Correction to: Multi-modal physiological sensing approach for distinguishing high workload events in remotely piloted aircraft simulation

Denny Yu¹ · Chantale Wilson Antonik² · Frederick Webber² · Eric Watz² · Winston Bennett Jr²

Published online: 29 July 2021

© Springer Nature Switzerland AG 2021

Abstract

Remotely piloted aircraft (RPA) operations are often characterized as highly taxing and dynamic. Physiological sensing technology can enhance personnel monitoring and training for these high-stress environments; however, work assessing the effectiveness of physiological sensors during RPA operations is limited. The proposed work tests two hypotheses: (1) physiological sensors can distinguish operator workload between scenario difficulty levels, and (2) the sensors can quantify the impact of RPA events on the operator workload. Twelve pilots completed RPA simulations at all three difficulty levels while physiological sensors collected electroencephalogram (EEG) and heart rate activity. Hypotheses were tested using mixed-effects models. Observed heart rate variability metrics did not differ among the three scenario difficulty levels except for LF/HF ratio. A 47% and 57% reduction in alpha band power was observed between easy and hard difficulty levels for the frontal and parietal channels, respectively. Abort and Reach objective events resulted in 0.2–0.3 dB lower beta activity and 66 ms increased heart rate, while losing sight of the objective (e.g., fog) had 0.72 dB increased high beta activity. Different physiological modalities (EEG and ECG) had varying effectiveness in distinguishing scenario difficulty and RPA events, suggesting a hybrid sensing approach may provide more insight than just using one modality. In conclusion, physiological sensing can distinguish operator response to scenario difficulty and events in high-fidelity RPA simulations.

Keywords Physiological psychology · Simulation and training · Remotely piloted aircraft · Workload

This manuscript is submitted for the Special Issue on “Human-Autonomy Teaming in Military Contexts”

The original article can be found online at <https://doi.org/10.1007/s42454-020-00016-w>.

Correction to: Human-Intelligent Systems Integration (2019) 1:89–99 <https://doi.org/10.1007/s42454-020-00016-w>

This article has been invited for inclusion in the special issue on “Human-Autonomy Teaming in Military Contexts” [Volume 1, Issue 2–4]

✉ Denny Yu
DennyYu@purdue.edu

¹ Purdue University, 315 N. Grant St., West Lafayette, IN 47907, USA

² Air Force Research Laboratory, Warfighter Readiness Research Division, Dayton, OH, USA

1 Introduction

Remotely piloted aircraft (RPA) personnel frequently operate in high-stress, unpredictable technical environments. Due to their versatility, RPAs provide support in many different mission areas, e.g., intelligence, surveillance, and reconnaissance; close-air support; air interdiction; combat search and rescue; and strike coordination and reconnaissance. RPAs are often characterized as highly taxing and dynamic, since RPA missions can induce a wide range of experiences from long periods of monotony to sudden, extremely demanding and difficult situations. Control of these unmanned vehicles have known issues in regard to human performance, and issues include multi-tasking, trust, situation awareness, and operator workload (Chen et al. 2010; W. Zhang et al. 2016).

Physiological sensing technology has the potential to detect and monitor aforementioned issues that affect human performance. Although work demonstrating the link between physiological metrics and constructs such as trust (Akash et al. 2018; Dong et al. 2015; Fu et al. 2018; M. Wang et al. 2018) and situation awareness (Catherwood et al. 2014; Kaur et al. 2019; Yeo et al. 2017; Zhang et al. 2021) is still developing, the ability of physiological sensors to measure workload has been well-established.

Many studies across various domains demonstrate that physiological sensors can be used to predict operator workload (Charles and Nixon 2019). Heart rate and heart rate variability (HRV) metrics have been correlated with workload in many studies and application areas (Abdelrahman et al. 2015; Cinaz et al. 2013; McDuff et al. 2014; Mehler et al. 2012; Moses et al. 2007; Sugimoto et al. 2016; Sztajzel 2004; Wei et al. 2014b; G. F. Wilson 2002). For example, investigators have shown that heart rate and heart rate variability can distinguish between downwind and takeoff/landing phases during flight (Scannella et al. 2018). Literature in neuroergonomics suggests a strong relationship between brain activity and user cognitive states (Ayaz et al. 2012b; Brouwer et al. 2012; Johnson et al. 2011; G. F. Wilson 2002; G. F. Wilson and Fisher 1995). For example, piloting tasks under low cognitive load have higher P300 amplitudes than when under high load (Causse et al. 2015). Frequency decomposition of brain waves into alpha, beta, theta, and delta bandwidths has also been associated with mental workload, wakefulness, sleeping, and deep sleep (Brouwer et al. 2012; Kumar and Kumar 2016). In flight simulations with pilots, these bandwidths have the potential to distinguish between varying flight tasks, e.g., takeoff, cruise, and tracking (G. F. Wilson 2002). Other researchers have proposed modeling EEG beyond the conventional bandwidths and event-related potentials (ERP); they suggest classifying EEG into metrics of engagement, distraction, and sleep onset

(Berka et al. 2007; Johnson et al. 2011). Despite various methods and application domains that support that physiological sensing can be used to predict workload, additional work is needed to understand how these patterns translate to more complex and cognitively demanding skills specific to RPA training and missions.

Exploratory studies on RPA and pilot operations have proposed several applications of workload sensing to individualize training and improve operator performance. In RPA mission simulations, pilot engagement (calculated using EEG) was predictive of scenario difficulty levels (C. Wilson et al. 2016). Several adaptive aiding approaches have also been proposed using physiological sensors (Afergan et al. 2014; Mark et al. 2017; Teo et al. 2019). Mark et al. (2017) proposed a neurofeedback-based training paradigm by adapting pilot task difficulty levels based on brain activity extracted from functional near-infrared spectroscopy (fNIRS). Afergan et al. (2014) demonstrated a real-time feedback technique that improved operator performance during unmanned aerial vehicle (UAV) simulation. Specifically, working memory load was measured in real time using fNIRS and used to dynamically adapt simulation demands through the addition/subtraction of aircraft units (Afergan et al. 2014). However, the study's focus on route-planning tasks that may not reflect the complexity of RPA operations which requires a pilot, sensor operator, and mission intelligence coordinator to work as a team. Despite these promising findings, the translation of physiological sensing to RPA training is currently limited due to the (1) inability to localize specific RPA task contributors and their effect on operator workload, e.g., audio/visual information processing, teamwork, search, recall, planning, and targeting, and (2) lack of clear physiological metrics that are relevant for predicting operator workload and performance in complex tasks and situations observed in RPA operations.

