Estimating Cognitive Load and Cybersickness of Pilots in VR Simulations via Unobtrusive Physiological Sensors

G. S. Rajshekar Reddy[1](✉), Cara A. Spencer[1], Kevin Durkee[2], Brennan Cox[3],
Olivia Fox Cotton[2], Sheila Galbreath[2], Sarah Meyer[2], Michael Natali[3],
Todd Seech[3], Gabriella Severe-Valsaint[4], Gavin Zimmerman[1],
and Leanne Hirshfield[1]

[1] Institute of Cognitive Science, University of Colorado, Boulder, CO, USA
{rajshekar.gudasubhash,cara.spencer,gavin.zimmerman,
leanne.hirshfield}@colorado.edu
[2] Aptima, Inc., Dayton, OH, USA
{kdurkee,ofox,sgalbreath,smeyer}@aptima.com
[3] U.S. Navy, Wright-Patterson Air Force Base, Dayton, OH, USA
michael.w.natali.mil@us.navy.mil, todd.seech1@navy.mil
[4] Naval Air Warfare Center Training Systems Division, Orlando, FL, USA
gabriella.severe-valsaint.civ@us.navy.mil

Abstract. Predicting real-time estimates of cognitive load in pilots assists intelligent flight systems in alleviating high workloads, thereby averting accidents and directly impacting safety in aviation. Virtual Reality (VR) flight simulations provide an immersive stage to evaluate physiological measures and identify their cognitive correlates. In this work, unobtrusive sensors such as eye-tracking, pupillometry, and photoplethysmography (PPG) record physiological data while six participants perform six flying tasks of varying complexity in VR. The extracted feature sets such as pupil diameter change, number of fixations and saccades, and heart rate variability (HRV) are compared to the Pilot Inceptor Workload (PIW) measures, specifically duty cycle and aggressiveness. The PIW, number of saccades, and the self-reported workload measures were significantly affected by the tasks. However, the number of saccades measure demonstrated a significant negative correlation with the PIW's measures, contradicting prior work. The remaining feature sets, including the pupil diameter change and the number of fixations, display a nearly identical trend to the PIW measure, though no significance was detected.

Keywords: Cognitive load · Cybersickness · Virtual reality · Aviation

1 Introduction

Modern-day aviation involves interfacing with incredibly complex technologies, and simulations help train pilots to operate these complex technologies efficiently.

J. Y. C. Chen and G. Fragomeni (Eds.): HCII 2022, LNCS 13318, pp. 251–269, 2022.
https://doi.org/10.1007/978-3-031-06015-1_18

For decades, flight simulators have been used in commercial pilot training and military aviation. Nevertheless, there are several significant challenges with traditional flight simulators: (i) the high costs to acquire, operate and maintain, (ii) large space requirements, and (iii) their limited capabilities that are specific to a particular aircraft.

These challenges make them a liability and hamstring the user in ways that VR training does not. VR has presented itself as a promising modality to offer a more immersive and adaptable simulation experience at a fraction of the cost of traditional flight simulators. VR flight training tools provide the opportunity to evaluate aviation systems and human performance through various new and informative approaches that improve upon the science of training effectiveness and human system integration.

Furthermore, piloting fighter aircraft and performing mid-air maneuvers under high G's are complex multi-tasking and cognitively demanding activities requiring working memory to satisfy task demands [30]. Simulating such resource-intensive VR scenarios and accurately predicting the pilots' cognitive state is of crucial importance for many reasons: (i) using VR can save money, time, and machinery (ii) provide pilots with enhanced learning experiences, (iii) provide instructors with guidance and evaluation metrics to best help students, (iv) and lead to the prevention of declines in performance through training flight assistance systems.

With physiological sensors integrated into VR Head-mounted Displays (HMDs) such as eye tracking and PPG, it is now possible to objectively estimate the users' mental workload in real-time. These unobtrusive and non-invasive measures are of particular value in complex environments, such as the cockpit, because they typically do not increase cognitive load and do not require shifting attention from the complex task at hand. In addition, obtrusive sensors may hinder the user's performance and delay critical reaction times which are crucial in aviation.

