FT2025
SWEDISH AEROSPACE AT THE CROSSROADS
The 12th Swedish Aerospace Technology Congress
FT2025 in Stockholm
October 14-15, 2025

Impact of HUD Design on Cognitive Load in UAV Control

Marina Ronconi de Oliveira[1], Ivan de Souza Rehder[1], Larissa Takei[1], Emilia Villani[1], and Moacyr Machado Cardoso Junior[1]

[1]*Aeronautics Institute of Technology/ CCM-ITA, Brazil*

E-mail:, marina.oliveira@ccm-ita.org.br, ivan.rehder@ccm-ita.org.br, larissa.takei@ccm-ita.org.br, emilia.villani@ccm-ita.org.br, moacyr.cardoso@gp.ita.br.

Abstract

This study analyzes the influence of mental workload on physiology of 24 individuals with flight deck experience during a remote control of an uncrewed aerial vehicle (UAV). The UAV operates without line-of-sight communication, so any command given by the pilot has a two-second delay due to the satellite link. To assist the pilot, two flight data display interfaces (HUDs) were developed. During the simulation, physiological responses were monitored using electrodermal activity (EDA) sensors, electrocardiograms (ECG) and eye tracking. Besides the physiological data, the research also uses subjective workload assessments such as the NASA Task Load Index (NASA-TLX), the Subjective Workload Dominance technique (SWORD), and the Instantaneous Self-Assessment (ISA). Results indicated that designed HUDs influenced cognitive load and flight accuracy leading to lower workload and performance improvement. These results highlight the need for more adaptable interfaces. As a future perspective, the usage of adaptive operator support systems is recommended, adjusting interfaces and automation levels according to the user's cognitive state, enabling a more efficient and safer interaction with advanced aviation systems.

Keywords: flight deck, HMI, mental workload, physiological monitoring

1 Introduction

Aviation safety has evolved over the decades, together with technological and operational advances in the sector. However, human error remains one of the causes of aviation accidents, despite the global reduction in incident rates [1]. Pilots' mental workload stands out among the factors contributing to such errors as it can make decision-making and response capacity both harder in critical situations [2, 3]. In this context, understanding pilot physiology becomes essential, particularly in light of the growing challenges imposed by highly automated modern flight decks.

With onboard automation, new cognitive demands are placed on operators. Although automation has benefits such as reducing repetitive tasks and increasing operational efficiency, it can also bring cognitive overload, especially in high-complexity or emergency scenarios in which the pilot must take over control of the aircraft [4, 5]. The Human-Machine Interface (HMI) in these situations requires attention to the pilot's cognitive state, in order to maintain their ability to respond at critical moments.

In this scenario, monitoring pilot physiology becomes a strategy to detect signs of mental overload in real time. The use of physiological sensors such as Electrocardiogram (ECG), Electrodermal Activity (EDA) and eye tracking has proven effective in inferring the operators' mental workload [6, 7]. Integrating these into adaptive systems could be a significant advancement in the design of future cockpits, making them more agile and personalised responses to pilots' cognitive demands [8].

Based on this premise, the present study aimed to evaluate methods and tools for understanding pilot physiology and mental workload, in order to serve as input for future flight deck designs. To this end, a controlled experiment involving pilots, using physiological sensors and subjective workload assessment scales such as NASA Task Load Index (NASA-TLX) [9], Subjective Workload Dominance (SWORD) [10], and Instantaneous Self-Assessment (ISA) was carried out.

The structure of this paper is as follows: Section 2 presents the related work; Section 3 describes the materials and methods; Section 4 discusses the experimental results; and Section

5 provides conclusions and recommendations for future work.

2 Related Work

Research on pilot physiology has increasingly focused on psychophysiological monitoring as a tool to assess mental workload and enhance operational safety. Studies conducted by Brazilian aerospace institutions such as the Department of Science and Aerospace Technology (*Departamento de Ciência e Tecnologia Aeroespacial*) (DCTA) and the Flight Testing and Research Institute (*Instituto de Pesquisa e Ensaios em Voo*) (IPEV) have explored the real-time monitoring of military pilots using advanced physiological sensors, including Electroencephalogram (EEG), Heart Rate (HR), and eye-tracking devices [11]. These investigations aim to detect early signs of cognitive overload or incapacitation, propose the development of adaptive alert systems, and suggest refinements in pilot selection and training processes to increase resilience under high workload conditions.

Further studies have analysed the use of physiological and subjective tools to assess workload in simulated and operational environments. For instance, investigations with ECG and Galvanic Skin Response (GSR) sensors during day and night missions using Night Vision Goggles (NVG) [12] found minimal workload variation between conditions, but emphasised the influence of pilot fatigue and rest routines. In Uncrewed Aerial Vehicle (UAV) operation contexts [13], eye-tracking combined with NASA-TLX and usability scales has shown that more experienced operators exhibit lower cognitive strain, with strong correlations between pupil dilation and perceived workload. Similarly, research with fighter pilots during IFR simulator tests demonstrated that Heart Rate Variability (HRV) is sensitive to task complexity, even when performance scores remain constant.

Finally, real-flight studies [14] with light aircraft have indicated that ocular metrics, especially saccadic rate, are more reliable than cardiac indicators for distinguishing between workload levels during different flight phases. These findings reinforce the importance of integrating physiological monitoring into the cockpit environment. Collectively, the reviewed works support the implementation of real-time, multimodal monitoring systems as a foundation for safer and more adaptive flight deck designs.

3 Materials and Methods

This study presents an experimental investigation of pilot mental workload in simulated flight scenarios, aiming to understand how different cockpit interface configurations—particularly involving the use of a Heads-Up Display (HUD)—affect cognitive load. Subjective and physiological metrics were employed to enable a multidimensional assessment of the participants' mental effort. The research was conducted at the Aeronautical Institute of Technology (*Instituto Tecnológico de Aeronáutica*) (ITA), within the Competence Center of Manufacturing (*Centro de Competência em Manufatura*) (CCM).