The purpose of this work is to determine the effectiveness and limitations of physiological sensing in high-fidelity simulated RPA operations. Specifically, this manuscript explores the ability of physiological metrics to distinguish operator workload between easy, medium, and hard scenario difficulty levels (Hypothesis 1) and between different types of RPA-specific events (Hypothesis 2).

2 Methods

A laboratory study was conducted to evaluate physiological response and workload during RPA simulations with pilot trainees.

Fig. 1 Snapshot of the RPA simulation environment as seen by participants tracking point of interest highlighted by red arrow (note that red arrow was not displayed during the simulation)

2.1 Participants

This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board. Informed consent was obtained from each participant. Twelve participants (10 males and 2 females) were recruited from the School of Aviation with a median of 40-h experience with RPA flight. Median experience with combined manned and unmanned flight was 170 h. Mean age was 24 years. Participants were compensated \$15/h for their participation.

2.2 RPA training simulation

The RPA training simulation was performed using the Predator/Reaper Research Integrated Combat Networked Environment (PRINCE), a high-fidelity RPA simulator developed at the Air Force Research Laboratory (Schill et al. 2014). As with live RPA operations, the PRINCE simulator requires two participants working together simultaneously; one as the pilot and one as the sensor operator. Each pair of participants attended in two experimental sessions together. For example, a participant will perform the role of the RPA pilot in one session then perform the role of the RPA sensor operator in the next session. The two sessions occurred on separate days, and the two-person team remained consistent for both sessions. Previous findings on the pilot role were presented at a conference (C. Wilson et al. 2016), and the present study focused on workload for the sensor operators. In addition, this work goes beyond scenario-level analysis and provides in-depth findings on the effects of RPA simulation events.

From subject matter expert (SME) interviews, scenarios representing three complexity levels (i.e., Easy, Medium, and Hard) were designed by varying tasks demands, time restrictions, and visual obstacles that obscured the points of interest (POIs) (Fig. 1). SMEs were experts that had previous experiences with RPA. The following steps were

Fig. 2 Locations and names of the EEG electrodes placement

9 of 9 electrode locations shown

followed for establishing the difficulty of the scenario (and further described in C. Wilson et al. 2016): (1) research team and SME discussed RPA skills (e.g., navigating (pilot) and tracking targets (sensor operator)), (2) the team and SME identified obstacles RPA's may encounter, (3) scenario tasks and obstacles were rated by SME from 1 to 5, (4) ratings were used by research team to develop scenarios, and (5) SMEs evaluated the scenarios again and any alterations suggested implemented. Because the subjects participated in two sessions (once as pilot and once as the sensor operator), two versions of the scenarios (Version A and Version B) with the same complexity level were created by SMEs to reduce potential learning effects between sessions.

From the SME interviews, scenarios overall centered in targeting tasks, e.g., locating waypoints and tracking POIs (C. Wilson et al. 2016), and the difficulty of these tasks were modulated by factors such as obstacles (e.g., no fly zones and visual obstructions), time limits, and target movement speeds. In all the scenarios, participants performed four tasks sequentially; each task had a specific point/target of interest. To complete these tasks, participants need to perform a combination of (1) receiving mission instructions from the “mission control” via headset, entering coordinates, acknowledging and executing the mission, flying the RPA to instructed locations, searching for POIs (e.g., vehicle), navigating around no-fly-zones, and tracking POIs.

2.3 Study design

There were three scenario complexity levels (easy, medium, and hard), and the order of the three scenario complexities was counterbalanced (six different scenario orders: 123, 132, 213, 231, 312, and 321) and randomly assigned to each pair of participants for the first session. Scenario version (version A or B) and order differed for participants' second session. Participants completed the second session 2 days after the first. Each session required approximately 2 h to complete.

2.4 Data collection

To capture physiological signals during the experiment, participants wore an electroencephalogram (EEG) system

(B-Alert X10, Advanced Brain Monitoring, Inc.). One channel of the EEG system collected electrocardiogram (ECG) data and the nine other channels measured brain activity located across the frontal, central, and parietal lobes (Fig. 2). Manufacturer provided Synapse cream was applied to each electrode. Physiological data were recorded using the NeuroBridge software developed by Intific, Inc. (Ramsey et al. 2015). Software recorded raw signals for all ten channels at 256 Hz. Although not implemented in this current study, other physiological metrics may also be helpful in monitoring RPA operators. For example, galvanic skin response, muscle activity, pupil diameter, and thermal imaging were other previously used physiological modalities (Charles and Nixon 2019; Wu et al. 2021; T. Zhang et al. 2021). However, the B-Alert system only had one extra channel (beyond the nine channels used for EEG), and we selected ECG for measurement.

Study researchers annotated start and stop times of the scenario, in addition to key events. Key events annotated include:

- 1) Visual obstructions that obscured targets and POIs (“Lost” events).
- 2) Observed crashes, no-fly-zone infractions, exceeding instruction guidelines/parameters, tracked incorrect targets, safety events, etc. (“Poor Performance” events).
- 3) Aborting task objective (“Abort” events)
- 4) Completing the task objective (“Reach” events).
- 5) Instances where the simulation crashed due to technical difficulties and the team had to restart and resume the study (“Pause” event).

Results focus on the five events: Lost, Poor Performance, Abort, Reach, and Pause events.

