Real-Time Neurophysiological and Subjective Indices of Cognitive Engagement in High-Speed Flight

Matthew D'Alessandro; Ryan Mackie; Tom Berger; Carl Ott; Christopher Sullivan; Ian Curry

INTRODUCTION: Managing cognitive demand is critical for aviation safety. Yet, accurately assessing pilot workload during complex flight maneuvers remains challenging. This study evaluated an integrated methodology combining real-time cognitive engagement indicators to provide a comprehensive assessment and assess the reliability of physiological and subjective measures for monitoring operator state.

METHODS: Six experienced U.S. Army rotary-wing pilots completed simulated high-workload flight scenarios like low-altitude, reconnaissance, and air threat avoidance maneuvers. Continuous wireless electroencephalography (EEG), heart rate data, and subjective workload ratings were recorded during the flights.

RESULTS: EEG engagement indices and heart rate variability metrics demonstrated reliable within-subject consistency across trials for individual pilots, with mean intraclass correlation coefficient values ranging from 0.59–0.69. Both measures exhibited synchronized fluctuations across pilots at key events, increasing during high workload segments and decreasing in lower demand periods. Subjective ratings also showed good within-subject reliability, with mean intraclass correlation coefficient values ranging from 0.74–0.85. These findings underscore the reliability of our measurements, instilling confidence in the validity of our research.

DISCUSSION: The findings of this study provide strong support for the feasibility of using a multi-measure approach that integrates EEG, heart rate variability, and subjective ratings. This approach can continuously monitor real-time cognitive workload fluctuations during simulated rotary-wing operations. While objective measures showed within-subject consistency, substantial between-subject variability highlights the importance of individualized neurocognitive profiling. The integration of neurophysiological, autonomic, subjective, and environmental data holds great promise for the future of pilot workload assessment despite the challenges posed by individual differences.

KEYWORDS: high-speed flight, engagement index, physiological variables.

Assessing physiological and performance metrics in military pilots is critical for aviation safety. There has been growing interest in evaluating "cognitive workload" in pilots. However, it is essential to recognize that "workload" represents a global construct emerging from concurrent task demands. The Human Performance Envelope concept considers the complex interactions of factors impacting performance variations that must be inferred through analyzing behaviors as well as psychological and physiological processes.[1,2] Currently, there is no single definition or direct measure of workload due to its multifaceted nature. A high workload can overload attentional resources, thus prompting errors and accidents, while a very low workload may cause boredom and reduced situational awareness. Still, objectively measuring and quantifying pilot

cognitive workload remains an ongoing challenge in human factors research. The approach must incorporate inputs from different physiological sources into a multifaceted assessment, yet the key sources and definitions require further delineation and consensus. Nonetheless, prior studies have demonstrated

the feasibility of using physiological and subjective metrics to provide insight into workload quantification in pilots.[3]

Previous studies evaluating pilot cognitive workload have used many different physiological metrics and devices to attempt to objectively measure this multifaceted construct[4]; for example, using multichannel electroencephalography (EEG) recordings, Antonenko and colleagues found that power spectral features in the theta band correlated with task difficulty.[5] Similarly, other findings have demonstrated EEG coherence measures could distinguish high vs. low workload segments in real flight operations.[6] However, as EEG specifically measures electrical brain activity, it can only directly interpret levels of engagement as reflected in different frequency waves. While EEG serves as an objective metric of overall cognitive engagement, its utility is limited to interpreting engagement levels and fails to encapsulate the multifaceted nature of workload fully. Nonetheless, EEG can directly provide information about cognitive engagement and alertness through an engagement index $[\beta/(\alpha + \theta)]$, positively correlating with concentration and task load.[7]

In addition to EEG, heart rate variability (HRV) has shown promise as a physiological indicator that may aid in evaluating workload.[8,9] HRV reflects dynamic autonomic changes modulated by cognitive demands.[10,11] In a flight simulator study, decreased HRV was associated with executing complex maneuvers, suggesting HRV may provide an objective workload metric.[12] Integrating EEG and HRV assessment could offer complementary central and autonomic nervous system perspectives on pilot engagement. A combination approach to assessing physiological indices and subjective ratings may enable more robust real-time quantification of mental and physiological metrics in pilots during actual or simulated flights.

This study aimed to demonstrate the feasibility and validity of using EEG engagement indices, HRV, and modified Bedford workload ratings as metrics for monitoring cognitive engagement during simulated rotary-wing flight. Furthermore, this study aimed to investigate the aggregated physiological and subjective factors comprising cognitive workload in military pilots during simulated flight scenarios rather than examining the effects of cognitive workload itself. We hypothesized that: 1) the EEG engagement index would positively correlate with mental demand while HRV would negatively correlate with mental demand; 2) physiological metrics (EEG and HRV) and modified Bedford ratings would exhibit concurrent validity across flight segments; and 3) physiological measures would sensitively capture engagement fluctuations across scenarios. Our findings could facilitate adopting more objective, multifaceted mental workload assessments to enhance aviation safety.

METHODS

Subjects

This study was approved by the U.S. Army Combat Capabilities Development Command Human Research Protection Program (#23-016) and all pilots provided consent before participation.

Pilots were briefed on the experimental conditions and the proprietary use of the data that was obtained from their participation. This study involved six experienced (over 1000 military flight hours) U.S. Army pilots (four experimental pilots and two operational pilots) who flew four rotary-wing aircraft configurations in the NASA Ames Vertical Motion Simulator (NASA, Moffet Field, CA, United States).

Procedure

The aircraft configurations comprised two full-authority fly-by-wire flight control systems and two partial-authority systems. The full-authority configurations included a lift-offset coaxial-compound longitudinal thrust axis (Coax) helicopter and a winged single main rotor helicopter (WSMR). The partial-authority configurations consisted of the partial-authority winged single main rotor helicopter (PAW) and a conventional single main rotor helicopter (UH-60). Two mission vignettes, each approximately 10 min long, were developed to simulate scout missions flown by a single pilot under daytime conditions. A modified workload assessment was conducted through physiological monitoring and pilots' in-situ self-reporting. Data was collected for 2 wk. Three pilots participated in the experiments in week one, and three participated in the second week. During each week, pilots first flew one full-authority aircraft configuration followed by one partial-authority aircraft configuration. This approach allowed pilots sufficient time to train on the different configurations and mission scenarios before data collection. The first pilot cohort (week 1) flew the full-authority Coax and the partial-authority UH-60 aircraft. The second pilot cohort (week 2) flew the full-authority and partial-authority winged single main rotor aircraft (WSMR and PAW).

The first flight scenario (further referred to as Dawson) required an advance to contact along the Cheat River to establish an observation post and attack by fire position. Flying the designated route initially at the best range airspeed of 150 kn indicated airspeed (KIAS), pilots needed to switch to the best endurance airspeed at 100 KIAS before transitioning to 80 KIAS nap-of-earth flight along the rolling river valley. At specified points, the mission involved vertical and lateral unmasking to quickly achieve a stable hover for simulated deployment of air launched effects. The main objectives included precise low-altitude flight, airspeed control within 10 KIAS of specified values, rapid unmasked hover transitions, and timed route completion under 600 s.