Eventually, the data from these measures can be used to construct a machine learning classifier such that flight systems are better informed when to assist the pilot, ultimately paving the way to safer aviation.

However, a significant downside to using immersive simulators is cybersickness caused by the visual-vestibular conflict. Users tend to feel nauseous, dizzy, and generally disoriented when they experience moving visual stimuli while the vestibular system in the inner ear cannot sense actual motion [32]. Real-time biomarkers such as electrodermal activity, respiration, PPG, and eye movement characteristics provide the necessary physiological measures closely related to cybersickness [9,28]. Identifying the precise correlations of these measures would enable future VR systems to mitigate cybersickness symptoms in real-time and consequently improve the VR experience.

Considering all these benefits of real-time prediction of the pilot's cognitive load and cybersickness, we undertook a study to examine the respective correlates and extend previous work by validating the feasibility of using unobtrusive physiological sensors in VR. We built a testbed to induce cognitive load and cybersickness using the Prepar3D [34] flight simulator, which entails performing six flying tasks of various complexity in succession. Specifically, the tasks involve flying an aircraft through virtual gates while executing maneuvers in between. Participants comprise university students who operate a trainer aircraft using a Hands-on Throttle-and-Stick (HOTAS). While the participants perform the task, a suite of physiological sensors capture various data metrics like pupil diameter change, number of fixations and saccades, and HRV; and are described further in Sect. 3.2. We then investigate the correlations of these metrics to the PIW's [18] parameters, namely duty cycle and aggressiveness, which serve as a reliable ground truth estimate of task complexity, and hence workload.

The key contributions of this paper are:

- Validating the use of unobtrusive physiological sensors in VR to estimate cognitive load in flight training scenarios.
- Examining correlations between behavioral and physiological parameters, thus corroborating prior work indicating that pupil dilation and fixations are a reliable workload indicator while also contradicting prior work by demonstrating saccades to be negatively correlated with workload.
- Designing and constructing a VR flight simulation testbed that induces cognitive load and cybersickness through a series of six flying tasks of varying complexity.

In the subsequent sections, we discuss other prior work in estimating a pilot's cognitive load, followed by a thorough description of the study undertaken in Sect. 3. Sections 4 and 5 report the results from the study and their interpretations. We then conclude and outline directions for future work.

2 Background

Mental overload plays a significant role in aviation safety and must be mitigated. It can be assessed through various means such as self-reported measures like the NASA-Task Load Index (TLX) [20] and the Bedford Workload Scale [46], or physiological measures. However, self-reported measures are limiting as humans may not always accurately document their physiological state. For instance, a study showed that motorists were inaccurate in their self-reported vigilance measures compared to the objective physiological measures [48]. In addition to being a real-time estimate, researchers have proved that physiological measures are more sensitive and have employed various physiological-based indices to estimate cognitive states; Lohani et al. [35] provide a detailed review of the measures used, including heart rate and HRV, blood pressure, skin conductance, and pupillometry, among others.

Grant and others [15] introduced a research tool, TOME, to collate and examine real-time neurophysiological data and predict workload. The TOME server analyzes the sensor streams and generates alerts, including the inferred cognitive states of the user. They validated their tool through a flight-sim study where workload is assessed through a suite of sensors measuring blood flow in the frontal cortex (fNIRS), respiration rate, heart rate, galvanic skin response, and pupil dilation. The tasks involved flying a plane while maintaining specific altitude levels under varying weather conditions. They found that pupil dilation correlated highly with task difficulty and was the strongest workload indicator.

In their study, Huttunen et al. [25] estimate cognitive load by observing prosodic features in the pilots' speech. They recorded fifteen military pilots' speech patterns and utterances while they performed a simulated combat flight. The tasks required participants to perform tactical maneuvers, beyond-visual-range interceptions, and so on while serving as a leader of a pair of F/A-18 jets. They showed that the mean fundamental frequency and vocal intensity increased as a function of cognitive load, establishing speech prosody as a viable indicator of cognitive state.