2.5 Data analysis

Custom scripts written in Matlab (r2018a, MathWorks, Inc.) extracted Neurobridge recordings and linked the physiological data streams with scenario task and event time markers. ECG signals were analyzed using Kubios HRV Premium software (v3.2) to derive five heart rate variability metrics: R-R interval (RR), standard deviation of RR intervals (SDNN), root mean square of successive RR interval differences (RMSSD), percentage of successive RRs > 50 ms (pNN50), and Low Frequency/High Frequency power ratio (LF/HF) (Tarvainen et al. 2014). All five metrics were calculated for the scenario analysis comparison (Hypothesis 1); however, only RR and SDNN were calculated for the event-level comparison (Hypothesis 2) due to a minimum of 30 s windows of ECG data for HRV calculation as recommended by Kubios, i.e., events windows were too brief.

From the EEG data, 36 response variables were calculated corresponding the four frequency bandwidths (alpha, low beta, high beta, and theta) for each of the nine channel locations. The open-sourced EEGLab (v14.1.1) (Delorme and Makeig 2004) software was used with Matlab to process and decompose EEG signals into four classically defined bands: theta band (4–8 Hz), alpha band (8–12 Hz), low beta band (12–16 Hz), and high beta band (16–25 Hz). Specifically, signals for each channel were referenced to their average and 0.5 Hz high pass FIR filter was applied. The CleanLine plugin (Mullen 2012) was used to remove line noise. Independent component analysis (ICA) was performed for each entire scenario, and components were rejected using the ADJUST plugin adapted for the B-Alert X10 system. ADJUST plugin detects artifacts based on the joint use of spatial and temporal features and optimized to capture blinks, eye movements, and generic discontinuities (Mognon et al. 2011).

Because each participant performed the scenarios at all complexity levels, a repeated-measures mixed-effects model was used to determine the impact of scenario complexity on physiological signals. Specifically, for scenario-level analysis (Hypothesis 1), the fixed effects were scenario difficulty, scenario order, and scenario version, repeated variable was order, and random effect was subjects. The impact of scenario events (Hypothesis 2) of Lost Sight, Poor Performance, Reach, and Abort (see “Data collection”) on physiological metrics were determined using mixed models that adjust for scenario difficulty and subject effects. Specifically, ECG data was averaged for 10 s immediately after the event following recent recommendations for psychophysiological research that 10-s windows can be sufficient for reliable estimates of time-domain HRV metrics (Laborde et al. 2017). EEG was windowed 3 s after the event to capture late event-related potential components and following on previous research indicating that 2.5-s windows were sufficient to classify differences in cognitive workload using EEG (Brouwer et al. 2012). In addition, EEG activity windows of 3 s or more can predict attention lapses in driving (Risser and Ware 1999). Therefore, EEG features were extracted from a 3-s time window. Five mixed models were similarly used to measure differences in physiological response before and after each event type (Hypothesis 2). The response variable was the difference in mean physiological signals before and after the event. The main effects for this model were the scenario difficulty and event type. Multiple comparisons of significant factors were adjusted with Bonferroni corrections. All statistical analyses were performed using SPSS Statistics software package (v22, IBM Corp.).

Table 1 Descriptive statistics of scenario lengths and sample sizes for each physiological measure by complexity level

Scenario difficulty	Sample size ECG	Sample size EEG	Mean (minutes)	Standard deviation (minutes)	Min (minutes)	Max (minutes)
Easy	9	10	23.1	3.4	20.2	31.5
Medium	8	8	25.6	5.3	19.6	36.9
Hard	8	9	21.6	3.9	12.6	27.1
Total	25	27	23.3	4.3	12.6	36.9

Fig. 3 Estimated marginal means and standard error for RR interval, standard deviation of RR intervals (SDNN), root mean square of successive RR interval differences (RMSSD), percentage of successive RRs > 50 ms (pNN50), and Low Frequency/ High Frequency power ratio (LF/ HF ratio) (an asterisk indicates significant differences between pairs after Bonferroni adjustment $p < 0.05$)

3 Results

Participants completed all scenario levels and attended both testing sessions. Since all participants ($n = 12$) performed the sensor operator role at each of the three complexity levels, a total sample size of 36 trials of physiological data were collected with ECG and EEG sensors. Results are summarized at the scenario level (Hypothesis 1). This is followed by results at the event level (Hypothesis 2), e.g., Lost and Reach.

3.1 Scenario-level analysis

Mean scenario times ranged from 20 to 37 min (Table 1). Although participants completed all scenarios, some physiological data were missing for some scenarios due to the software not recording file ($n = 1$), a faulty ECG connection ($n = 2$, affects ECG sample size only), the software stopping re- cording ($n = 3$), and mismatched ($n = 2$) or missing event markers ($n = 3$). As shown in Table 1, mean task times were shortest for the Hard scenario and longest for the Medium scenario. The Medium scenario was also observed to have the largest variance in task time.

Fig. 4 Mean and standard deviation for electroencephalography (EEG) channels and bands that differed across scenario complexity (an asterisk indicates significant differences between pairs after Bonferroni adjust- ment $p < 0.05$)

Out of the five heart rate variability (HRV) metrics, only LF/HF ratio differed between the three scenario difficulty levels (Fig. 3). RR intervals averaged 812 to 835 ms across the three scenario difficulty levels, and estimated marginal means for RR on Easy scenarios trended 23 ms longer than Medium scenarios ($p = 0.06$). For the LF/HF ratio, significant differences of 0.74 ($p = 0.031$) were observed between the medium and hard difficulty (Fig. 3).