The second scenario (further referred to as Ojai) consisted of a reconnaissance route along Highway 33 to identify enemy vehicles for surveillance handoff to organically controlled autonomous systems. This profile mandated that pilots maintain 100 KIAS and restrict above-ground level altitude to below 150 ft (46 m), all while flying a narrow, sinuous canyon. Enemy air defense systems in fictionalized threat scenarios necessitated the extremely low altitude. Pilots were instructed to deploy air-launched effects upon visually acquiring enemy vehicles before continuing the route to a specified endpoint. The main objectives included airspeed control within 5 KIAS, altitude conformance within strict margins, visually identifying

all simulated enemy entities, and avoiding ground collisions despite the hazardous terrain proximity.

Materials

Pilots provided real-time workload ratings in both flight scenarios at 30-s intervals through cockpit interfaces. Quantifiable performance data recorded by the simulator included aircraft position and control information, flight technical errors, run durations, and achievement of key timed mission tasks (data not shown). Continuous wireless recordings of brain electrical activity (EEG) and heart rate variability monitored physiological workload responses. Pilots were encouraged to minimize vocalizations to only necessary ones to reduce additional artifacts for the physiological data.

This study recorded EEG data using the 10-channel B-Alert EEG (B-Alert ×10, Biopac Systems Inc., Goleta, CA, United States) system, as EEG recordings have been successfully used in physiological assessment of military-relevant tasks.[13] We recorded continuous wireless EEG during the simulated flights and extracted spectral features to calculate the $\beta/(\alpha + \theta)$ band power ratios previously validated for workload assessment.[14] While all 10 EEG channels were recorded during the flights, only the frontal channels were used to calculate the $\beta/(\alpha + \theta)$ index. EEG data provides several objective metrics for quantifying pilots' engagement across varying task demands.

One effective index is the engagement index $\beta/(\alpha + \theta)$ (EEG-BAT), defined by Prinzel and colleagues. This index calculates the ratio of beta power to the sum of alpha and theta power in frontal EEG channels. A higher engagement index indicates greater cognitive engagement and alertness. Studies have demonstrated that this index positively correlates with task load, making it very effective for one metric of the ever-complex interpretation of workload quantification.[7] The engagement index has also been shown to reduce over time during sustained attention tasks, demonstrating the deterioration of engagement.[15] The EEG-BAT index is highly recommended for adaptive system design due to its sensitivity. Studies of complex piloting tasks have found higher values in fronto-central and parietal EEG regions, indicating greater engagement demands.[16]

This study recorded electrocardiography (ECG) data using the Polar H10 heart rate monitor (Polar Electro, Helsinki, Finland). The Polar H10 chest strap monitor provides continuous, accurate heart rate measurement via Bluetooth connectivity. Through the use of interference-preventing algorithms, the heart rate monitor can provide beat-to-beat precision. This validated monitor provides the necessary measurements to quantify heart rate variability during the simulated flight sessions. HRV was calculated from the recorded ECG data using a 60-s rolling window.

We employed a modified version of the Bedford Workload Scale to assess subjective workload. The original Bedford Workload Scale uses 10 rating points; however, this study reduced the number of rating options to 5—this adaptation aimed to facilitate real-time workload assessments by pilots during the execution of flight scenarios. The modified scale allowed pilots to provide subjective workload ratings ranging from 1 to 5, with 1 representing negligible effort and insignificant workload and 5 indicating an extremely high workload with no spare capacity. Five buttons in a line were placed on the instrument panel in a direct forward eye line to obtain these ratings. Every 30 s, the buttons would flash to cue the pilot to make a selection.

Using a modified Bedford Workload Scale with fewer rating options facilitated the pilots' ability to provide subjective workload assessments in real time without interrupting their primary task of flying the mission scenarios. By incorporating visual cues and a straightforward rating system, this study design aimed to minimize the cognitive load associated with the workload assessment process, thereby increasing the reliability and validity of the subjective workload measures. This approach allowed researchers to capture pilots' perceived workload fluctuations throughout the mission scenarios, enabling a more comprehensive understanding of the subjective workload experienced during various phases and events. The instantaneous workload assessments could then be analyzed with objective measures, such as performance metrics or physiological data, to gain deeper insights into the factors influencing workload and their potential implications for task performance and pilot well-being. Here, we compared the modified Bedford ratings to HRV and EEG-BAT using correlations to assess the relationship between the pilot's subjective interpretation of workload and their physiological engagement.

Statistical Analysis

All statistical analyses were performed using R (R Foundation for Statistical Computing, Vienna, Austria) and R Studio (Posit Software PBC, Boston, MA, United States).[17] All statistical tests were evaluated at a significance level of 0.05. Mean values reported in the text will include 95% confidence intervals in parentheses. To analyze changes in EEG and HRV over time, each flight scenario was divided into sections ranging from 30–150 s in length (**Table I**). Each section's start and stop times

Table I. Description and Average Time for Flight Scenario Sections.

SCENARIO & FLIGHT SECTION	TIME (s)	DESCRIPTION
Dawson		
1–2	88.0 (3.7)	Maintain 150 KIAS. Degraded visual environment from fog (Section 2 only).
3	125.4 (11.9)	Maintain 150 KIAS. Degraded visual environment from fog.
4	78.5 (15.7)	Decelerate to hover. Unmask and deploy air launched effects.
5	97.2 (12.3)	Accelerate and maintain 80 KIAS.
6	94.0 (19.6)	Decelerate to hover. Unmask and deploy air launched effects.
Ojai		
1	80.8 (11.4)	Maintain 150 KIAS.
2–4	109.6 (5.6)	Canyon flight. Maintain 100 KIAS. Identify enemy vehicles Deploy air launched effects.
5	54.9 (12.8)	Exit canyon and decelerate to hover.

Time values are mean and standard deviation. KIAS: knots indicated air speed.

were defined separately for individual flights because pilots did not complete flight sections in the same amount of time. All statistical analyses were performed separately for the two flight scenarios.

EEG-BAT and HRV data were averaged for each flight section. To account for individual differences between pilots, average EEG-BAT and HRV values were analyzed separately using mixed-effects linear regression models. Each regression model consisted of fixed categorical variables for aircraft (four levels) and flight section (5 or 6 levels for Ojai and Dawson, respectively), an interaction effect for aircraft and flight section, and a random intercept for each pilot. Regression model residuals were checked for normality and homoscedasticity to validate the linear regression models. Omnibus tests for aircraft, flight section, and the interaction effect were analyzed using the ANOVA function applied to the linear regression model.

When the omnibus test for the flight section was statistically significant, pairwise comparisons were made using the emmeans R package, with *P*-values adjusted using the Benjamini-Hochberg method to control the false discovery rate and balance controlling for Type 1 and Type 2 errors. To further reduce the chances of a Type 1 error, pairwise comparisons were only made against flight section 1 for each flight scenario. Using the initial active flying phase (section 1) as the reference point for analysis allowed us to isolate and understand the impact of subsequent events and tasks on the pilots' cognitive and physiological responses beyond the demands of simply operating the aircraft during routine flight conditions.

Modified Bedford workload ratings were collected at 30-s intervals during each flight. To create a continuous time-series variable, each rating was carried forward until the next rating. However, this method assumes that the workload changes instantaneously after each rating, which is unrealistic. Therefore, the workload ratings were shifted back by 15 s, allowing each 30-s window to be centered on when the pilot provided the rating. The continuous, shifted modified Bedford ratings were then averaged for each flight section, similar to EEG-BAT and ECG data processing. These average values were used for all analyses presented in this report.