Babu [2] investigated the correlation between PIW against three factors of a pilot's workload: flying quality, secondary task difficulty, and the closing-opposing boundary size levels. The investigation involved a simulator study and an in-flight study, where military pilots performed a pitch-tracking task of a moving target. All the three factors listed above had significant effects on the duty cycle and aggressiveness, the two variables representing PIW. It was also revealed that PIW highly correlated with the Workload Buildup Flight Test Technique, which is based on the Boundary Avoidance Theory [16,17], thereby establishing PIW as a reliable ground truth measure of pilot workload.

Heart rate and HRV are common measurements for flight simulator studies due to their sensitivity to autonomic driven states, especially sympathetic arousal which prepares the central nervous system for activation [45,49]. Villafaina et al. [58] monitored HRV as an indicator of autonomic responses induced by combat flight maneuvers and when combined with the Electroencephalography (EEG) recordings, they were able to discriminate which maneuvers required more attention and which were more mentally demanding. Itoh et al. [26], with similar measures to the current study, used HRV and subjective rankings to determine that mental demand was not different between piloting plane models in their simulators. Mohanavelu et al. [39], employed visibility manipulations and secondary tasks to examine cognitive workload and using an array of HRV metrics with subjective measures, they were able to determine which condition elicited higher cognitive workload.

Chen and Epps [6] look at eye activity to measure cognitive load changes and improve patient care in a healthcare setting. A series of arithmetic-based tasks indicated that pupil features were the most robust discriminator of cognitive load followed by cumulative blinks. They further revealed that saccade amplitude also increased with increased task difficulty. Krejtz et al. [29] analyze whether microsaccades and pupil diameter can serve as a reliable discriminant of cognitive load. Microsaccades are involuntary jerk-like eye movements that typically happen during prolonged fixations [61]. Their study involving three difficulty levels of mental calculation tasks revealed a significant effect of the three tasks on the microsaccade magnitude, but only a marginal main effect on the microsaccade rate. Detecting microsaccades requires a high-frequency eye tracker, and since we are limited by 120 Hz eye tracker in the VR HMD, we do not investigate it in this study.

The role of ocular parameters in estimating a pilot's cognitive load was investigated by Babu et al. [3]. Specifically, they analyze pupil dilation and eye movements such as fixations, saccades, and saccadic intrusions of military pilots during different flight phases and maneuvers. It was found that all parameters were significantly different for the three task difficulty levels. Moreover, the number of fixations and saccades strongly correlated with variations of PIW in the second and third tasks. They further undertook an in-flight study that revealed that the fixation rate was higher during take-off and landing than while cruising. However, our study contradicts this work regarding saccades being a positive workload indicator.

Hebbar et al. [21] explored the correlations between performance-based metrics and physiological measures like EEG and ocular parameters in pilots. Their flight-simulation study involved tasks with three difficulty conditions, the final one requiring participants to perform a secondary selection task on a multi-function display while maintaining their altitude within a range of 1000 ft. They found that the tasks significantly impacted the EEG signal in the low-beta and theta bands. They also found a significant impact on ocular parameters such as the Nearest Neighbour Index (NNI), which represents the distribution pattern of fixations, and the L1 Norm of Spectrum (L1NS) derived from the pupil dilation signal. Moreover, they established a significant correlation between L1NS and NNI and observed positive correlations between duty cycle and EEG and between duty cycle and the ocular parameters.

In summary, several studies have shown that physiological indicators, in particular ocular and heart rate parameters have been proven to be a valuable and pragmatic indicator of task complexity and consequently, cognitive state. However, most of the above studies use obtrusive sensors that may hamper task performance and slow critical reaction times. Our study focuses on examining whether unobtrusive sensors in VR can be used to estimate workload and cybersickness reliably.

3 Methodology

3.1 Experiment Setting

This study aims to induce cognitive load and cybersickness through a series of six flying tasks of various complexity. The weather conditions were kept constant during all the tasks and were simulated to be bright, clear, sunny, and with no wind. The tasks involve flying a Boeing T-45C jet through six virtual gates placed at regular intervals and are described in further detail in Table 1. A vertical inside loop maneuver is "when the pilot pulls the plane up into the vertical, continues around until they are heading back in the same direction, like making a 360° turn", and a barrel roll maneuver is "a combination of a loop and a roll. The flight path during a barrel roll has the shape of a horizontal corkscrew" [60].