3.1 The Experiment

During a simulated flight between two cities, a critical failure required an emergency landing at an unmapped airfield. In this scenario, the pilot was unable to remotely configure the landing parameters and had to take manual control of the aircraft. A two-second delay in satellite communication — between the pilot's command inputs and the aircraft's response — added further complexity to the task.

3.2 Participants

The experiment involved 24 volunteer participants with professional flight deck experience who acted as pilots in simulated flight scenarios. All participants were informed about the objectives of the study and provided written informed consent by signing the Informed Consent Form (ICF), in accordance with ethical principles for research involving human subjects (Ethics Submission Registration Number: CAAE 77429824.7.0000.5503). Prior to the experimental sessions, participants underwent initial training to become familiar with the simulation environment and the equipment used.

3.3 Experimental scenarios

This experiment is part of a broader initiative called the Air Domain Systems (ADS), structured according to a design–test–analysis protocol. The evaluations were carried out using the ADS Simulator (Figure 1 [15], a computational platform developed to replicate aircraft behavior in a controlled flight environment.

Figure 1: The ADS simulator computer setup

A real-time flight test was conducted in a Simulink®–FlightGear® environment, with the cockpit graphics rendered via Unity®. Each session involved a visual approach to São José dos Campos airport, starting from cruise altitude and incorporating a fixed 2s communication latency. The only manipulated factor was the HUD configuration. Flight durations averaged 6–7 minutes, and the sequence of the three HUD conditions was assigned randomly for each participant to prevent practice effects.

- No HUD (Control condition): Only the Primary Flight Display was visible to the pilots, experiencing the same

2 seconds delay on all critical flight parameters, such as heading, airspeed, altitude and pitch angle.

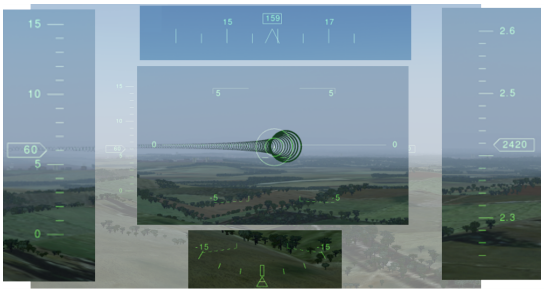

Figure 2: No HUD.

- HUD 1.0 (Numerical cues): In this first head-up display setup, the forward-looking view was augmented with numeric readouts of bank angle and roll rate, numeric readouts of pitch angle and pitch-rate and a delay-compensated trajectory prediction, depicted as a series of blue squares, computed by the onboard flight-dynamics model.

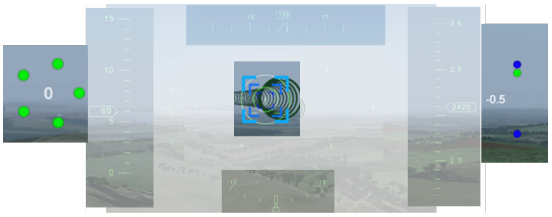

Figure 3: HUD 1.0.

- HUD 2.0 (Graphical and mode-based display): Building on HUD 1.0, the enhanced version added a graphical tapes for bank angle and roll rate, a graphical tapes for pitch angle and pitch-rate, a predictive flight-path indicator shown as a magenta "flight-path ball" (using the same compensation algorithm), a chevron-style throttle status indicator and two selectable modes—Cruise and Landing—tailored to different flight phases.

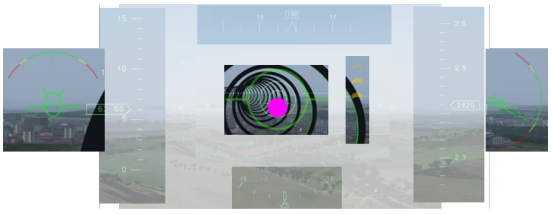

Figure 4: HUD 2.0.

3.4 Equipments and instrumentation

The experimental platform was instrumented with both objective physiological sensors and subjective assessment tools. Physiological data were acquired via the CAPTIV system (TEA), which continuously recorded EDA, peripheral temperature (TEMP) and an ECG. Concurrently, participants wore Tobii Pro Glasses 2 for mobile eye tracking. This system captured high-resolution measurements of pupil diameter and fixation patterns across both cockpit instruments and the external scene.

Immediately following each predefined flight segment, participants used the ISA to record their perceived cognitive workload in real time. After completing each full scenario, they filled out the NASA-TLX, which evaluates six dimensions of workload—mental demand, physical demand, temporal demand, performance, effort, and frustration—via a multidimensional rating scale. Once all three scenarios were finished, pilots applied the SWORD method to perform pairwise comparisons among the HUD conditions and establish a ranking of relative subjective workload.

In addition to subjective assessments, objective performance was evaluated throughout each flight based on the deviation from the center of predefined target areas (green circles). Participants performed better when their control inputs kept the aircraft closer to the center.

3.5 Data Analysis

3.5.1 Objective physiological measurements

For the EDA signals, a low-pass filter with a cutoff frequency of 5 Hz was initially applied to eliminate high-frequency components and preserve the phasic component of the signal. From the filtered signal, metrics such as the number of identified peaks—representing discrete physiological events—and the sum of the amplitudes of these peaks—reflecting the overall intensity of the autonomic response—were extracted. For both variables, the percentage difference relative to each participant's individual baseline was calculated.

In the case of the ECG, processing began with the identification of R-wave peaks in the pre-filtered signal. Based on this detection, RR intervals—corresponding to the time between successive heartbeats—were computed. These intervals were then converted into NN intervals by removing non-physiological outliers. From the valid NN intervals, the following Heart Rate Variability (HRV) metrics were extracted: HR, the Standard Deviation of NN Intervals (SDNN), the Root Mean Square of Successive Differences (RMSSD), and the Low-frequency to High-frequency Ratio (LF/HF). As with the EDA metrics, the percentage change from the individual baseline was calculated for each of these parameters, allowing for the assessment of relative variations in physiological state across participants.