Four EEG features (i.e., POz alpha, POz low beta, F3 al- pha, F4 alpha, and F4 theta bands) showed statistically signif- icant differences ($p < 0.05$) across scenario difficulty

Table 2 Frequency of events by scenario difficulty in the electrocardiography (ECG) and electroencephalography (EEG) datasets

Event	ECG data				EEG data			
	Easy	Medium	Hard	Total	Easy	Medium	Hard	Total
Pause	3	2	0	5	3	2	0	5
Lost Sight	14	10	15	39	16	10	15	41
Abort POI	8	9	2	19	9	9	2	20
Poor Performance	13	6	6	25	13	6	6	25
Reach POI	18	35	18	71	20	35	19	74

Fig. 5 Boxplots of standard deviation of RR intervals (SDNN) measured **Fig. 6** Comparison of mean RR interval (ms) before and after an event, where an asterisk indicates significant differences ($p < 0.05$) studies have argued that this interpretation is incorrect and that interpretation of LF/HF ratio remains controversial (Billman 2013; Shaffer et al. 2014)

levels. None of the central electrodes (Cz, C3, and C4) approached statistical significance. Similarly, high beta bands, regardless of channel location, did not differ between scenario difficulty. Subsequent multiple comparisons of significant features with Bonferroni adjustment showed pairwise differences in POz and F4 channels. Specifically, POz alpha and POz low beta bands were 47% and 41%, respectively, and were lower for Hard than Easy scenarios. For the

Fig. 7 Difference (after-before) and standard error in electroencephalogram (EEG) activity for channel locations and bandwidths that differed significantly ($p < 0.05$) by event type after adjusting for scenario difficulty

F4 alpha band, activity during the Hard scenario was 57% lower than the Easy scenario (Fig. 4).

3.2 Event-level analysis

Physiological signals were analyzed for each event. The frequency of these events observed in the ECG and EEG data is shown in Table 2. Simulation crashes (“Pause”) occurred five times in the entire study. Reaching POI was the most common event, followed by Lost Sight events. Poor Performance was observed more frequently on Easy scenarios than Medium and Hard.

The effects of observed events (Table 2) on HRV metrics were analyzed using a mixed model with scenario difficulty and event type as fixed effects and subject as a random effect (Fig. 5). There were no significant differences for event type on observed RR, but event type did impact observed SDNN ($p < 0.05$; Fig. 5). Mean SDNN after Pause events were 35 and 41 ms longer ($p < 0.05$) than Lost Sight and Abort events, respectively. Mean SDNN after Lost Sight, Abort, and Reach events were 24–37 ms shorter ($p < 0.01$) than the average SDNN during the entire scenario. Poor performance SDNN did not differ from other events.

The mixed model was also used to determine whether changes in ECG signals before and after an event differed (Fig. 6). For mean RR, only Abort and Reach events resulted in statistically different before and after ECG signals. Specifically, RR was 65 ms longer ($p < 0.05$) after an Abort

event than before during Hard scenarios, with no differences observed for Easy and Medium scenarios. Mean RR was 34 ms longer ($p < 0.05$) after a Reach event than before a Reach event for Medium scenarios. The change in SDNN was not statistically different for any events.

Analysis was conducted to compare EEG activity before and after an event type occurred using a mixed-effects model containing event type and scenario difficulty as main effects (Fig. 7). EEG activity decreased ($p < 0.05$) between 0.24 and 1.14 after Abort events for parietal and central channels in the alpha and low beta bandwidths. EEG activity increased ($p < 0.05$) by 0.15–0.72 in the beta bands (high and low) after Lost Sight events in both frontal and parietal channels. EEG activity similarly had a 0.34 increase ($p < 0.05$) in low beta activity after Pause events. Finally, decreases ($p < 0.05$) of 0.20–0.26 in EEG activity were observed after Reach events for the central and frontal channels at the high beta bandwidth.

4 Discussion

As technology, interface complexity, and military tactics advance, operation environments will become more complex, and the mental and physical workloads on military personnel and military teams will increase. Physiological sensors can continuously measure indices of physiological activation and provide an approach for inferring operator workload. In the RPA environment where demands can dynamically fluctuate, these sensors can potentially help distinguish between specific tasks and events. However, many metrics, 10 s after each event type and SDNN averaged across the entire scenario. Matching letters indicate statistically different pairs of event types, e.g., Pause events differed from Lost Sight and Abort techniques, and analysis methods can be derived from physiological sensing. As the findings in the current study show, physiological metrics have varying effectiveness in distinguishing scenario difficulty and RPA-specific events during high-fidelity RPA simulation.

4.1 Physiological signals at the scenario level

When utilizing physiological measures to study workload, a common approach is summarizing (e.g., mean, median, and maximum) the physiological data across the entire session, operation, or scenario (Abdelrahman et al. 2015; Ayaz et al. 2012a; Ayaz et al. 2012b; Böhm et al. 2001; Cinaz et al. 2013; Cinaz et al. 2010; Rieger et al. 2014; C. Wilson et al. 2016; Yu et al. 2017). When summarized across scenarios in the present study (Fig. 3), most of the ECG variables (i.e., heart rate, SDNN, RMSSD, and pNN50) were not effective in distinguishing scenario difficulty. Only the HRV metric LF/HF showed reduction between the medium and

hard difficulty levels. The LF/HF ratio has been proposed to reflect sympatho/vagal balance (Camm et al. 1996); however, recent studies have argued that this interpretation is incorrect and that interpretation of LF/HF ratio remains controversial (Billman 2013; Shaffer et al. 2014).

A recent systematic review on physiological sensors to detect mental workload found that HR and HRV metrics were the most commonly used modality (Charles and Nixon 2019); however, the evidence supporting HR and HRV as predictors for predicting workload was mixed. Some studies showed that ECG metrics (1) can distinguish workload between resting and task states, (2) showed mixed results distinguishing mental workload exercises (e.g., Stroop, n-back, and arithmetic), and (3) had limited success in distinguishing between different types of tasks. For example, comparing SDNN between rest and mental tasks (i.e., Mensa test), Taelman and others (2009) observed no differences in SDNN and LF/HF, but differences in RR and pNN50 (Taelman et al. 2009). In more advanced tasks like driving, some investigators similarly showed that SDNN failed to differentiate between varying difficulty levels during n-back memory recall tasks (Mehler et al. 2011); however, Mehler et al. (2011) did suggest that other time-domain HRV metrics (e.g., RMSSD) may be more useful in cognitive demand during driving, concluding that physiological measures need to be selected carefully depending on what aspects of workload are of interest. Specific to aviation, several studies focused on flight tasks and flight simulation (Dussault et al. 2004, 2005; Skibniewski et al. 2015; Wei et al. 2014a) and concluded that HRV was not as sensitive to varying cognitive demands as other physiological signals (G. F. Wilson 2002). From the systematic review (Charles and Nixon 2019), the general conclusion on HR and HRV was that their ability to distinguish mental workload was influenced by application domain, task duration, and task factors.