The simulation is built entirely in Lockheed Martin's Prepar3D [34] flight simulator, which allows precise customization and provides Steam VR [57] support. Moreover, Prepar3D allows for creating scenarios where developers can

Table 1. Description of the flying tasks

Task no.	Task name	Task description
1	Flying straight	Requires flying in a straight line path while making sure to pass virtual gates at regular intervals
2	Minor turns	Requires performing minor turns between gates. A virtual arrow points to the next gate
3	180-degree turns	Requires performing 180-degree turns between gates. A virtual arrow points to the next gate but requires participants to look over their shoulders to notice the gate
4	Loop maneuver	Requires performing 180-degree turns between gates and executing a total of 5 vertical inside loop maneuvers upon hearing an audio prompt
5	Barrel roll maneuver	Requires performing 180-degree turns between gates and executing a total of 5 barrel roll maneuvers upon hearing an audio prompt
6	Nearsight/Farsight focusing	Requires performing 180-degree turns between gates and reporting the flight's current airspeed, altitude, and heading upon hearing an audio prompt. Participants are instructed to read this information off of the head-up display

customize the flight simulation's environment and provide additional guidance to users through user interface elements, sound, object visibility, and more. It also enables logging flight data through additional scripting. Upon completing each task, participants fill out two questionnaires, pertaining to their workload and cybersickness. These questionnaires were administered within the VR environment by placing a 2D plane showing survey items in a 3D space. Participants were able to use the HOTAS controls to toggle through and select survey items.

The HP Reverb G2 Omnicept Edition headset [24] is utilized in this study as it packs a suite of unobtrusive physiological sensors, namely, Eye-tracking, Pupillometry, and a PPG sensor. The headset bears a resolution of 2160 × 2160 per eye, a 90 Hz refresh rate display, and a field of view of 114°. Participants control the aircraft by operating a Thrustmaster Warthog HOTAS [55] and the Thrustmaster Pedals [56] to manipulate the rudder. Figure 1(a) shows the user flying the jet in VR and Fig. 1(b) depicts a screen capture of the testbed in Prepar3D.

3.2 Measured Feature Sets

The cognitive load of a pilot can be assessed through various techniques and is classified into three categories: (i) Objective, (ii) Subjective, and (iii) Physiological [5,53]. Our study employs measures from all three categories and are described below.

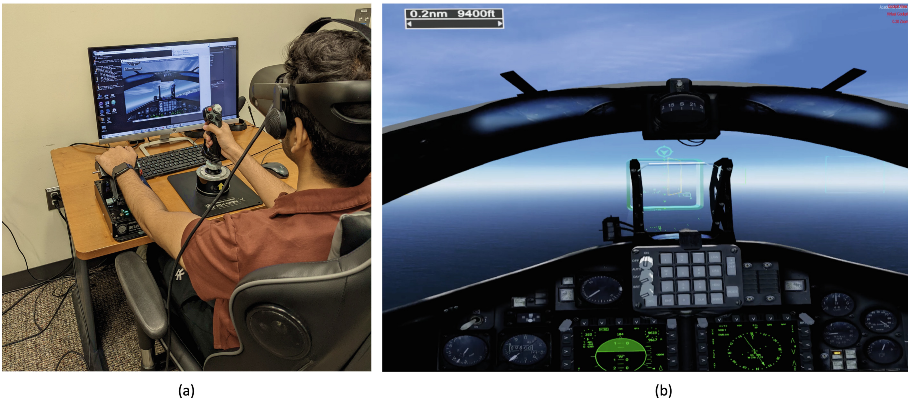


Fig. 1. (a) A user flying the Boeing T-45C jet using a HOTAS, and (b) a screen capture of the testbed in Prepar3D.

Pilot Inceptor Workload. PIW is the workload expended to control the aircraft through the primary flight controls [18]. It is characterized by two individual measures: duty cycle and aggressiveness, which are closely associated with the pilot's appreciation of their workload [41].