Regarding the processing of the data obtained through the eye tracker, the values of the right and left pupil diameter were extracted over time within the selected period. The average pupil diameter was calculated by taking the arithmetic mean of both eyes and subsequently averaging it over the defined time window. The percentage difference relative to each participant's individual baseline was then computed for this average. Additionally, heat maps were generated to visually indicate the areas of the screen that received the most visual attention, and gaze plots were created to represent the temporal sequence of fixation points, enabling the analysis of visual

exploration patterns during the task.

3.5.2 Subjective workload assessments

The NASA-TLX was administered at the end of each simulation scenario and consisted of two main stages. In the first stage, participants performed pairwise comparisons among the six workload dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration), indicating which aspects were more relevant to their perceived cognitive effort. The number of times each dimension was selected during the comparisons determined its relative weight. In the second stage, participants rated each dimension on a scale from 1 to 10, reflecting the perceived intensity of workload for that particular aspect. Using the weights and the ratings, a weighted average was calculated to obtain the final NASA-TLX score.

The ISA was applied continuously during the simulated flights. Approximately every 30 seconds, a visual signal prompted participants to report their current level of mental workload using a scale from 1 (minimal) to 5 (maximum). All responses recorded during each scenario were used to compute an average score per participant and per scenario.

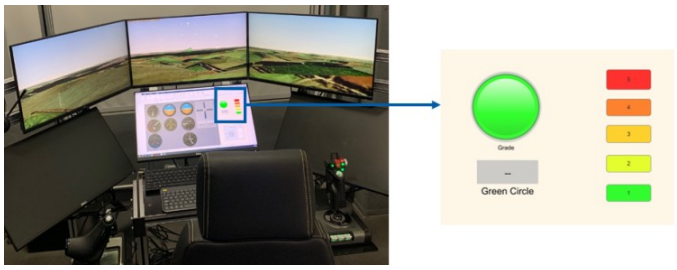

Figure 5: ISA section.

The SWORD method was administered at the end of all simulation trials to directly compare the different interface configurations. Participants made pairwise comparisons between the three interface conditions (No HUD, HUD 1.0, and HUD 2.0), assigning a value on a workload dominance scale for each pair. This scale ranged from 1 (equal workload) to 9 (absolute dominance), according to each participant's perception. The assigned values were organized into a judgment matrix, which was then column-normalized. The mean of each row produced the principal eigenvector, representing the relative weights of mental workload for each task. Subsequently, the eigenvalue of the matrix was computed, and a consistency check was performed to assess the logical coherence of the comparisons.

3.5.3 Statistical Analysis

To compare the effect of the three HUD conditions on each dependent variable, a blocked Analysis of Variance (ANOVA) was used, considering each participant as a blocking factor to control for individual variations. Prior to the test, the assumptions of normality of residuals and homogeneity of variances were verified using the Shapiro-Wilk and Levene's tests, respectively.

For variables that did not meet the ANOVA assumptions even after data transformation attempts (e.g., logarithmic), the equivalent non-parametric test, the Friedman Test, was employed. In all cases, a significance level of $\alpha = 0.05$ was adopted. When a significant main effect of the treatment was found, post-hoc tests such as Tukey's HSD (for ANOVA) or non-parametric multiple comparison tests were conducted to identify specific pairwise differences.

Additionally, to investigate the intrinsic relationship between the different workload and physiological metrics, a Spearman's rank correlation matrix was calculated.

4 Results and discussions

The results presented in this chapter are derived from an experimental approach whose primary objective was to identify which of the evaluated HUD interfaces imposes the lowest mental workload on participants.

4.1 NASA-TLX

The NASA-TLX results are summarized in Figure 6 and Table 1. In these visualizations, red represents the scenario with the highest mental workload, yellow indicates an intermediate level, and green corresponds to the lowest workload. The "No HUD" scenario presented the highest average mental workload, followed by "HUD 1.0", while "HUD 2.0" had the lowest workload values. The data also indicate greater variability in participants' responses for the "No HUD" condition, suggesting divergent perceptions, whereas "HUD 2.0" results were more consistent. Overall, the findings suggest that HUD 2.0 was the most cognitively efficient interface among the three evaluated.

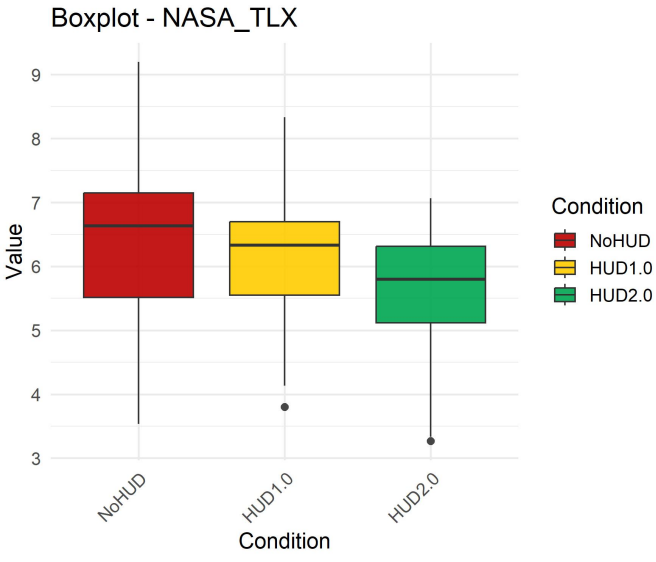

Figure 6: NASA-TLX Results.

Condition	Average NASA-TLX Score
No HUD	6.63
HUD 1.0	6.33
HUD 2.0	5.80

Table 1: Average NASA-TLX scores for each scenario.

To assess the effect of the HUD conditions on the NASA-TLX score, a blocked ANOVA was conducted, with each participant treated as a block. The assumptions for the test were met. The analysis revealed a statistically significant effect of the HUD condition ($F(2,46) = 4.54, p = 0.016$), as detailed in Table 2.

A post-hoc Tukey's HSD test was used for pairwise comparisons. The results, shown in Table 3, indicate that the workload in the `No HUD` condition was significantly higher than in the `HUD 2.0` condition ($p \approx 0.05$).