Intraclass correlation coefficients (ICC) were used to examine test-retest reliability for EEG-BAT, HRV, and modified Bedford workload data collected during repeated flights. ICC estimates were calculated using the psych R package, based on a single measurement, two-way mixed-effects model, as is appropriate for test-retest validation. Interpreting ICC values can be subjective and depends on the context of the data being analyzed. Generally, ICC values between 0.50 and 0.75 suggest moderate reliability, values between 0.75 and 0.90 suggest good reliability, and values above 0.90 suggest excellent reliability. When interpreting ICC values, it is also important to note that coefficients reflect the degree of correlation and agreement between repeat measurements. Data with low correlation values will have low ICC values, even when the data is reproducible.[18]

RESULTS

A total of 52 Dawson flights and 48 Ojai flights were completed between the 6 pilots. Four Ojai flights were removed from all analyses due to the pilot not completing the flight as directed. Results from the analyses are presented in **Fig. 1**, **Fig. 2**, **Fig. 3**, and **Fig. 4**.

This study evaluated the EEG-BAT engagement index in pilots across four aircraft configurations and two flight scenarios. Despite between-subject variability in absolute EEG-BAT values, within-subject consistency emerged across repeated flights (Figs. 1A, 1B). The EEG-BAT patterns demonstrated minimal deviation across runs with the same aircraft and scenario for a given pilot. Repeated flights in the Dawson and Ojai scenarios had mean ICC values of 0.59 (0.41, 0.76) and 0.69 (0.52, 0.87), respectively (Fig. 2A). This suggests moderate test-retest reliability of the EEG-BAT index within individuals under similar conditions. Examining the EEG traces, common inflection points emerged across pilots during key phases of the flight scenarios. For all aircraft types and scenarios, increases and decreases in the EEG-BAT engagement index occurred at relatively consistent timeframes across pilots. For example, during the low-altitude nap-of-earth portions of the Dawson scenario, EEG-BAT spiked upwards at similar time points for all pilots, reflecting heightened attentional demands during this challenging flight segment. Conversely, a downward shift in EEG-BAT was observed in the straight and level transit phases, indicating reduced engagement.

The time-locked changes in EEG-BAT across individuals suggest that the brain's electrical activity reliably tracks variations in engagement associated with changing task demands and flight segments. This within-subject consistency supports the feasibility of objectively using EEG-derived engagement indices to quantify real-time workload fluctuations during simulated flight operations. The similar inflection patterns across pilots further demonstrate that all pilots experienced comparable engagement index profiles throughout the different flight scenarios. This provides concurrent validity for the sensitivity of the EEG-BAT index to accurately capture relative variations in task engagement and offer insights into overall cognitive demand.

We analyzed EEG-BAT values using mixed-effects linear regression models to examine between-subject consistency. The EEG-BAT regression model for the Dawson scenario showed that the interaction effect between section and aircraft was not statistically significant [$F(15, 306.0) = 1.16$, $P = 0.31$]. The interaction effect was removed and the regression model was re-evaluated. The updated regression model showed that the section and aircraft coefficients were statistically significant (**Table II**). Pairwise comparisons for sections showed statistically significant differences between section 1 and sections 2 and 4 (**Table III** and Fig. 1C).

The results of the EEG-BAT regression model for the Ojai scenario showed that the interaction, section, and aircraft coefficients were all statistically significant (Table II). Pairwise
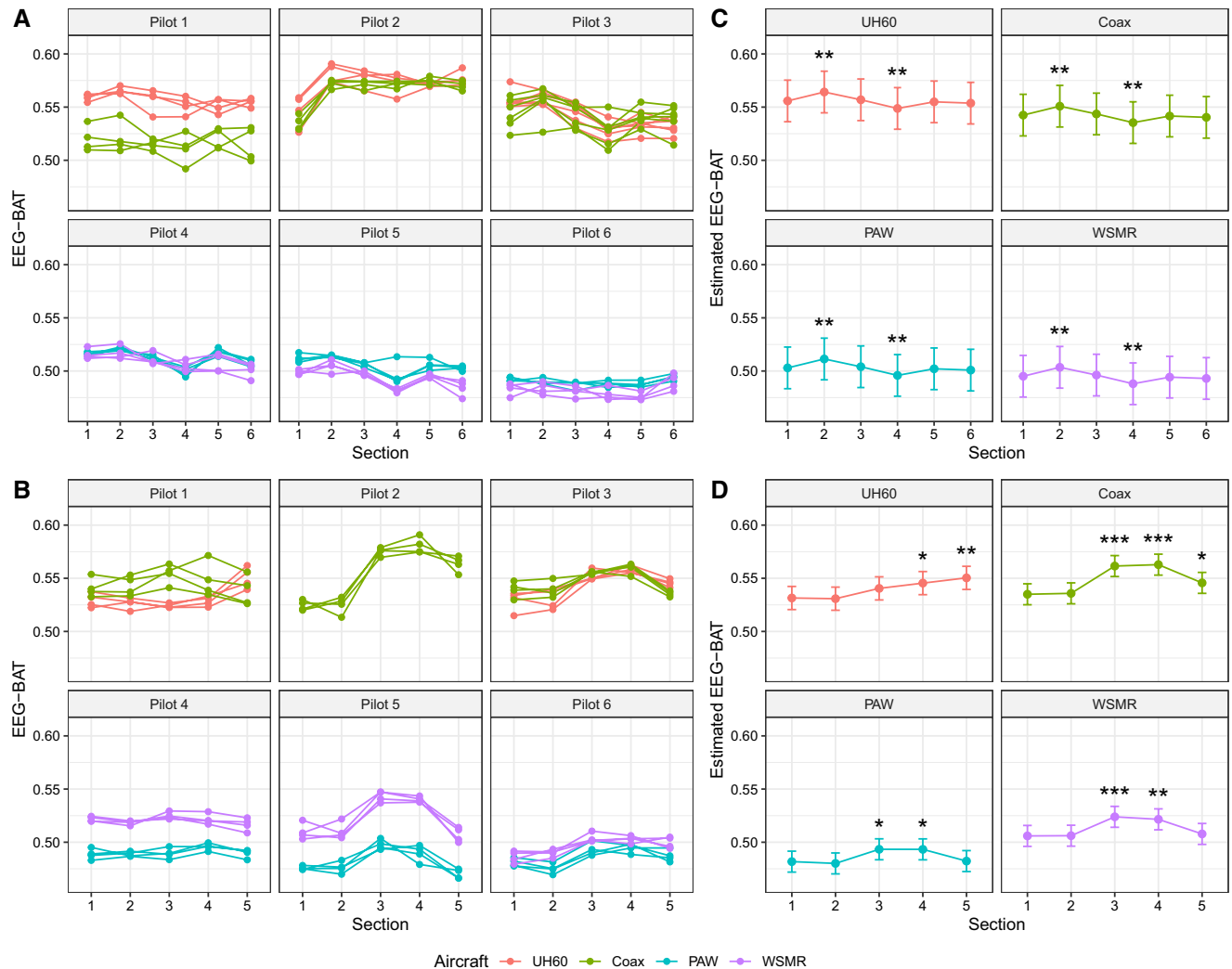
Fig. 1. EEG-BAT raw values and estimated values from mixed-effects linear regression models. A) Dawson raw values. B) Ojai raw values. C) Dawson estimated means and 95% CI. D) Ojai estimated means and 95% CI. Asterisks indicate statistically significant changes compared to section 1 (*$P \le 0.05$, **$P \le 0.01$, ***$P \le 0.001$).

comparisons for the sections showed significant differences for all four aircraft (Table III and Fig. 1D). The Coax aircraft showed significant differences between section 1 and sections 3, 4, and 5. The PAW aircraft showed significant differences between section 1 and sections 3 and 4. The UH-60 aircraft showed significant differences between section 1 and sections 4 and 5. The WSMR aircraft showed significant differences between section 1 and sections 3 and 4.