One contribution of this study was exploratory HRV findings specific to the RPA domain. Our findings showed limited effectiveness of HR and HRV for distinguishing difficulty in training simulations, which aligns with the current conclusions regarding HRV. Although some of the previously mentioned studies distinguished workload with ECG, several moderating factors likely influenced the prediction of workload in the present study. First, the RPA simulation tasks were complex, dynamically changing, which differed from the mental skills and secondary tasks of previous work. Second, ECG was shown to reliably predict large effect sizes (e.g., rest vs. task or large differences in task difficulty). The variations in RPA task difficulty were designed for non-expert RPA operators and may not be distinguishable. Finally, scenarios in this study averaged 23 min in length, and averaging ECG signals for the entire scenario may not adequately capture the many team interactions and events that contribute to perceived workload and task complexity.

In the next section, the applications of HRV for distinguishing workload at the event level will be discussed.

A consistent relationship observed was decreased EEG signals at the alpha bandwidth between Easy and Hard scenarios in the parietal and frontal EEG channels. Many studies suggest that the alpha band represents states of relaxation and development of drowsiness, with alpha power decreasing with mental load (Harmony, 1999; Klimesch 1999; Liu et al. 2015; Liu et al. 2017; Smith and Gevins 2005; Spüler et al. 2016; S. Wang et al. 2016; G. F. Wilson and Fisher 1995; Young et al. 2014). Work in flight simulation observed reductions in alpha band activity between baseline and flight tasks (G. F. Wilson 2002). A review of EEG in driving and piloting work also suggested that alpha band activity decreased during concurrent task execution, working memory load, and cognitively demanding tasks in the parietal regions (Borghini et al. 2014). More specifically, the review suggested prefrontal areas target decision-making while parietal areas were related to processing space relations. In RPA simulations, the complexity of the visual search and task demands were primary contributors to scenario difficulty, and observed alpha band activity may suggest that the alpha band activity (frontal and parietal channels in our results) can be used to distinguish workload between scenario difficulties in RPA environments.

In contrast to the alpha band, other EEG bands (e.g., theta and beta) had limited effectiveness in distinguishing between the three scenario difficulty levels (Fig. 4). For example, increases in theta band activity had been proposed to reflect time pressure, focused attention, and mental workload (Borghini et al. 2014; Charles and Nixon 2019; Kumar and Kumar 2016). Although these were all key elements in our RPA simulations, no differences were observed. An EEG review acknowledged that reviewed studies were mixed on whether theta band activity can distinguish several levels of complexity in multi-tasking environments (Borghini et al. 2014). Since the RPA simulations in the present study required sustained attention and multi-tasking, these factors may explain the lack of significant differences on the theta band variables. In addition, fatigue may also attenuate the theta and beta band activity. Specifically, transitions from mental workload to mental fatigue were characterized by decreases in the beta band power (Zhao et al. 2012), and this was observed in the POz channel (Fig. 4). In the present study, each scenario can last up to 37 min, and the current techniques using mean bandwidth activity may not be sufficient in capturing the dynamical changes mental demands during the entire scenario (i.e., a Hard scenario can contain segments of low and high workloads).

In summary, both EEG and HRV responses appeared to reflect physiological behaviors observed in previous literature when applied to varying levels of scenario difficulty. However, statistically significant and consistent changes

were only observed with EEG data. This suggests that ECG is not as effective as EEG measurements, and EEG measurements may be best able to distinguish differences in scenario difficulty when assessing workload at the scenario level.

4.2 Physiological signals by events

Extracting workload responses for specific events focused on two methods: (1) comparing responses immediately after an event with responses during other event markers and (2) comparing workload responses at time windows before and after events.

The SDNN metric was significantly lower for most events (i.e., Lost Sight, Abort, and Reach) than the scenario average (Fig. 5). A recent meta-analysis showed that SDNN decreased with stress (Kim et al. 2018). In high-fidelity flight simulation, lower SDNN were observed during flight simulation than before the flight, with the lowest SDNN observed during difficult flight maneuvers, e.g., steep turns and circle to land (Cao et al. 2019). The lower SDNN during events suggests that workload during RPA events were higher than average workload experienced throughout the scenario during RPA operations. Potential implications of this finding can be for detecting key events that impact operators by identifying instances of faster SDNN response, e.g., when operators encounter targets obscured by buildings or smoke (“Lost Sight event”). In addition, SDNN was able to distinguish between Pauses and events like Lost Sight and Abort POI. As pauses in the scenario occurred during technical difficulties, the participants were not under any workload while researchers attempted to restart the simulation. Differences observed suggest that lower SDNN can indicate RPA events or tasks that increase workload (e.g., Lost Sight).

Although the RR metric was not able to distinguish between events like SDNN, before/after comparisons showed that RR changed significantly for abort and reach events. Specifically, mean RR increased immediately after an Abort and Reach event (Fig. 6). Because RR is inversely related to heart rate, this can be interpreted as heart rate slowing after a Reach or Abort event occurred. Both Reach and Abort events represent the conclusion of an objective, either failure (Abort) or success (Reach). The slowing of heart rate may reflect the recovery effect after the end of an objective but before another is assigned or the scenario ended. These findings suggest RR intervals may be able to detect physiological response after ending an objective. However, HRV metrics were not able to distinguish performance-related information (e.g., success/failure or observed poor performance events) which may limit physiological signals application in automatically assessing operator performance.

EEG activity showed promise in detecting changes using the “before and after” analysis approach. For events that

signaled the conclusion of an objective (i.e., reach and abort), beta band activity immediately after the event was reduced. Since participants were actively engaged with an RPA objective before these events and then await an objective immediately after, there is likely higher mental workload prior to Reach or Abort, and lower workload immediately after. As mentioned previously, this aligned well with the literature suggesting increased beta activity with the stimulus that increased workload (Kumar and Kumar 2016). In contrast to physiological activation patterns observed for reach and abort events, lost sight events elicited a different pattern for the beta band. Specifically, lost sight occurred while participants were actively tasked with an objective. Upon losing sight of the POI (e.g., simulated fog or buildings that obscured line-of-sight), workload and alertness were expected to increase as participants actively attempt to overcome the obstacle. Aligned with this expectation, beta band activity increased at both frontal and parietal channels during lost sight events. In summary, RPA events can be distinguished with physiological signals, and the observations appeared aligned with physiological signals observed in other tasks and domains (Kumar and Kumar 2016; Liu et al. 2017; Young et al. 2014).