Table 2: Analysis of Variance (ANOVA) Results for the NASA-TLX Variable.

	sum_sq	df	F	PR(>F)
Condition	8.79	2.0	4.54	0.016
ID	56.72	23.0	2.55	0.003
Residual	44.52	46.0		

Table 3: Pairwise comparison results (Tukey HSD) for the NASA-TLX Variable.

group1	group2	p-adj	reject
HUD1.0	HUD2.0	0.29	False
HUD1.0	NoHUD	0.62	False
HUD2.0	NoHUD	0.05	True

4.2 ISA

Figure 7 presents the ISA results using a boxplot. As in the previous figure, the color scheme follows the same pattern: red represents the scenario with the highest perceived mental workload, yellow indicates an intermediate level, and green corresponds to the lowest workload. The average scores for each experimental condition are shown in Table 4. Since the ISA scale ranges from 1 to 5 — with higher values indicating greater cognitive effort — the data reveal that the HUD 2.0 condition required the least mental workload from participants. This was followed by HUD 1.0, which showed a moderate level of workload, while the No HUD condition was the most cognitively demanding. Despite these differences in average scores, the distributions across the three scenarios were relatively similar, suggesting that changes in HUD configuration had only a limited effect on the subjective perception of mental workload during task execution.

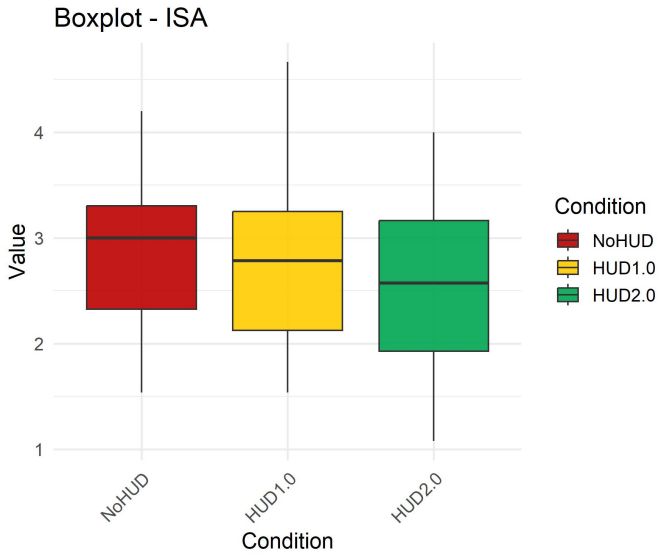

Figure 7: ISA Results.

Condition	Average ISA Score
No HUD	2.87
HUD 1.0	2.79
HUD 2.0	2.57

Table 4: ISA Average.

The effect of the HUD conditions on the ISA score was evaluated using a blocked ANOVA, as its assumptions were met. A significant main effect was found ($F(2,46) = 3.41, p = 0.041$), suggesting an overall difference among the conditions 5. However, the post-hoc Tukey's HSD test did not have sufficient statistical power to identify significant differences between any specific pair of conditions (see Table 6). This discrepancy may occur when the p-value from the ANOVA is close to the significance threshold or when there is high variability in the data, which limits the statistical power of the post-hoc test.

Table 5: ANOVA Results for the ISA Variable.

	sum_sq	df	F	PR(>F)
Condition	1.21	2.00	3.41	0.04
ID	31.34	23.00	7.68	0.00
Residual	8.16	46.00		

Table 6: Pairwise comparison results (Tukey HSD) for the variable ISA.

group1	group2	p-adj	reject
HUD1.0	HUD2.0	0.56	False
HUD1.0	NoHUD	0.90	False
HUD2.0	NoHUD	0.34	False

4.3 SWORD

The analysis of the data from the SWORD questionnaire was conducted based on the percentage evaluation of the scenarios identified by participants as those requiring the least and the highest mental workload, as illustrated in Figures 8 and 9, respectively.

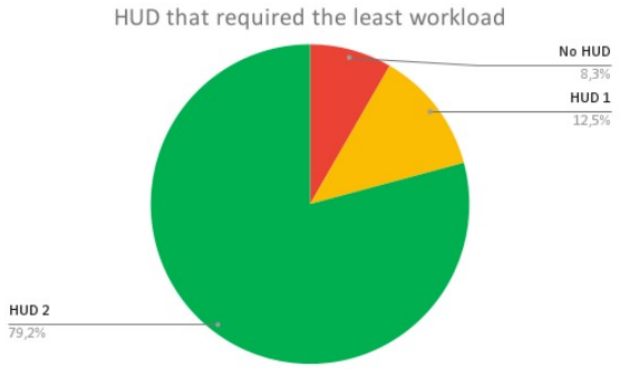

Figure 8: HUD that required the least mental workload.

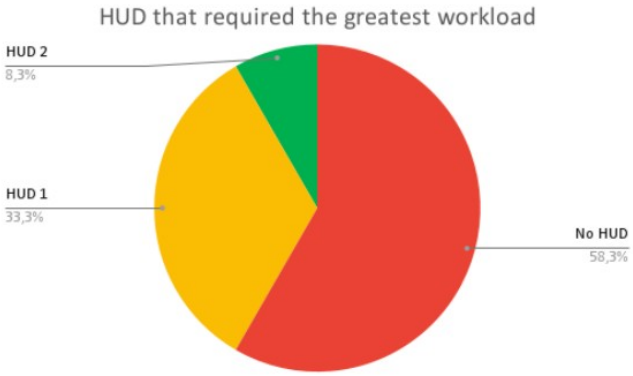

Figure 9: HUD that required the greatest mental workload.

Figure 8 shows that the majority of participants (79.2%) indicated the HUD 2.0 scenario as the least cognitively demanding. In contrast, only 8.3% of participants perceived the No HUD scenario in the same way, suggesting that it was generally considered more mentally taxing.

Figure 9 reveals that the No HUD scenario was perceived as the most mentally demanding by 58.3% of participants, followed by HUD 1.0 with 33.3%, while only 8.3% attributed the highest workload to HUD 2.0.