Duty Cycle: Duty cycle is defined as "the percentage of time the pilot changes his input on the inceptor" [19].

Aggressiveness: Aggressiveness is defined as "the root-mean-squared per-second average of the inceptor position rate of change with respect to time" [18,19].

Here, inceptor refers to the flight stick used in the study. These behavioral metrics are analyzed to identify their correlations to cognitive load and cybersickness, and they serve as a ground truth measure of task complexity. Prepar3D outputs the stick deflection measure in the X and Y axis, from which duty cycle and aggressiveness are computed as follows [41]:

$$\text{Duty cycle} = 100\% * \frac{1}{t_n - t_2} \sum_{i=2}^{n} x_i$$

$$\text{Here, } x_i = \begin{cases} 0, & \text{if } \frac{\delta_i - \delta_{i-1}}{t_i - t_{i-1}} < noise\ threshold\ and\ |\delta_i| < \delta_{max} \\ 1, & \text{otherwise} \end{cases}$$

$$\text{Aggressiveness} \quad = \sqrt{\frac{1}{n-1} \sum_{i=2}^{n} (\frac{\delta_i - \delta_{i-1}}{t_i - t_{i-1}})^2}$$

where δ_i refers to the discrete value of the stick deflection, t_i refers to the discrete value of time, and n is the number of discrete data points.

Heart Rate and Heart Rate Variability. Omnicept, via the PPG sensor, provides Heart Rate (HR) and Heart Rate Variability (HRV), which are consistently used in workload research due to their sensitivity to autonomic responses and the relationship to cognitive functions [37,49].

Root Mean Square of Successive Differences (RMSSD): The RMSSD is a robust time-domain measure known to reflect vagal nerve mediated HRV changes (i.e. autonomic nervous system influences) and is not sensitive to respiration mediated HRV changes like other measures, such as HR Max - HR Min [38,50,51,59].

High Frequency (HF): HF is the appropriate frequency-domain measure for the time windows chosen for this study as compared to lower frequencies which are more sensitive to longer physiological rhythms over longer time windows [40]. Lower HF has also been associated with parasympathetic activity driven by stress, panic, anxiety and worry, as well as inflexibility to changing environmental demands, and is connected to attentional and emotional self-regulatory systems [12,49,54].

Eye Tracking and Pupillometry. Eye-tracking offers the most dependable and non-invasive estimate of cognitive load [8]. The Omnicept system enables access to various eye-tracking metrics, although this study utilizes only the pupil dilation data and the combined eye-gaze vectors.

Pupil Diameter Change: Pupil dilation activity is a physiological signal directly linked to the peripheral nervous system [6]. Regardless, changes in pupil dilation are susceptible to changes in ambient light intensity [4] and off-axis distortion [36], and the changes that are related to cognitive activities are relatively small [6]. Hence, pupil dilation should be measured with respect to an average dilation measure during a baseline trial. In that regard, a baseline average was calculated during the period the participants read the study instructions in VR.

The pupil dilation signal is passed through 4 Hz Butterworth low-pass filter to remove artifacts and high-frequency noise. The left and right eye pupil diameters are averaged to further reduce noise in the signal. The baseline average is then subtracted from this signal to get the pupil diameter change measure.

Number of Fixations and Saccades: Fixations are a type of eye movement wherein the eye focuses on a central foveal vision such that the visual system can gather information about the object in focus. Saccades are a type of eye movement used to move the central focus rapidly from one point to another. Fixations and saccades are linked to cognitive activity as they are encoded by neural signals from cortical and subcortical systems [6]. Fixations and saccades are extracted from the combined gaze vector data through the Velocity-Threshold Identification (I-VT) algorithm [42,47], which classifies gaze vectors into either a fixation or a saccade depending on their point-to-point angular velocities. The angular velocity threshold used in this study is 30 degrees per second and was chosen based on the study by Olsen and Matos [43].

Omnicept's Cognitive Load Measure. The Omnicept system also entails a machine learning model, which takes as input all of the Omnicept's physiological measures and outputs a real-time cognitive load metric, ranging from 0 to 1. Their model was trained based on a dataset including behavioral and physiological indicators while users performed mentally demanding tasks in VR. HP details this system in their white paper [52].