In addition to EEG, HRV was continuously monitored during the flights to provide an autonomic perspective on workload. Analysis of the HRV data revealed a similar pattern of within-subject consistency across repeated flights. For individual pilots, minimal deviations occurred in HRV values between runs with the same aircraft and scenario (Fig. 3A and 3B). Repeated flights in the Dawson and Ojai scenarios had mean ICC values of 0.65 (0.54, 0.77) and 0.62 (0.48, 0.76), respectively (Fig. 2B). This suggests moderate test-retest reliability for the HRV measures within a given pilot under similar conditions. Examination of the HRV

traces indicated common points of inflection that were consistent across pilots during key phases. For all aircraft, increases and decreases in HRV values occurred at relatively similar timeframes for each pilot. For example, during the high workload nap-of-earth segments in the Dawson scenario, HRV decreased at comparable points for pilots. This reflects heightened sympathetic activation and reduced HRV associated with increased task demands.

Conversely, straight and level transit phases elicited increases in HRV, indicating lower cognitive workload and greater parasympathetic activation. The synchronized HRV changes across pilots suggest that autonomic activity reliably tracks variations in workload between flight segments. The similar HRV fluctuation patterns across pilots provide concurrent validity that HRV can sensitively capture relative changes in cognitive workload and physiological arousal. The comparable inflection points also suggest that all pilots experienced similar workload profiles throughout the scenarios based on the autonomic response.
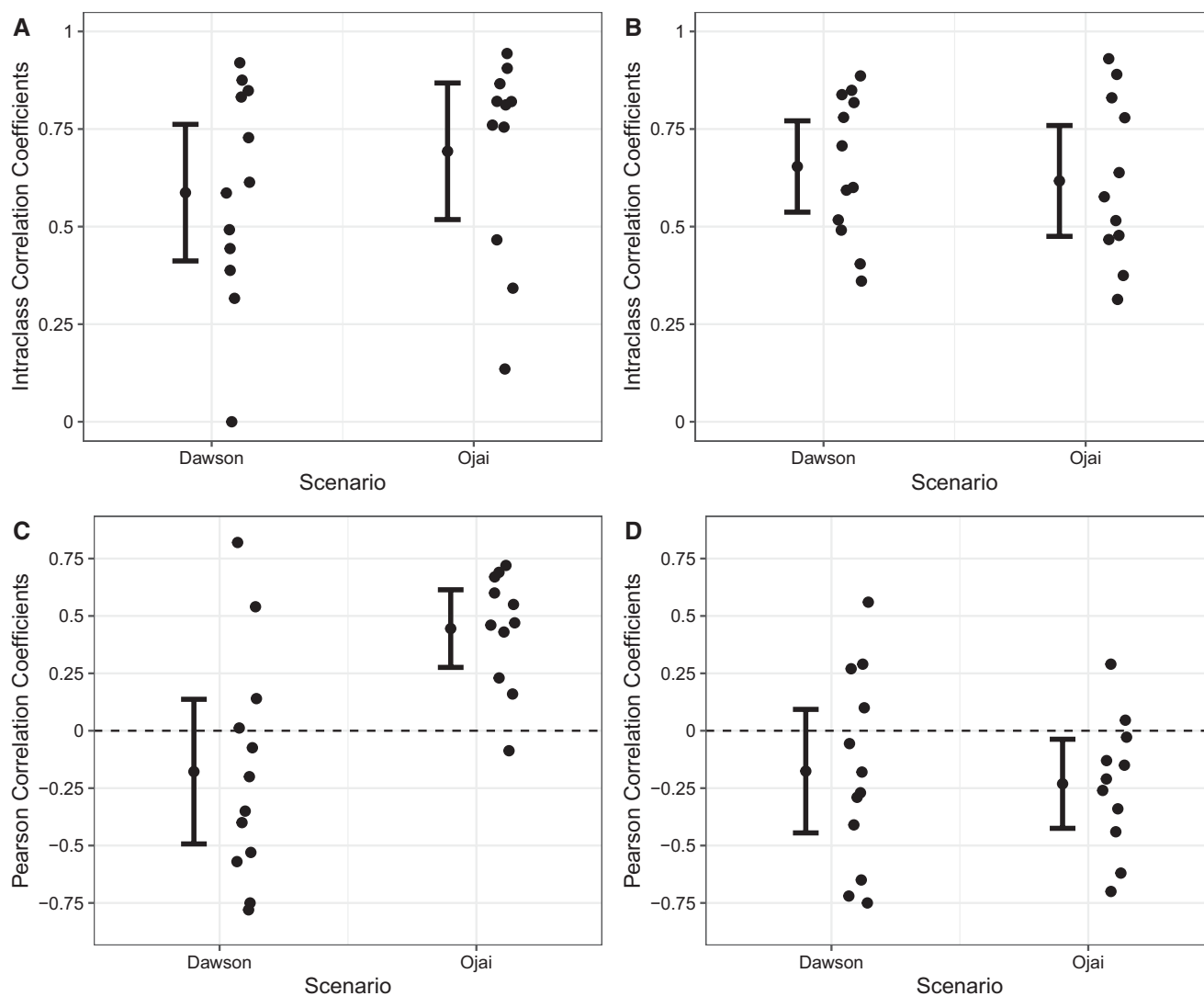
Fig. 2. A) Intraclass correlation coefficients for EEG-BAT. B) Intraclass correlation coefficients for HRV. C) Pearson correlation coefficients for Bedford workload ratings and EEG-BAT. D) Pearson correlation coefficients for Bedford workload ratings and HRV. Each subfigure shows raw data with associated mean and 95% CI.

As with the EEG data, the between-subject consistency of HRV was evaluated using mixed-effects linear regression models. The HRV regression model for the Dawson scenario showed that the interaction effect between section and aircraft was not statistically significant [$F(15, 306.0) = 1.04$, $P = 0.41$]. The interaction effect was removed and the regression model was re-evaluated. The updated regression model showed that section and aircraft fixed effects were statistically significant (Table II). Pairwise comparisons for sections showed statistically significant differences between section 1 and sections 4 and 5 (Table III and Fig. 3C). The Ojai HRV regression model showed that the interaction, section, and aircraft coefficients were all statistically significant (Table II). Pairwise comparisons for sections showed significant differences for two out of four aircraft (Table III and Fig. 3D). The PAW aircraft showed significant differences between section 1 and sections 2, 3, 4, and 5. The WSMR aircraft showed one significant difference between sections 1 and 2. Overall, analysis of the HRV data demonstrates

within-subject consistency and between-subject synchronization of workload-related HRV changes. This supports the feasibility of using HRV to objectively quantify a component of cognitive workload fluctuations in real-time flight settings. Additionally, HRV provides a complementary autonomic perspective to the EEG engagement indices.