The data for the SWORD scores violated the assumptions for parametric testing. Therefore, the non-parametric equivalent, the Friedman Test, was used to compare the three HUD conditions. The test revealed a highly significant difference in perceived workload among the scenarios ($\chi^2(2) = 17.70, p < 0.001$).

4.4 Performance

The radius of the green circles was fixed at 20 meters, and the average distances were calculated based on displacements along the x and y axes for each experimental scenario, reflecting how closely participants maintained their trajectory relative to the center of the target areas during the flight.

The average distances to the center obtained for each condition are presented in Table 7, while Figure 10 displays the data using a boxplot. As in the previous figures, the colors follow the same pattern for representing mental workload.

Condition	Distance Average
No HUD	5.42
HUD 1.0	5.18
HUD 2.0	5.03

Table 7: Average distance from the center for each scenario.

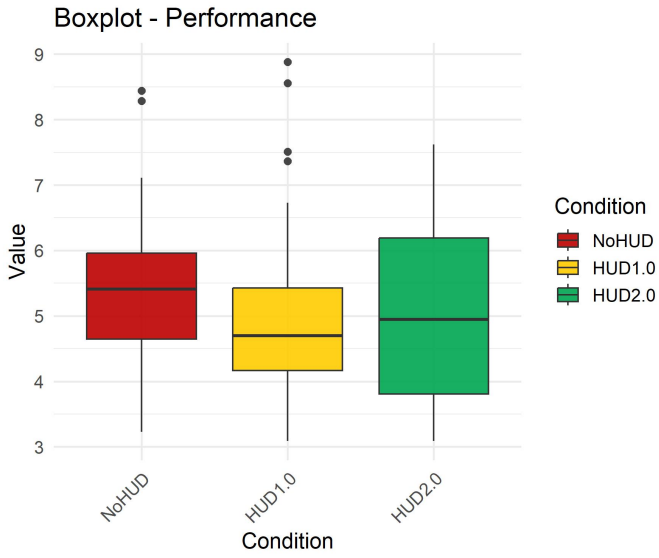

Figure 10: Performance.

The results show that the HUD 2.0 scenario yielded the best average performance, that is, the smallest deviations from the center, followed by HUD 1.0. The No HUD condition exhibited the largest deviations, reflecting the lowest accuracy among the three.

Despite the average in the HUD 2.0 scenario meaning a lighter workload, greater data dispersion was observed, indicating increased variability in performance among participants. This pattern suggests that while the HUD 2.0 interface contributed positively to most individuals' performance, some participants experienced specific difficulties during the task, which may be attributed to individual differences in adapting to the interface.

The raw performance data (distance from the center) did not meet the assumptions for a standard ANOVA. A logarithmic transformation was successfully applied to the data to satisfy these assumptions. Subsequently, a blocked ANOVA was performed on the transformed data. The analysis found no statistically significant effect of the HUD conditions on flight performance ($F(2,46) = 0.83, p = 0.44$).

4.5 EDA

For the analysis of the data obtained through the EDA sensor, three parameters related to the phasic component of each participant's signal were considered: the number of peaks detected (Num_peaks), the sum of signal values (Sum_data), and the sum of the amplitudes of these peaks (Sum_amp). Subsequently, the percentage difference of each parameter relative to the individual baseline of each participant was calculated, as described in Equation 1, with the aim of standardizing physiological responses and enabling comparisons across experimental conditions.

$$\text{Percentage Difference} = \left(\frac{\text{Scenario Average} - \text{Baseline Average}}{\text{Baseline Average}} \right) \times 100\% \quad (1)$$

To facilitate the interpretation of results, boxplots were constructed (Figures 11 to 13), following the same color scheme used in the other analyses.

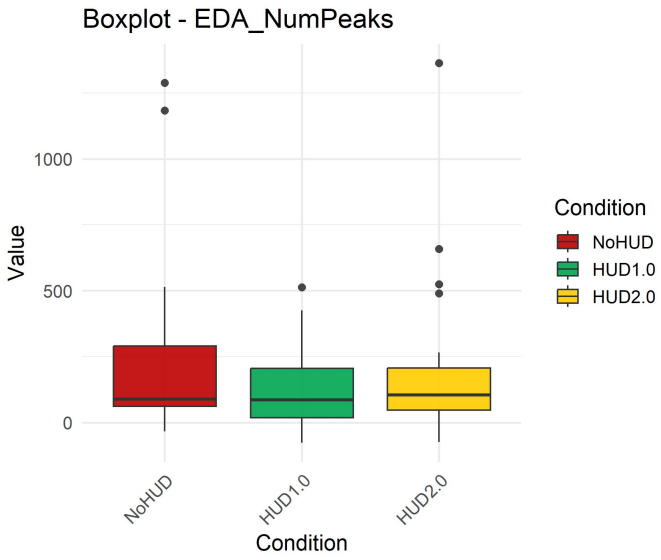

Figure 11: Number of peaks with outliers.

During the analysis, some outliers were observed. This may be attributed to the high sensitivity of the EDA to emotional and cognitive fluctuations, which makes this type of signal particularly susceptible to uncontrolled variations in the experimental environment. Factors such as skin hydration level, relative humidity, and ambient temperature can significantly affect the skin's electrical conductance and, consequently, the EDA signal. Furthermore, the sensor was positioned on the fingers, a region prone to involuntary movements or postural adjustments throughout the task, which may introduce noise into the signal or momentarily distort the measurements.

The average percentage difference from each parameter's baseline is presented in Table 8.