Self-reported Measures. The self-report measures administered in VR after each task were designed to assess participants' mental workload and cybersickness. These surveys included the mental demand sub-item from the NASA-TLX [20] questionnaire, which prompts them to rate how mentally demanding the task was on a scale of one to ten, and the Virtual Reality Sickness Questionnaire (VRSQ) [27] which asks participants to rate on a 4-point Likert scale any VR symptoms they experienced, including General discomfort, Fatigue, Eyestrain, Difficulty focusing, Headache, Fullness of the head, Blurred vision, Dizziness, and Vertigo.

Moreover, participants also completed a Motion Sickness Susceptibility Questionnaire (MSSQ) [14] prior to the study, asking them to rate how nauseated or sick they felt during various scenarios in the past, on a 4-point Likert scale, and the Motion Sickness Assessment Questionnaire (MSAQ) [13] after the study, where they rate their experience on different criteria related to motion sickness on a 9-point Likert scale. The scores from these two questionnaires can be examined to determine if VR induces cybersickness in participants who have increased susceptibility to motion sickness and in what effect.

3.3 Study Design

This study used only one independent variable: the flying tasks with six levels of complexity. The study used a wholly within-subjects design, with each participant performing all six flying tasks in succession. A total of 6 participants comprising students from a university in the western United States, of which all 6 were males and with ages ranging from 23 to 26 (M = 24.7, SD = 1.2), took part in the study. Out of the 6 participants, only 1 participant had prior experience with VR. All participants had experience playing video games and averaged 2.9 h (SD = 1.8) of video games played in a week.

Participants first provide informed consent to participate in the study, complete a demographics questionnaire, followed by the MSSQ assessment. Following this, they were debriefed on the tasks they had to perform and were shown video tutorials on how to perform a loop and a barrel roll maneuver. They were also trained on using the HOTAS to operate the aircraft. Participants are informed to pass through the gates as efficiently as possible and interpret missing a gate as a flight safety violation. They complete a tutorial round to become familiar with the simulator and controls. The tutorial is similar to task 1, although participants were allowed to practice maneuvers between gates. The participants were allowed to take breaks between tasks as and when required. To acquire accurate eye-gaze data, an eye calibration was performed using HP's Eye Calibration software [22], each time the headset was taken off and worn again.

Between each task, participants fill out two questionnaires in VR. The first one is the VRSQ for self-reporting cybersickness. The second questionnaire is a sub-item from NASA-TLX for self-reporting workload. After completing the study, participants fill out the MSAQ and are also advised to list any comments they might have related to the study.

4 Results

We examined the responses of the feature sets listed in Sect. 3.2 to the six flying tasks. The data were averaged over a 30-s window, starting when a participant crossed a gate and was averaged across five gates. This window was specifically chosen considering that much of the effort applied during the tasks was during this period as participants had to orient themselves to the next gate, perform loops or barrel rolls, and so on. Figure 3 illustrates the comparison of the mean measures for all tasks, while the error bars depict the 95% confidence interval of these means. We tested for normality using the Shapiro-Wilk tests for all feature sets. The results were not significant, indicating that the feature sets did not deviate from normality. Considering the different task conditions as the independent variable, we performed a one-way repeated measures ANOVA followed by Bonferroni post hoc tests of the feature sets to test for significance. Due to a technical difficulty described in Sect. 5.2, heart rate data could not be collected for three of the six participants and is not investigated further.

The results show that the duty cycle (F(5, 25) = 4.27, $p < .01$, $r = .46$) and aggressiveness (F(5, 25) = 7.55, $p < .001$, $r = .6$) were significantly affected by the tasks, with Mauchly's test indicating that the assumption of sphericity had been met for both duty Cycle ($\chi^2(14) = 26.78$, $p > .05$) and aggressiveness ($\chi^2(14) = 11.29$, $p > .1$). Plotting aggressiveness versus duty cycle is commonly known as the PIW plot, and its general layout [18] is shown in Fig. 2(a). High aggressiveness and low duty cycle represent occasional fast inputs (upper-left corner), low aggressiveness and high duty cycle represent constant slow inputs (bottom-right corner), and high aggressiveness and high duty cycle represent constant fast inputs (upper-right corner) and are where the pilot gain is the highest. Figure 2(b) illustrates the PIW plot for the six tasks undertaken in our study.