4.3 Limitations and future work

The current research explored physiological responses of pilot trainees in the sensor operator role in these scenarios. Although these initial findings suggest that physiological sensing can provide relevant information on sensor operator response to scenario difficulty and events for RPA training, further work is needed to enable translation of findings to RPA operator training. Specifically, many studies noted that age, gender, and experience differences can influence physiological responses in both ECG and EEG and larger sample sizes and experienced RPA operators to confirm generalizability of the current findings. However, we believe the characteristics of the targeted population of RPA trainees are represented in this study's population at the aviation school. In addition, our study population seem to also reflect current Air Force demographics, i.e., ~21% females (Center 2020). However, further balanced and controlled studies are needed to better understand the influence of gender in RPA physiological monitoring.

Although it is highly desirable to detect the physiological patterns leading up to and after Poor Performance events, these events did not approach statistical significance in any analysis. One potential explanation is that poor performance events were observer-annotated, and more objective and reliable annotation may improve sensor ability to distinguish signatures of poor performance. Another method would be to examine moments where physiological activation increased and exploring RPA tasks contributors to the

increase. Many metrics and methods exist for examining physiological metrics during training, and further work can explore additional metrics (e.g., eye-tracking, engagement index (Berka et al. 2004), sleep onset, and data fusion (Zhou et al. 2020)) and optimize time windows for workload prediction. For example, physiological response times or delays in response times can impact inferences drawn from these signals, especially for the short time windows of the studied RPA events. Additional data analytics experiments are needed to account for these latent temporal effects. Beyond heart rate and brain activity sensors, other physiological metrics may also be helpful in monitoring RPA operators, and many studies have used galvanic skin response, muscle activity, pupil diameter, and thermal imaging as physiological monitoring modalities (Charles and Nixon 2019). This study focused on EEG and ECG as previous work in multiple domains suggest that these modalities are promising. Moving towards the long-term goal of more effective training systems, task design, and trainee learning, additional work is needed to identify and measure the contributors to workload in RPA operations with more controlled events and systematic adjustments to work demands in order to model a physiology-based workload/performance curve.

In conclusion, the present work demonstrates that physiological sensing can distinguish workload in the sensor operator role for RPA training scenarios. More specifically, various modalities (EEG and ECG) had varying effectiveness in distinguishing scenario difficulty and RPA events, suggesting a hybrid sensing approach may provide more insight than just using one modality. This study bolsters support for the notion that physiological sensing can offer specific, objective, and real-time insight into trainee responses to training elements, events, and objectives to help educators and training developers better design and execute effective dynamic training programs.

Acknowledgements The authors would like to acknowledge the Air Force Research Laboratory's Summer Faculty Fellowship Program and the Airman Systems Directorate for their support of this work. The authors would also like to thank research team members for all their efforts in conducting this study.

Funding The authors would like to acknowledge the Air Force Research Laboratory's Summer Faculty Fellowship Program for their support of the work.

Data availability De-identified datasets may be available upon institutional approval.

Code availability Not applicable.

This research complied with the American Psychological Association Code of Ethics and was approved by the Institutional Review Board. Informed consent was obtained from each participant.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