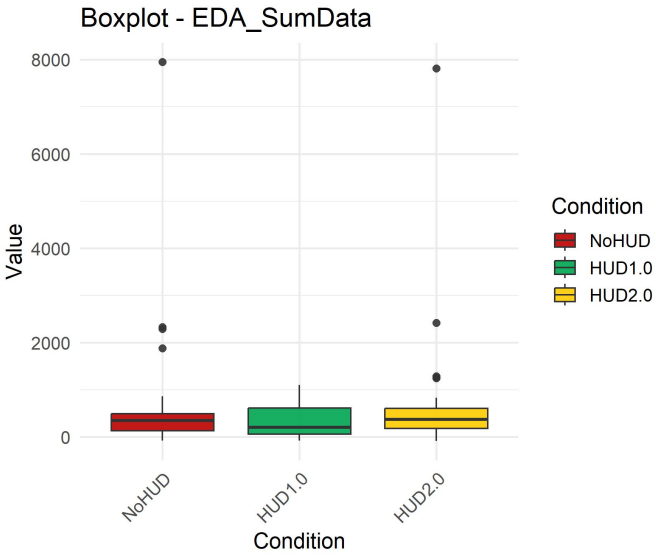

Figure 12: Sum of signal value with outliers.

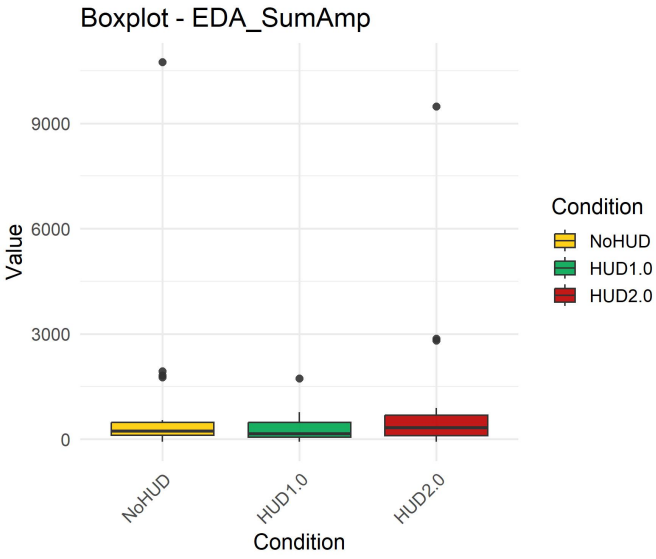

Figure 13: Sum of the amplitudes of these peaks.

Condition	Num_peaks	Sum_amp	Sum_data
No HUD	238.76	864.11	843.13
HUD 1.0	131.36	305.70	338.35
HUD 2.0	202.36	920.85	819.90

Table 8: EDA – Average percentage difference from baseline for each parameter (%).

The data for all electrodermal activity (EDA) metrics (`Num_peaks`, `Sum_amp`, and `Sum_data`) severely violated the assumptions for parametric testing. The non-parametric Friedman Test was therefore used for each metric to compare the HUD conditions. The analysis did not find a statistically significant effect for any of the EDA metrics (`Num_peaks`:

$p = 0.45$; `Sum_amp`: $p = 0.88$; `Sum_data`: $p = 0.88$).

4.6 ECG

For the analysis of the data obtained from the ECG signal, the same methodological approach used in the EDA analysis was adopted. Initially, the percentage differences of various cardiac parameters relative to each participant's baseline were calculated, with the aim of normalizing the physiological responses.

The boxplots were constructed for each parameter (Figures 14 to 17) in order to facilitate the visualization of data distribution across the experimental scenarios. The color scheme used was the same as in the other analyses, ensuring visual and interpretative consistency.

Figure 14: Heart Rate.

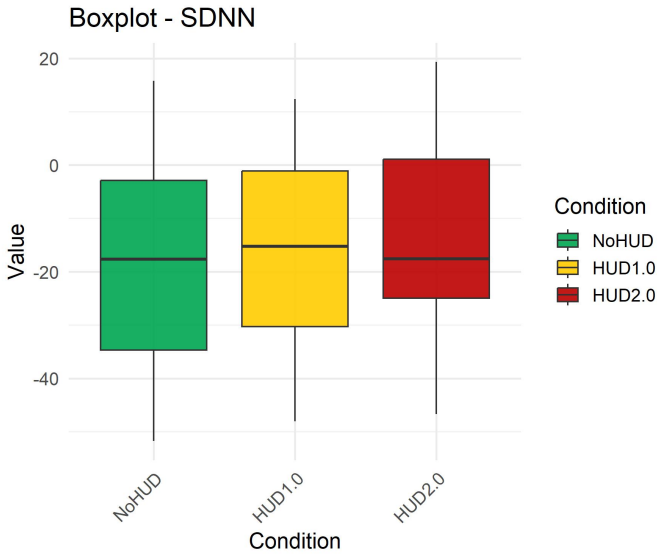

Figure 15: Standard Deviation of NN Periods.

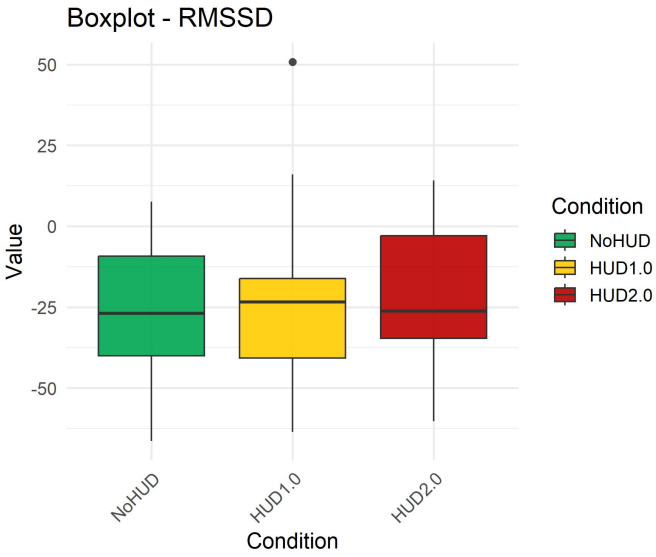

Figure 16: Root Mean Square of Successive Differences.