The modified Bedford workload ratings demonstrated good consistency within individual pilots across the different flight scenarios. While pilots varied in their overall workload ratings, each tended to rate similar phases of flight consistently as higher or lower workload when comparing across different aircraft (Fig. 4). Repeated flights in the Dawson and Ojai scenarios had mean ICC values of 0.85 (0.81, 0.89) and 0.74 (0.62, 0.85), respectively, suggesting moderate to good test-retest reliability. Between pilots, the trends in workload ratings showed similarities, though there was less consistency compared to the within-subject ratings. This suggests that while all pilots experienced increases and decreases in workload at approximately the same
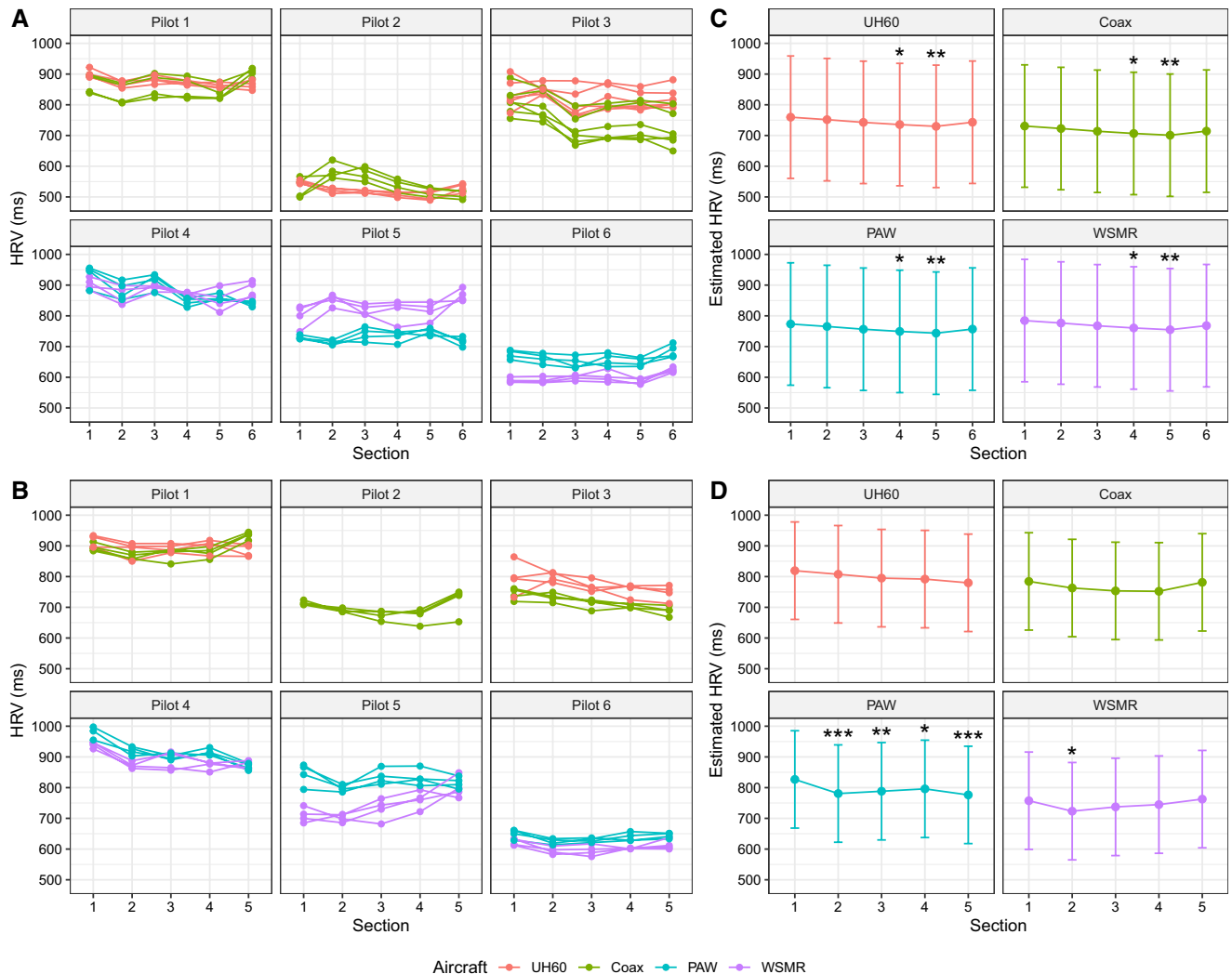
Fig. 3. HRV raw values (ms) and estimated values from mixed-effects linear regression models. A) Dawson raw values. B) Ojai raw values. C) Dawson estimated means and 95% CI. D) Ojai estimated means and 95% CI. Asterisks indicate statistically significant changes compared to section 1 (*$P \leq 0.05$, **$P \leq 0.01$, ***$P \leq 0.001$).

points of the flights, their subjective experience and rating of the workload demands varied more across individuals. Despite this between-subject variability, one clear pattern emerged: the lowest workload rating for nearly every pilot on almost every flight occurred during the beginning phase of operations (section 1). This aligns with an expected baseline of moderate workload associated with operating the aircraft during routine flying conditions before additional tasks or complexities were introduced later in each flight scenario.

To further explore the relationship between subjective workload assessments and objective measures of cognitive engagement, we conducted a correlation analysis between the pilots' ratings on the modified Bedford Workload Scale and each of the physiological engagement indices (EEG-BAT and HRV). Pearson correlation coefficients were calculated for each unique pilot, aircraft, and scenario combination. Figs. 2C and 2D show the Pearson correlation coefficients for each physiological index and scenario and the mean correlation coefficients with 95% confidence intervals. The modified Bedford-EEG correlations

for the Ojai scenario showed the best consistency. All correlation coefficients were positive apart from one value, and most correlations were close to or above 0.5. The mean correlation for Ojai was 0.45 (0.28, 0.61), suggesting a low to moderate positive correlation between the modified Bedford ratings and EEG-BAT values. The modified Bedford-EEG correlations for the Dawson scenario were less consistent and showed the opposite trend compared to the Ojai scenario. Out of 12 total correlation coefficients, 8 were negative. The mean correlation for Dawson was −0.18 (−0.49, 0.14), suggesting a low negative correlation. This is the opposite of what was hypothesized based on existing literature showing that EEG positively correlates with workload measures.