- Abdelrahman AM, Bingener J, Yu D, Lowndes BR, Mohamed A, McConico AL, Hallbeck MS (2015) Impact of single-incision laparoscopic cholecystectomy (SILC) versus conventional laparoscopic cholecystectomy (CLC) procedures on surgeon stress and workload: a randomized controlled trial. *Surg Endosc* 30(3):1205–1211. <https://doi.org/10.1007/s00464-015-4332-5>
- Afegan D., Peck EM, Solovey ET, Jenkins A, Hincks SW, Brown ET, ... Jacob RJ (2014). Dynamic difficulty using brain metrics of workload. Paper presented at the Proceedings of the 32nd annual ACM conference on Human factors in computing systems
- Akash K, Hu W-L, Jain N, Reid T (2018) A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8(4):1–20
- Ayaz H, Çakir MP, İzzetoglu K, Curtin A, Shewokis PA, Bunce SC, Onaral B (2012a) Monitoring expertise development during simulated UAV piloting tasks using optical brain imaging. Paper presented at the Aerospace Conference, 2012 IEEE
- Ayaz H, Shewokis PA, Bunce S, Izzetoglu K, Willems B, Onaral B (2012b) Optical brain monitoring for operator training and mental workload assessment. *Neuroimage* 59(1):36–47. <https://doi.org/10.1016/j.neuroimage.2011.06.023>
- Berka C, Levendowski DJ, Cvetinovic MM, Petrovic MM, Davis G, Lumicao MN, Zivkovic VT, Popovic MV, Olmstead R (2004) Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *Int J Hum Comput Interact* 17(2):151–170
- Berka C, Levendowski DJ, Lumicao MN, Yau A, Davis G, Zivkovic VT et al (2007) EEG correlates of task engagement and mental work- load in vigilance, learning, and memory tasks. *Aviat Space Environ Med* 78(5):B231–B244
- Billman GE (2013) The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance. *Front Physiol* 4:26. <https://doi.org/10.3389/fphys.2013.00026>
- Böhm B, Rötting N, Schwenk W, Grebe S, Mansmann U (2001) A prospective randomized trial on heart rate variability of the surgical team during laparoscopic and conventional sigmoid resection. *Arch Surg* 136(3):305–310
- Borghini G, Astolfi L, Vecchiato G, Mattia D, Babiloni F (2014) Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neurosci Biobehav Rev* 44:58–75. <https://doi.org/10.1016/j.neubiorev.2012.10.003>
- Brouwer A-M, Hogervorst MA, Van Erp JBF, Heffelaar T, Zimmerman PH, Oostenveld R (2012) Estimating workload using EEG spectral power and ERPs in the n-back task. *J Neural Eng* 9(4):045008
- Camm AJ, Malik M, Bigger J, Breithardt G, Cerutti S, Cohen RJ et al (1996) Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. *Eur Heart J* 17(3):354–381
- Cao X, MacNaughton P, Cadet LR, Cedeno-Laurent JG, Flanagan S, Vallarino J et al (2019) Heart rate variability and performance of commercial airline pilots during flight simulations. *Int J Environ Res Public Health* 16(2). <https://doi.org/10.3390/ijerph16020237>
- Catherwood D, Edgar GK, Nikolla D, Alford C, Brookes D, Baker S, White S (2014) Mapping brain activity during loss of situation awareness: an EEG investigation of a basis for top-down influence on perception. *Hum Factors* 56(8):1428–1452
- Causse M, Fabre E, Giraudet L, Gonzalez M, Peysakhovich V (2015) EEG/ERP as a measure of mental workload in a simple piloting task. *Procedia Manuf* 3:5230–5236. <https://doi.org/10.1016/j.promfg.2015.07.594>
- Center, A. F. P. (2020). Military demographics: June
- Charles RL, Nixon J (2019) Measuring mental workload using physiological measures: a systematic review. *Appl Ergon* 74:221–232. <https://doi.org/10.1016/j.apergo.2018.08.028>
- Chen JY, Barnes MJ, Harper-Sciarni M (2010) Supervisory control of unmanned vehicles. Retrieved from
- Cinaz B, La Marca R, Arnrich B, Tröster G (2010) Monitoring of mental workload levels. Paper presented at the International Conference on e-Health. sn: IADIS
- Cinaz B, Arnrich B, La Marca R, Tröster G (2013) Monitoring of mental workload levels during an everyday life office-work scenario. *Pers Ubiquit Comput* 17(2):229–239. <https://doi.org/10.1007/s00779-011-0466-1>
- Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J Neurosci Methods* 134(1):9–21
- Dong, S.-Y., Kim, B.-K., Lee, K., & Lee, S.-Y. (2015). A preliminary study on human trust measurements by EEG for human-machine interactions
- Dussault C, Jouanin J-C, Guezennec C-Y (2004) EEG and ECG changes during selected flight sequences. *Aviat Space Environ Med* 75(10):889–897
- Dussault C, Jouanin J-C, Philippe M, Guezennec C-Y (2005) EEG and ECG changes during simulator operation reflect mental workload and vigilance. *Aviat Space Environ Med* 76(4):344–351
- Fu C, Zhang Z, He J, Huang S, Qiu J, Wang Y (2018) Brain dynamics of decision-making in the generalized trust game: evidence from ERPs and EEG time-frequency analysis. *Acta Psychol Sin* 50(3):317–326
- Harmony TA, Fernández TA, Silva J, Bosch J, Valdés P, Fernández-Bouzas A et al (1999) Do specific EEG frequencies indicate different processes during mental calculation? *Neurosci Lett* 266(1):25–28
- Johnson RR, Popovic DP, Olmstead RE, Stikic M, Levendowski DJ, Berka C (2011) Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biol Psychol* 87(2):241–250
- Kaur A, Chaujar R, Chinnadurai V (2019) Effects of neural mechanisms of pretask resting EEG alpha information on situational awareness: a functional connectivity approach. *Hum Factors* 62:1150. <https://doi.org/10.1177/0018720819869129>
- Kim HG, Cheon EJ, Bai DS, Lee YH, Koo BH (2018) Stress and heart rate variability: a meta-analysis and review of the literature. *Psychiatry Investig* 15(3):235–245. <https://doi.org/10.30773/pi.2017.08.17>
- Klimesch W (1999) EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res Rev* 29(2):169–195. [https://doi.org/10.1016/S0165-0173\(98\)00056-3](https://doi.org/10.1016/S0165-0173(98)00056-3)
- Kumar N, Kumar J (2016) Measurement of cognitive load in HCI systems using EEG power spectrum: an experimental study. *Procedia Comput Sci* 84:70–78. <https://doi.org/10.1016/j.procs.2016.04.068>
- Laborde S, Mosley E, Thayer JF (2017) Heart rate variability and cardiac vagal tone in psychophysiological research – recommendations for experiment planning, data analysis, and data reporting. *Front Psychol* 8:213. <https://doi.org/10.3389/fpsyg.2017.00213>
- Liu Y, Ayaz H, Onaral B, Shewokis PA (2015) Neural adaptation to a working memory task: a concurrent EEG-fNIRS Study
- Liu Y, Ayaz H, Shewokis PA (2017) Multisubject “Learning” for mental workload classification using concurrent EEG, fNIRS, and