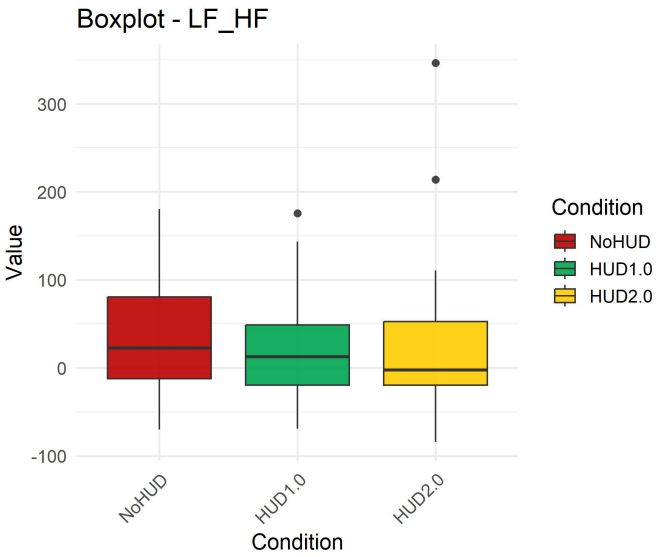

Figure 17: Ratio of Low Frequency and High Frequency.

Condition	HR	SDNN	RMSSD	LF/HF
No HUD	9.19	-17.50	-25.19	38.75
HUD 1.0	8.31	-16.60	-23.32	23.15
HUD 2.0	7.85	-12.27	-22.01	31.14

Table 9: ECG – Average percentage difference from baseline.

As shown in Table 9, it can be observed that, with the exception of the LF/HF parameter, the other physiological indicators showed that the scenario with the lowest mental workload was HUD 2.0, while the No HUD condition exhibited the highest variation values and was therefore associated with the highest perceived mental workload.

The ECG revealed no statistically significant physiological response to the different HUD conditions. A blocked ANOVA was applied to HR and SDNN, as they met parametric assumptions, but found no significant effects ($F(2,46) = 0.50, p = 0.61$ for HR; $F(2,46) = 1.71, p = 0.19$ for SDNN). Similarly, for the RMSSD and LF/HF ratio metrics, which violated these assumptions, the non-parametric Friedman Test also found no significant differences ($\chi^2(2) = 1.75, p = 0.42$ for RMSSD; $\chi^2(2) = 0.58, p = 0.75$ for the LF/HF ratio).

4.7 Eye Tracker

The pupil diameters of the participants' right and left eyes were recorded over time. Based on these data, the average pupil diameter was calculated, and the percentage variation relative to the baseline condition was subsequently determined. As a result, Figure 18 was obtained, and the corresponding averages are presented in Table 10.

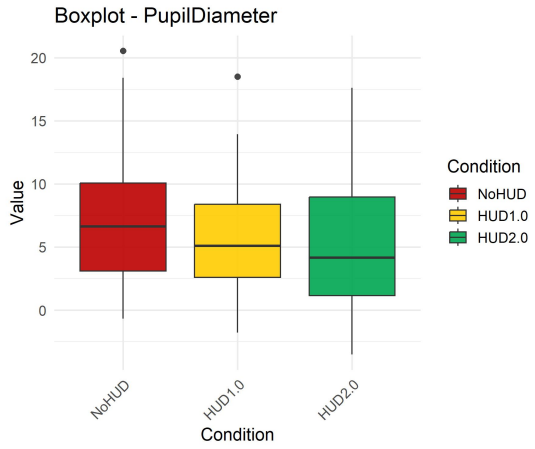

Figure 18: Pupil Diameter

Condition	Average Pupil Diameter (%)
No HUD	7.64
HUD 1.0	5.84
HUD 2.0	5.35

Table 10: Pupil Diameter- Average percentage difference from baseline.

A blocked ANOVA was conducted to evaluate the effect of the HUD conditions on the average pupil diameter, as the data met the necessary assumptions. The analysis indicated a statistically significant main effect ($F(2,46) = 4.85, p = 0.012$), suggesting that the interfaces had an overall impact on pupil size. However, the subsequent post-hoc Tukey's HSD test did not find significant differences between any specific pair of conditions.

4.8 Statistical Analysis

4.8.1 Spearman Correlation

With the database collected in the experiment, it was possible to perform a statistical analysis. The one chosen for this experiment was Spearman's. This method was chosen for its robustness in assessing monotonic relationships, which are not necessarily linear, making it suitable for the nature of the psychophysiological and subjective data. The resulting correlation matrix, presenting the correlation coefficients (ρ) for each pair of variables, is shown in Figure 19.

Based on the data presented, a moderate positive correlation was observed between the subjective workload scale NASA-TLX and the SWORD metric (ρ = 0.45), indicating that as perceived workload increases, so does the subjective dominance score. This reinforces the consistency of the self-reported responses.

As expected, variables of the same physiological nature showed strong correlations with each other. The heart rate variability metrics, SDNN and RMSSD, exhibited a strong positive correlation ($\rho = 0.64$), reflecting their common origin in the analysis of RR intervals. Similarly, the different parameters of electrodermal activity (EDA) were highly correlated with one another.

A moderate positive correlation was found between Heart Rate (HR) and Pupil Diameter ($\rho = 0.50$). This association suggests a co-activation of the sympathetic nervous system in response to cognitive effort, where both indicators tend to increase simultaneously.

On the other hand, the negative correlation between NASA-TLX and the physiological parameters EDA (SumAmp, ρ = -0.33) suggests that higher levels of perceived workload are associated with lower cumulative electrodermal activity. This likely reflects the complexity of the psychophysiological response. This can be attributed to different physiological factors - like the tendency for responses to decrease with familiarity, adaptation to the task, and differences in how each participant's body handles pressure — to technical limitations in signal acquisition, such as poor sensor contact or environmental interference.

Overall, the highest correlation coefficients were observed between variables of the same nature (subjective–subjective, physiological–physiological). This finding supports the hypothesis that the response to cognitive workload is multidimensional, and that integrated approaches are more appropriate for its quantification.

5 Conclusion

The design of the HUD significantly impacts the perceived cognitive workload of pilots, with the HUD 2.0 proving to be the most effective in reducing workload during the simulated task. This conclusion is supported by statistically significant results from NASA-TLX and SWORD, where HUD 2.0 is ranked as the least demanding interface; furthermore, it is corroborated by the pupillometry.