Modified Bedford-HRV correlations were less consistent compared to the modified Bedford-EEG correlations. Overall, there were more negative correlations than positive and the mean correlations were also negative for both the Ojai [−0.23 (−0.43, −0.04)] and Dawson [−0.18 (−0.45, 0.09)] scenarios. This aligns well with existing literature showing a negative
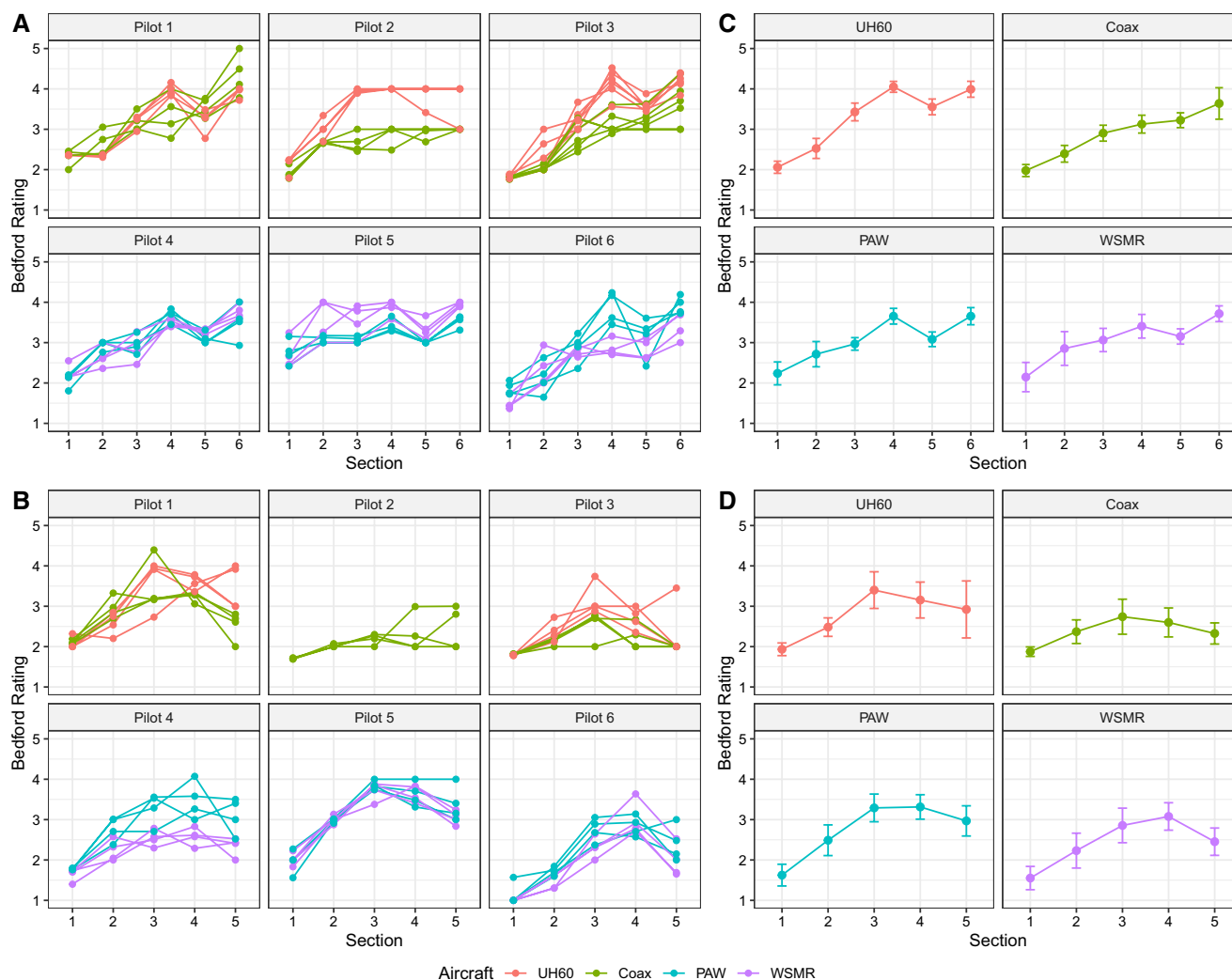
Fig. 4. Bedford workload rating raw values and mean values. A) Dawson raw values. B) Ojai raw values. C) Dawson means and 95% CI. D) Ojai means and 95% CI.

correlation between HRV and workload measures. However, the modified Bedford-HRV correlations were not as strong as the modified Bedford-EEG correlations. Only 6 HRV correlations had an absolute value ≥ 0.5, whereas 11 EEG correlations had an absolute value ≥ 0.5.

Table II. Results for EEG-BAT and HRV Mixed-Effects Linear Regression Models.

METRIC	SCENARIO	COEFFICIENT	DF-b	DF-w	*F*	*P*
EEG-BAT	Dawson	Section	5	306.0	9.71	< 0.001
		Aircraft	3	14.0	32.29	< 0.001
	Ojai	Interaction	12	213.9	2.70	0.002
		Section	4	213.9	27.95	< 0.001
		Aircraft	3	14.8	107.01	< 0.001
HRV	Dawson	Section	5	306.0	3.34	0.006
		Aircraft	3	13.7	7.38	0.003
	Ojai	Interaction	12	214.0	2.43	0.006
		Section	4	214.0	7.71	< 0.001
		Aircraft	3	13.6	38.23	< 0.001

Columns report between-groups (DF-b) and within-groups degrees of freedom (DF-w), *F*-statistics, and *P*-values.
EEG-BAT: engagement index $\beta/(\alpha + \theta)$; HRV: heart rate variability.

DISCUSSION

This study evaluated experienced U.S. Army pilots during simulated rotary-wing flight scenarios to investigate the aggregated physiological and subjective factors that comprise cognitive workload. EEG engagement indices revealed consistent patterns within individual pilots across repeated trials of the same scenario. This observation, along with moderate ICC values, demonstrates the test-retest reliability of these metrics under similar conditions for a given pilot, aligning with prior research on the within-subject reliability of EEG for workload assessment.[19] Simultaneously, substantial variability emerged between pilots regarding absolute baseline values and workload response profiles. This highlights the need to establish individualized cognitive state profiles through baseline EEG assessments before evaluating workload, as individuals differ in arousal, attention, and cognitive processing due to underlying neurophysiological and personality differences.[20,21] Thus, relative changes from baseline are more informative than absolute cross-subject comparisons.

Table III. Results for EEG-BAT and HRV Pairwise Comparisons from Mixed-Effects Linear Regression Models.

METRIC	SCENARIO	AIRCRAFT	SECTIONS	*P*
EEG-BAT	Dawson	All	1 vs. 2	0.001
		All	1 vs. 4	0.005
	Ojai	Coax	1 vs. 3	< 0.001
		Coax	1 vs. 4	0.001
		Coax	1 vs. 5	0.025
		PAW	1 vs. 3	0.037
		PAW	1 vs. 4	0.037
		UH-60	1 vs. 4	0.028
		UH-60	1 vs. 5	0.003
		WSMR	1 vs. 3	< 0.001
		WSMR	1 vs. 4	0.001
HRV	Dawson	All	1 vs. 4	0.033
		All	1 vs. 5	0.007
	Ojai	PAW	1 vs. 2	< 0.001
		PAW	1 vs. 3	0.005
		PAW	1 vs. 4	0.028
		PAW	1 vs. 5	< 0.001
		WSMR	1 vs. 2	0.025

Columns report aircraft type, sections that showed significant differences from section 1, and *P*-values. *P*-values were adjusted using the Benjamini-Hochberg method. EEG-BAT: engagement index $\beta/(\alpha + \theta)$; HRV: heart rate variability; Coax: lift-offset coaxial-compound longitudinal thrust axis helicopter; PAW: partial-authority winged single main rotor helicopter; UH-60: conventional single main rotor helicopter; WSMR: winged single main rotor helicopter.