- physio- logical measures. *Front Hum Neurosci* 11:389. <https://doi.org/10.3389/fnhum.2017.00389>
- Mark J, Thomas N, Kraft A, Casebeer WD, Ziegler M, Ayaz H (2017) Neurofeedback for personalized adaptive training
- McDuff D, Gontarek S, Picard R (2014) Remote measurement of cognitive stress via heart rate variability. Paper presented at the Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE
- Mehler B, Reimer B, Wang Y (2011) A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload
- Mehler B, Reimer B, Coughlin JF (2012) Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task: an on-road study across three age groups. *Hum Factors* 54(3):396–412
- Mognon A, Jovicich J, Bruzzone L, Buiatti M (2011) ADJUST: an automatic EEG artifact detector based on the joint use of spatial and temporal features. *Psychophysiology* 48(2):229–240
- Moses ZB, Luecken LJ, Eason JC (2007) Measuring task- related changes in heart rate variability. Paper presented at the 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society
- Mullen T (2012) CleanLine EEGLAB plugin. San Diego, CA: Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC)
- Ramsey L, Bloomer J, Watz E, Neubauer P (2015) Pilot test of the Predator Research Integrated Networked Combat Environment Workload Indices via Neurobridge Kit. Warfighter Readiness Science And Technology Program
- Rieger A, Fenger S, Neubert S, Weippert M, Kreuzfeld S, Stoll R (2014) Psychophysical workload in the operating room: primary surgeon versus assistant. *Surg Endosc*:1–9
- Risser MR, Ware JC (1999) Driving simulation with EEG monitoring in normals and obstructive sleep apnea patients
- Scannella S, Peysakhovich V, Ehrig F, Lepron E, Dehais F (2018) Assessment of ocular and physiological metrics to discriminate flight phases in real light aircraft. *Hum Factors* 60(7):922–935. <https://doi.org/10.1177/0018720818787135>
- Schill, N. P., Leah, J., Gyovai, B. L., Joralmon, D. Q., Schneck, A. J., & Woudstra, D. A. (2014). Operational alignment in predator training research. Retrieved from
- Shaffer F, McCraty R, Zerr CL (2014) A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. *Front Psychol* 5:1040
- Skibniewski FW, Dziuda Ł, Baran PM, Krej MK, Guzowski S, Piotrowski MA, Truszczyński OE (2015) Preliminary results of the LF/HF ratio as an indicator for estimating difficulty level of flight tasks. *Aerosp Med Human Perform* 86(6):518–523
- Smith ME, Gevins A (2005) Neurophysiologic monitoring of mental workload and fatigue during operation of a flight simulator. Paper presented at the Biomonitoring for Physiological and Cognitive Performance during Military Operations
- Spüler M, Walter C, Rosenstiel W, Gerjets P, Moeller K, Klein E (2016) EEG-based prediction of cognitive workload induced by arithmetic: a step towards online adaptation in numerical learning. *ZDM* 48(3):267–278. <https://doi.org/10.1007/s11858-015-0754-8>
- Sugimoto I, Kitamura K, Murai K, Wang Y, Wang J (2016) Study on relation between operator and trainee's mental workload for ship maneuvering simulator exercise using heart rate variability. Paper presented at the Systems, Man, and Cybernetics (SMC), 2016 IEEE International Conference on
- Sztajzel J (2004) Heart rate variability: a noninvasive electrocardiographic method to measure the autonomic nervous system. *Swiss Med Wkly* 134(35–36):514–522
- Taelman J, Vandeput S, Spaepen A, Van Huffel S (2009) Influence of mental stress on heart rate and heart rate variability
- Tarvainen MP, Niskanen J-P, Lipponen JA, Ranta-Aho PO, Karjalainen PA (2014) Kubios HRV—heart rate variability analysis software. *Comput Methods Prog Biomed* 113(1):210–220
- Teo G, Matthews G, Reinerman-Jones L, Barber D (2019) Adaptive aiding with an individualized workload model based on psychophysiological measures. *Human Intelligent Systems Integration*. <https://doi.org/10.1007/s42454-019-00005-8>
- Wang S, Gwizdzka J, Chaovalitwongse WA (2016) Using wireless EEG signals to assess memory workload in the \$ n \$-Back task. *IEEE Trans Human Mach Sys* 46(3):424–435
- Wang M, Hussein A, Rojas RF, Shafi K, Abbass HA (2018) EEG-based neural correlates of trust in human-autonomy interaction
- Wei Z, Zhuang D, Wanyan X, Zhang H, Liu C (2014a) A theoretical model of mental workload in pilots based on multiple experimental measurements. In: Harris D (ed) *Engineering psychology and cognitive ergonomics: 11th International Conference, EPCE 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22–27, 2014. Proceedings*. Springer International Publishing, Cham, pp 104–113
- Wei Z, Zhuang D, Wanyan X, Zhang H, Liu C (2014b) A theoretical model of mental workload in pilots based on multiple experimental measurements. Paper presented at the International Conference on Engineering Psychology and Cognitive Ergonomics
- Wilson GF (2002) An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *Int J Aviat Psychol* 12(1):3–18. https://doi.org/10.1207/S15327108IJAP1201_2
- Wilson GF, Fisher F (1995) Cognitive task classification based upon topographic EEG data. *Biol Psychol* 40(1–2):239–250. [https://doi.org/10.1016/0301-0511\(95\)05102-3](https://doi.org/10.1016/0301-0511(95)05102-3)
- Wilson C, Ramsey L, Bloomer J (2016) Multimodal assessment of workload for predator training research. *Interservice/Industry Training, Simulation, and Education Conference*
- Wu C, Cha J, Sulek J, Sundaram CP, Wachs J, Proctor RW, Yu D (2021) Sensor-based indicators of performance changes between sessions during robotic surgery training. *Appl Ergon* 90:103251. <https://doi.org/10.1016/j.apergo.2020.103251>
- Yeo LG, Sun H, Liu Y, Trapsilawati F, Sourina O, Chen C-H, Ang WT (2017) Mobile EEG-based situation awareness recognition for air traffic controllers
- Young MS, Brookhuis KA, Wickens CD, Hancock PA (2014) State of science: mental workload in ergonomics. *Ergonomics* 58(1):1–17. <https://doi.org/10.1080/00140139.2014.956151>
- Yu D, Dural C, Morrow MMB, Yang L, Collins JW, Hallbeck S, ... Forsman M (2017) Intraoperative workload in robotic surgery assessed by wearable motion tracking sensors and questionnaires. *Surg Endosc* 31(2):877–886. <https://doi.org/10.1007/s00464-016-5047-y>
- Zhang W, Feltner D, Shirley J, Swangnetr M, and Kaber D (2016) Unmanned aerial vehicle control interface design and cognitive workload: a constrained review and research framework.
- Zhang T, Yang J, Liang N, Pitts B, Prakah-Asante K, Curry R, ... Yu D (2021) Physiological measurements of situation awareness: a systematic review. *Hum Factors*. <https://doi.org/10.1177/0018720820969071>
- Zhao C, Zhao M, Liu J, Zheng C (2012) Electroencephalogram and electrocardiograph assessment of mental fatigue in a driving simulator. *Accid Anal Prev* 45:83–90
- Zhou T, Cha JS, Gonzalez G, Wachs JP, Sundaram CP, Yu D (2020) Multimodal physiological signals for workload prediction in robot-assisted surgery. *J Hum Robot Interact* 9(2):Article 12. <https://doi.org/10.1145/3368589>