However, the reduction in perceived workload did not translate into a statistically significant improvement in flight performance, suggesting that under the tested conditions, the benefits of the HUDs were related more to operator comfort and perceived effort than to task execution accuracy. Moreover,

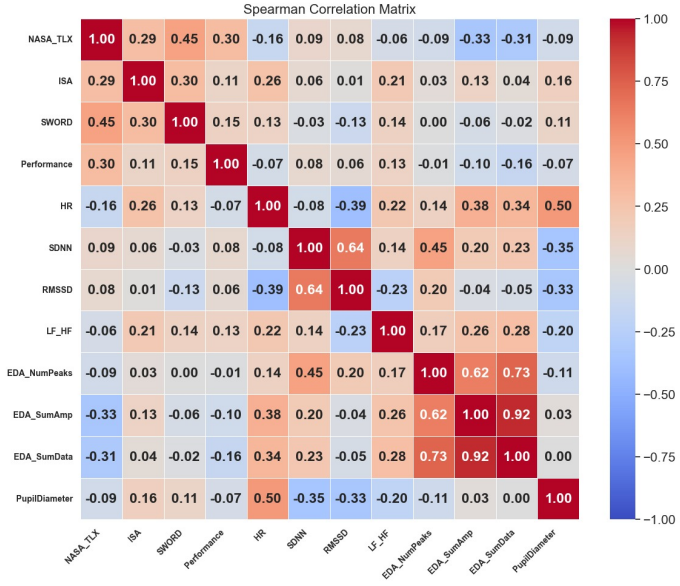

Figure 19: Spearman Correlation

a majority of the physiological sensors, including those for electrocardiography (ECG) and electrodermal activity (EDA), were not sensitive to detect significant differences between the interface conditions. This divergence underscores that different measurements capture different aspects of the human state, and a reduction in cognitive load does not mean a change across all physiological or performance metrics.

These findings have important implications for HMI design, reinforcing that graphical, context-aware interfaces can alleviate pilot strain. The study also serves as a caution, demonstrating that a multi-modal approach is essential for a comprehensive understanding of operator state. Future work should investigate these interfaces under conditions of higher task difficulty to determine if the benefits in perceived workload eventually translate to tangible performance gains. Additionally, the counterintuitive negative correlation found between subjective workload and electrodermal activity warrants further investigation to understand the complex psychophysiological dynamics at play. Ultimately, this research contributes to the development of safer and more human-centered aviation systems by providing a nuanced view of the interplay between interface design and pilot cognitive load.

References

[1] Boeing. Statistical summary of commercial jet airplane accidents, 2024. Boeing's Technical Report.

[2] T. Adams, R. Johnson, and S. Lee. Cognitive workload in aviation: Impacts and measurement methods. *Journal of Aviation Psychology*, 12(1):15–29, 2019.

[3] D. Williams and M. Patel. Human error and mental overload in flight operations. *Aerospace Human Factors Review*, 8(3):44–56, 2022.

[4] A. Miller and K. Thompson. Automation and cognitive load in modern flight decks. *Journal of Aerospace Systems*, 10(2):101–110, 2020.

[5] C. Lee and J. Kwon. Pilot interaction with automated systems: A cognitive load perspective. *International Journal of Aviation Psychology*, 15(1):25–40, 2023.

[6] J. B. Brookings, G. F. Wilson, and C. R. Swain. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, 42(3):361–377, 1996.

[7] A. F. Kramer. Physiological metrics of mental workload: A review of recent progress. *Multiple-task performance*, pages 279–328, 1991.

[8] N. Pongsakornsathien et al. Eye tracking in aerospace: Technologies and human-machine interaction. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 101–105, 2019.

[9] Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, volume 52, pages 139–183. Elsevier, 1988.

[10] Michael A Vidullch, G Frederic Ward, and James Schueren. Using the subjective workload dominance (sword) technique for projective workload assessment. *Human factors*, 33(6):677–691, 1991.

[11] J.R.S. Scarpari, M.W. Ribeiro, C.S. Deolindo, et al. Quantitative assessment of pilot-endured workloads during helicopter flying emergencies: an analysis of physiological parameters during an autorotation. *Scientific Reports*, 11:17734, 2021.

[12] M. H. O. C. da Silva, T. F. Macêdo, C. de Carvalho Lourenço, I. de Souza Rehder, A. A. da Costa Marchiori, M. P. Cesare, R. G. Cortes, M. M. Cardoso Junior, and E. Villani. Mental workload assessment in military pilots using flight simulators and physiological sensors. In *Human Mental Workload: Models and Applications. H-WORKLOAD 2021*, volume 1493 of *Communications in Computer and Information Science*, pages 99–115, Virtual, Online, 2021. 5th International Symposium on Human Mental Workload, Models and Applications (H-WORKLOAD 2021), Springer, Cham.

[13] A. C. Russo, A. Sarmento, I. S. Rehder, M. M. Cardoso-Junior, and E. Villani. Assessing mental workload and interface usability in military pilots: An advanced eye-tracking methodology. In *Proceedings of the 34th Congress of the International Council of the Aeronautical Sciences (ICAS)*, Florence, Italy, 2024. ICAS. Presented at ICAS 2024.

[14] Silvia Scannella, Carlo Chiorri, and Mickaël Causse. Assessment of ocular and physiological metrics to discriminate flight phases in real light aircraft. *Interna-*

tional Journal of Aerospace Psychology, 28(1-2):1–14, 2018.

[15] Andrew Gomes Pereira Sarmento, Thiago Rosado de Paula, Abner Souza de Oliveira, Edmar Thomaz da Silva, João Possamai, Henrique Costa Marques, Moacyr Machado Cardoso Junior, and Emilia Villani. A human-machine interface analysis for teleoperation of uav overtime delay. In *Proceedings of the International Council of the Aeronautical Sciences (ICAS)*. ICAS, 2022.