Heart rate variability data reinforced this concept, with moderate within-subject consistency but wide between-subject variations. The reliability of HRV as a workload indicator was further evidenced by synchronized changes at key mission points, aligning with prior research on HRV as a physiological marker of mental workload.[6] During high-intensity scenarios like low-altitude nap-of-earth flight segments, EEG-BAT and HRV shifted in the direction indicating heightened engagement and physiological arousal.[9,21] However, terrain complexity and momentary mission relevancy also impacted pilot focus, complicating workload assessments. For example, metrics showed greater variability during transit through wide-open canyon areas. Conversely, cognitive engagement intensified when navigating narrow, enclosed areas requiring greater precision. This illustrates how workload metrics may fluctuate independently from flight technical demands based on the perception of situational importance, consistent with cognitive theories that situational factors shape resource allocation.[1,22] The same terrain features can impose different cognitive loads depending on the mission phase. This underscores the need for holistic integration of physiological data with flight technical parameters, performance metrics, and subjective appraisals.[19,20]

Pearson correlation coefficients were computed to investigate the potential association between subjective workload ratings and objective engagement indices. By establishing these correlations, we sought to validate the subjective workload assessments obtained through the modified Bedford Workload Scale against the objective physiological measurements. A strong correlation would support the convergent validity of the subjective ratings, indicating that they accurately reflect the pilots' cognitive engagement levels during the mission scenarios. Furthermore, correlation analysis could provide insights

into the sensitivity and responsiveness of the subjective workload scale in capturing variations in cognitive workload, as well as the potential relationship between subjective perceptions of workload and objective physiological measures of engagement.[23]

The correlation analysis between subjective workload ratings obtained through the modified Bedford Workload Scale and objective measures of cognitive engagement yielded contrasting results across the two mission scenarios investigated in this study. In the Ojai scenario, a statistically significant positive mean correlation was observed between the pilots' subjective workload ratings and the EEG-BAT index (Fig. 2C). This positive correlation suggests that as the perceived workload increased, as reported by the pilots using the modified Bedford Workload Scale, the objective measures of cognitive engagement, as reflected by the EEG-BAT index, also tended to increase. This finding supports the convergent validity of the subjective workload assessments, indicating that the pilots' subjective workload experiences aligned with their physiological states of cognitive engagement during this mission scenario.

Conversely, the Dawson flight scenario showed a negative but nonsignificant mean correlation between the subjective workload ratings and the EEG-BAT index. This lack of correlation implies that the pilots' subjective perceptions of workload did not consistently correspond with the objective measures of cognitive engagement derived from the EEG data. This discrepancy could arise from various factors, such as:

1. Task characteristics: the nature and demands of the Dawson mission scenario may have involved cognitive processes or workload dimensions that were not accurately captured by EEG, leading to a dissociation between subjective experiences and objective physiological measures.
2. Individual differences: variations in individual characteristics, such as cognitive strategies, workload management skills, or physiological responses, could have influenced the relationship between subjective workload ratings and objective engagement measures in the Dawson scenario.
3. Temporal dynamics: the temporal dynamics of subjective workload perceptions and physiological responses may have needed to be perfectly synchronized, potentially contributing to the lack of correlation observed in the Dawson scenario.

Interestingly, neither of the flight scenarios exhibited statistically significant mean correlations between the subjective modified Bedford workload ratings and HRV (Fig. 2D). HRV is a measure of autonomic nervous system activity and is often used as an indicator of physiological stress and mental workload. The absence of correlations with HRV data suggests that the subjective workload ratings and the EEG-BAT index may have captured different aspects of cognitive workload, potentially reflecting distinct underlying processes or mechanisms. The divergent patterns observed across the two scenarios highlight the complex nature of workload assessment and the potential influence of task characteristics, individual differences, and measurement modalities on the relationships between subjective and objective workload measures. These findings underscore the

importance of multiple data sources and the need to consider contextual factors when interpreting workload assessments in complex operational environments.

Future research could further investigate the factors contributing to the observed discrepancies, such as task analysis, individual differences in workload perception and management strategies, and the temporal dynamics of subjective and objective workload measures. Additionally, exploring the integration of subjective and physiological measures could provide a more comprehensive understanding of the multidimensional nature of cognitive workload and its implications for human performance in operational settings. Our findings support the feasibility of combining EEG, HRV, and subjective workload ratings to evaluate multidimensional workload, which aligns with previous recommendations for using a combination assessment approach.[14,24]

However, substantial processing is required to translate raw physiological signals into meaningful workload constructs. Individualized baselines and calibrations are essential to account for neurocognitive variability.[21,25] Hence, an integrated methodology may prove promising to provide continuous real-time monitoring of the complex interactions between individual pilot cognitive states, flight environments, aircraft systems, and control interfaces. This can significantly enhance aviation safety and training by optimizing equipment and procedures for human-system performance.[1,19] However, addressing individual variability while considering contextual influences will remain a critical challenge for managing workload in real time.[20,22]

This study has some fundamental limitations to consider. The small sample size of only six experienced pilots limits generalizability, necessitating larger and more diverse pilot samples in future work. The fixed order of flight scenarios and aircraft configurations may have introduced order effects and bias as pilots gained familiarity over time, which could influence later performance independently of aircraft differences. Randomizing the order of flight scenarios and aircraft across pilots would mitigate such biases. However, the lack of randomization did not impact the assessment of measure reliability, which was the core focus of this study. Addressing these limitations with expanded sample sizes, randomization, and controlling for pilot experience levels will strengthen the interpretability and generalizability of findings from subsequent studies.

The flight simulations involved full motion in six axes, potentially introducing artifacts into the EEG recordings. Pilot head movements and vocalization during the simulated flights may have also contributed to artifact contamination of the EEG signals. Careful sensor placement and advanced artifact removal techniques were implemented to minimize the influence of such artifacts on the EEG data and the derived engagement index calculations. However, the residual artifact cannot be completely ruled out and may have impacted the EEG results. Future studies could incorporate additional artifact detection and removal methods and complementary neurophysiological measures that are less susceptible to motion artifacts to further improve the robustness and reliability of the cognitive workload assessments in dynamic operational environments.

It should be noted that the B-Alert ×10 EEG system used in this study captures brain activity from a limited set of electrode locations (Fz, F3, F4, Cz, C3, C4, POz, P3, and P4) based on the International 10-20 System. While these locations provide valuable information about cognitive engagement and attentional processes, they do not provide comprehensive coverage of other brain regions that may also contribute to cognitive workload. Specifically, the B-Alert ×10 array lacks sensors at locations such as FP1, FP2, F7, F8, T3, T4, T5, T6, O1, and O2, which are known to capture neural activity related to emotional and contextual attention, working memory beyond frontal regions, and language processing—all of which are integral components of cognitive workload. For this study, we focused solely on the engagement index derived from the electrode recordings from the frontal cortex electrodes only. This limited spatial resolution of the EEG recordings is a notable limitation of the current study, as it may have failed to capture the full extent of brain dynamics underlying the multifaceted construct of cognitive workload. Future research employing higher-density EEG systems or complementary neuroimaging techniques could provide a more comprehensive assessment of the neural correlates of cognitive workload in complex operational environments.

Another significant limitation to consider is the fixed placement of sensors on the plastic band of the B-Alert ×10 system. This design may not accommodate varying head sizes accurately, leading to potential inaccuracies in sensor positioning, especially for rear-mounted sensors like POz. Furthermore, the less-than-snug fitting of the sensor band could allow for undesirable movement and shifting during the dynamic flight simulations, introducing additional errors, particularly at sites like POz. Again, only data obtained from frontal lobe sensors were used for analysis for this study. However, researchers need to be aware that such inaccuracies in sensor placement and movement artifacts could compromise the spatial precision of the EEG recordings, affecting the reliability and interpretability of the derived engagement indices. Future studies should consider employing EEG systems with more flexible, adjustable sensor arrangements or cap-based electrode arrays that conform more precisely to individual head sizes and shapes, minimizing positioning errors and movement-related artifacts.

Moreover, this study employed a modified version of the Bedford Workload Scale, reducing the rating options from 10 to 5. While this adaptation aimed to facilitate real-time workload assessments by pilots during flight execution, it deviates from the original, normed version of the scale. Modifying the scale in this manner raises concerns regarding the validity and comparability of the subjective workload ratings obtained in this study with established norms and findings from previous research using the standard Bedford Workload Scale. The reduced number of rating options may have limited the granularity and sensitivity of the subjective assessments, potentially affecting the ability to capture subtle variations in perceived workload. Furthermore, the absence of normative data and validation studies for the modified scale makes it challenging to interpret the subjective ratings in a broader context. Future research should consider employing the original Bedford Workload Scale or conducting

thorough validation studies for any modified versions to ensure the reliability and validity of subjective workload measures.