Moreover, it was found that the tasks significantly affected the subjective NASA-TLX ratings (F(5, 25) = 9.0, $p < .001$, $r = .64$) and the Omnicept's cognitive load measure, (F(5, 25) = 12.31, $p < .001$, $r = .71$). The effect size indicated that the effect of tasks on the Omnicept's cognitive load measure was substantial. There was no significant affect of the tasks on the number of fixations (F(5, 25) = 2.14, $p = .094$, $r = .30$), the pupil diameter change (F(5, 25) = 2.16, $p = .092$, $r = .30$) and the VRSQ scores (F(5, 25) = 1.37, $p = .27$, $r = .21$). Meanwhile, the number of saccades measure was significantly affected by the tasks (F(1.51, 7.58) = 5.13, $p < .05$, $r = .50$) while having violated the sphericity assumption indicated by Mauchly's test ($\chi^2(14) = 29.14$, $p < .05$); therefore the degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\epsilon = .30$).

Bonferroni post hoc tests revealed a significant difference in the duty cycle only between tasks 1 and 2 (CI.95 = -16.68 (lower) -4.0 (upper), $p < .01$). No other comparisons were significant (all $ps > .05$). For aggressiveness, Bonferroni post hoc tests revealed a significant difference between tasks 1 and 2 (CI.95 = $-.13$ (lower) $-.001$ (upper), $p < .05$), tasks 1 and 3 (CI.95 = $-.15$ (lower) $-.004$ (upper), $p < .05$), and tasks 1 and 5 (CI.95 = $-.14$ (lower) $-.011$ (upper), $p < .05$). The post hoc tests for Omnicept's cognitive load measure and the number of saccades measure did not reveal any significant differences.

