambridge (2007). https://doi.org/10.1017/CBO9780511546396
25. Kim, H.G., Cheon, E.J., Bai, D.S., Lee, Y.H., Koo, B.H.: Stress and heart rate variability: a meta-analysis and review of the literature. Psychiatry Investig. 15(3), 235–245 (2018). https://doi.org/10.30773/pi.2017.08.17
26. Ghiasi, S., Greco, A., Barbieri, R., et al.: Assessing autonomic function from electrodermal activity and heart rate variability during cold-pressor test and emotional challenge. Sci. Rep. 10, 5406 (2020). https://doi.org/10.1038/s41598-020-62225-2
27. Liang, Y., et al.: Effect of acute stress on working memory in pilots: investigating the modu latory role of memory load. PLoS ONE 19(1), e0288221 (2024). https://doi.org/10.1371/journal.pone.0288221
28. Wang, Y., Fan, X., Shao, S., Fu, C., Wang, Y.: Research on the psychological stress of pilots under complex flight interaction tasks. In: 2nd International Conference on Artificial Intelli gence, Human-Computer Interaction and Robotics (AIHCIR), Tianjin, China, pp. 1–5 (2023). https://doi.org/10.1109/AIHCIR61661.2023.00007
29. Bustamante-Sánchez, Á., Clemente-Suárez, V.J.: Psychophysiological response in night and instrument helicopter flights. Ergonomics 63(4), 399-406 (2020). https://doi.org/10.1080/00140139.2020.1718772
30. Causse, M., Dehais, F., Faaland, P., Cauchard, F.: An analysis of mental workload and psychological stress in pilots during actual flight using heart rate and subjective measurements (2012). In: 5th International Conference on Research in Air Transportation (ICRAT 2012), 22 May 2012- 25 May 2012 (Berkeley, United States).
31. Bruna, O., Levora, T., Holub, J.: Assessment of ECG and respiration recordings from simu lated emergency landings of ultra-light aircraft. Sci. Rep. 8, 7232 (2018). https://doi.org/10.1038/s41598-018-25528-z
32. Villafaina, S., Fuentes-García, D.J.P., Gusi, N., Tornero-Aguilera, J.F., Clemente-Suárez, V.J.: Psychophysiological response of military pilots in different combat flight maneuvers in a flight simulator. Physiol. Behav. 238, 113483 (2021). https://doi.org/10.1016/j.physbeh.2021.113483
33. Ruseva, B.K., Vezenkov, S.R., Tonchev, P.T., Mihaylov, I.I., Manolova, V.R., Radoyski, U.G.: Comparative analysis of heart rate variability in aviator cadets and instructors during ground training. Arch. Balk. Med. Union 60(1), 21-31 (2025). https://doi.org/10.31688/ABMU.2025.60.1.02
34. ALSIM: AL250 Instructor Station Manual (2019)
35. Polar Electro Oy: Polar H10 Heart Rate Sensor. Polar. https://www.polar.com/ca-en/sensors/h10-heart-rate-sensor
36. Makowski, D., et al.: NeuroKit2: a Python toolbox for neurophysiological signal processing. Behav. Res. Methods 53(4), 1689–1696 (2021). https://doi.org/10.3758/s13428-020-01516-y
37. Shaffer, F., Ginsberg, J.P.: An overview of heart rate variability metrics and norms. Front. Public Health 5, 258 (2017). https://doi.org/10.3389/fpubh.2017.00258
38. Montgomery, S.M., Nair, N., Chen, P., Dikker, S.: Validating EmotiBit: an open-source multi modal sensor for capturing research-grade physiological signals from anywhere on the body. Measur. Sens. 32, 101075 (2024). https://doi.org/10.1016/j.measen.2024.101075
39. Kothe, C., et al.: The lab streaming layer for synchronized multimodal recording. Imaging Neurosci. 3 (2025). https://doi.org/10.1162/IMAG.a.136

40. Moncion, B., Cao, S., Kearns, S.: Creating competency-based assessment grade sheets and a rubric for private pilot license training. Waterloo Institute for Sustainable Aeronautics, University of Waterloo, Report 2024-001 (2024)
41. Kourtesis, P., Collina, S., Doumas, L.A.A., MacPherson, S.E.: Validation of the virtual reality neuroscience questionnaire: maximum duration of immersive virtual reality sessions without the presence of pertinent adverse symptomatology. Front. Hum. Neurosci. 13, 417 (2019). https://doi.org/10.3389/fnhum.2019.00417
42. Matthews, G., Wohleber, R.W., Lin, J.: Stress, skilled performance, and expertise: overload and beyond. In: Ward, P., Schraagen, J.M., Gore, J., Roth, E. (eds.) The Oxford Handbook of Expertise, pp. 490–524. Oxford University Press, Oxford (2019). https://doi.org/10.1093/oxfordhb/9780198795872.013.22
43. Polak, E., Ślugaj, R., Gardzińska, A.: Postural control and psychophysical state following a flight simulator session in novice pilots. Front. Public Health 10, 788612 (2022). https://doi.org/10.3389/fpubh.2022.788612
44. Muñoz, J.E., Lavoie, J.A., Pope, A.T.: Psychophysiological insights and user perspectives: enhancing police de-escalation skills through full-body VR training. Front. Psychol. 15, 1390677 (2024). https://doi.org/10.3389/fpsyg.2024.1390677
45. Xu, R., et al.: An in-flight multimodal data collection method for assessing pilot cognitive states and performance in general aviation. MethodsX 15, 103589 (2025). https://doi.org/10.1016/j.mex.2025.103589
46. Fettrow, T., et al.: Human contributions to safety data testbed flight simulation study: data methods, processing, and quality. Sci. Data 12, 1247 (2025). https://doi.org/10.1038/s41597-025-05336-7