The study design also may not have effectively distinguished changes in cognitive workload within each flight scenario. Capturing workload fluctuations across phases of flight could strengthen future experimental approaches in this area. While incorporating EEG and HRV data, additional real-time physiological measures like eye tracking, pupillometry, skin impedance, or other biomarkers may provide more significant insights into pilots' mental states. Relying solely on subjective workload ratings leaves findings susceptible to response biases. Overall, while the unique engagement and experience of pilots represent a strength, several limitations should be addressed in follow-up work to bolster the interpretability and generalizability of the findings. Incorporating larger subject samples, additional metrics, design enhancements, and a wider variety of aircraft will ultimately provide a deeper perspective on the cognitive impacts of advanced flight systems.

Importantly, this study provides insight into the feasibility of using EEG, HRV, and subjective workload ratings to quantify cognitive workload in real time in a military aviation setting. The results highlight the importance of accounting for individual variability and contextual influences when interpreting physiological data. Within-subject consistency supports the feasibility of EEG and HRV for tracking workload fluctuations, but individual differences necessitate personalized baselines to capture relative changes. Workload metrics also vary with terrain complexity and momentary mission relevance, emphasizing the need to integrate physiological data with environmental and performance data. Although advanced processing is required to translate signals into meaningful constructs, this approach shows promise for the continuous real-time monitoring of complex interactions between pilot cognitive states, flight contexts, aircraft systems, and control interfaces.

Combining objective physiological metrics with subjective workload assessments may enhance our ability to optimize aviation training and safety by quantifying the multidimensional nature of workload. However, accounting for individual neuro-cognitive variability while also considering situational influences remains crucial, requiring a move from generalized models toward individually calibrated interfaces tailored to each pilot's cognitive capabilities and the specific environmental context. Furthermore, incorporating additional monitoring techniques like pupillometry and eye tracking could provide further insights into workload fluctuations. The findings from this study represent an essential step toward developing perceptive aircraft systems that synthesize physiological, environmental, performance, and subjective data to facilitate optimal human-machine cooperation.

ACKNOWLEDGMENTS

Authors and Affiliations: Matthew D'Alessandro, Ph.D., Research Psychologist, Warfighter Performance Group, Ryan Mackie, M.S., Christopher Sullivan, M.S., and Ian Curry, BM, BS, U.S. Army Aeromedical Research Laboratory, Fort Novosel, AL, United States; and Tom Berger, Ph.D., and Carl Ott, M.S., U.S. Army Combat Capabilities Development Command Aviation & Missile Center, Moffett Field, CA, United States.

REFERENCES

1. McKendrick R, Feest B, Harwood A, Falcone B. Theories and methods for labeling cognitive workload: classification and transfer learning. Front Hum Neurosci. 2019; 13:295.
2. Tao D, Tan H, Wang H, Zhang X, Qu X, Zhang T. A systematic review of physiological measures of mental workload. Int J Environ Res Public Health. 2019; 16(15):2716.
3. Masi G, Amprimo G, Ferraris C, Priano L. Stress and workload assessment in aviation—a narrative review. Sensors (Basel). 2023; 23(7):3556. Erratum in: Sensors (Basel). 2024; 24(2):690.
4. Nixon J, Charles R. Understanding the human performance envelope using electrophysiological measures from wearable technology. Cogn Technol Work. 2017; 19(4):655–666.
5. Antonenko P, Paas F, Grabner R, Gog T. Using electroencephalography to measure cognitive load. Educ Psychol Rev. 2010; 22(4):425–438.
6. Dussault C, Jouanin JC, Philippe M, Guezennec CY. EEG and ECG changes during simulator operation reflect mental workload and vigilance. Aviat Space Environ Med. 2005; 76(4):344–351.
7. Prinzel LJ, Freeman FG, Scerbo MW, Mikulka PJ, Pope AT. A closed-loop system for examining psychophysiological measures for adaptive task allocation. Int J Aviat Psychol. 2000; 10(4):393–410.
8. Hinde K, White G, Armstrong N. Wearable devices suitable for monitoring twenty-four hour heart rate variability in military populations. Sensors (Basel). 2021; 21(4):1061.
9. Luft CD, Takase E, Darby D. Heart rate variability and cognitive function: effects of physical effort. Biol Psychol. 2009; 82(2):186–191.
10. Durantin G, Gagnon JF, Tremblay S, Dehais F. Using near-infrared spectroscopy and heart rate variability to detect mental overload. Behav Brain Res. 2014; 259:16–23.
11. Wang Z, Zhu K, Kaur A, Recker R, Yang J, Kiourti A. Quantifying cognitive workload using a non-contact magnetocardiography (MCG) wearable sensor. Sensors (Basel). 2022; 22(23):9115.
12. Cao A, Chintamani KK, Pandya AK, Ellis RD. NASA TLX: software for assessing subjective mental workload. Behav Res Methods. 2009; 41(1):113–117.

13. Diaz-Piedra C, Sebastián MV, Di Stasi LL. EEG theta power activity reflects workload among army combat drivers: an experimental study. Brain Sci. 2020; 10(4):199.

14. Hogervorst MA, Brouwer AM, van Erp JB. Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. Front Neurosci. 2014; 8:322.

15. Ko LW, Komarov O, Hairston WD, Jung TP, Lin CT. Sustained attention in real classroom settings: an EEG study. Front Hum Neurosci. 2017; 11:388.

16. Ismail LE, Karwowski W. Applications of EEG indices for the quantification of human cognitive performance: a systematic review and bibliometric analysis. PLoS One. 2020; 15(12):e0242857.

17. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, et al. Welcome to the tidyverse. J Open Source Softw. 2019; 4(43):1686.

18. Koo TK, Mae YL. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Chiropr Med. 2016; 15(2):155–163. Erratum in: J Chiropr Med. 2017; 16(4):346.

19. Borghini G, Astolfi L, Vecchiato G, Mattia D, Babiloni F. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neurosci Biobehav Rev. 2014; 44:58–75.

20. Gevins A, Smith ME. Neurophysiological measures of cognitive workload during human-computer interaction. Theor Issues Ergon Sci. 2003; 4(1–2):113–131.

21. Hankins TC, Wilson GF. A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. Aviat Space Environ Med. 1998; 69(4):360–367.

22. Cegarra J, Chevalier A. The use of Tholos software for combining measures of mental workload: toward theoretical and methodological improvements. Behav Res Methods. 2008; 40(4):988–1000.

23. Cao X, MacNaughton P, Cadet LR, Cedeno-Laurent JG, Flanigan S, et al. Heart rate variability, and performance of commercial airline pilots during flight simulations. Int J Environ Res Public Health. 2019; 16(2):237.

24. Hsu BW, Wang MJ, Chen CY, Chen F. Effective indices for monitoring mental workload while performing multiple tasks. Percept Mot Skills. 2015; 121(1):94–117.

25. Herold F, Törpel A, Hamacher D, Budde H, Zou L, et al. Causes and consequences of interindividual response variability: a call to apply a more rigorous research design in acute exercise-cognition studies. Front Physiol. 2021; 12:682891.