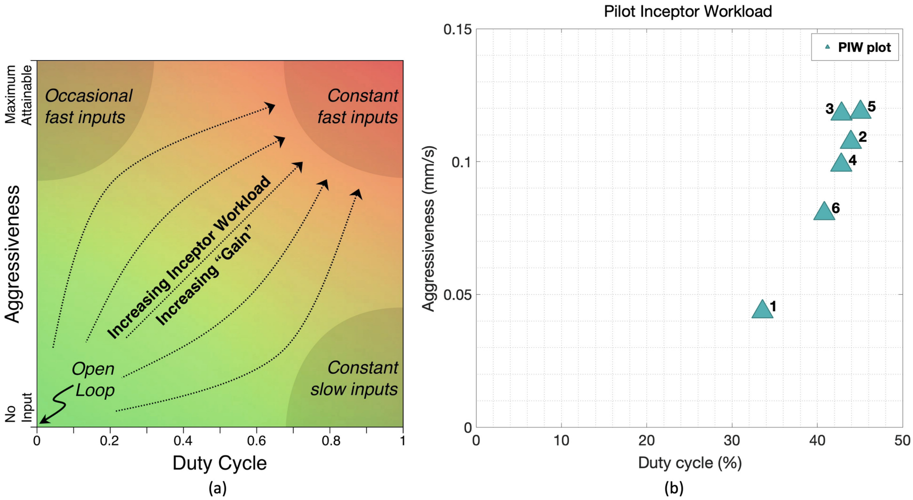
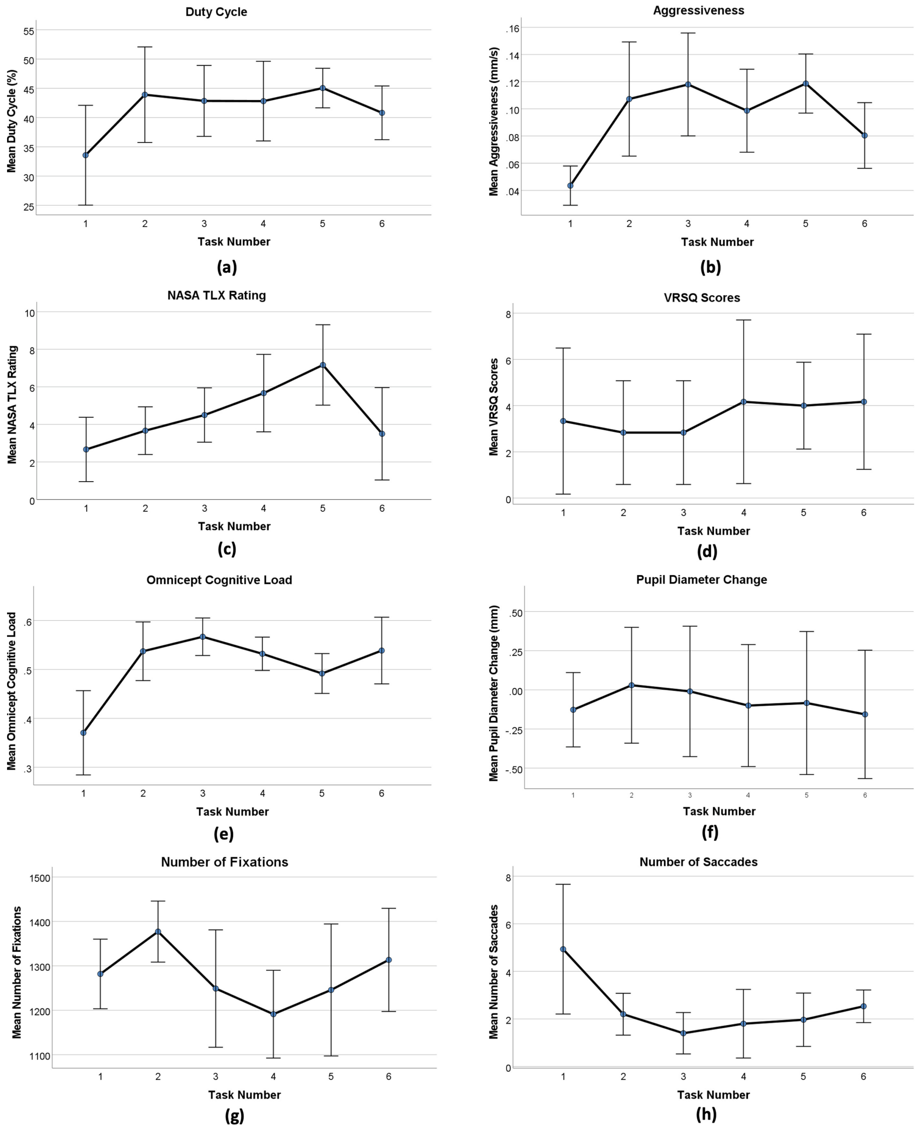


Fig. 2. (a) General layout of a Pilot Inceptor Workload (PIW) plot [18], and (b) PIW plots of the six flying tasks, averaged across all participants.

Additionally, we carried out a Pearson's correlation between the feature sets, listed in Sect. 3.2, the results of which are presented as a correlation matrix in Fig. 4 with the highlighted cells denoting a significant correlation. A strong positive correlation was observed between the ground truth measures: duty cycle and aggressiveness ($\rho = .80$, $p < .001$). Aggressiveness also exhibited a significant positive correlation with Omnicept's cognitive load measure ($\rho = .35$, $p < .05$) and the NASA TLX rating ($\rho = .43$, $p = .01$). Contrary to expectations, the number of saccades measure demonstrated a statistically significant negative correlation with duty Cycle ($\rho = -.52$, $p = .001$), aggressiveness ($\rho = -.45$, $p < .01$), and Omnicept's cognitive load measure ($\rho = -.64$, $p < .001$).



Duty Cycle

(a)

Aggressiveness

(b)

NASA TLX Rating

(c)

VRSQ Scores

(d)

Omnicept Cognitive Load

(e)

Pupil Diameter Change

(f)

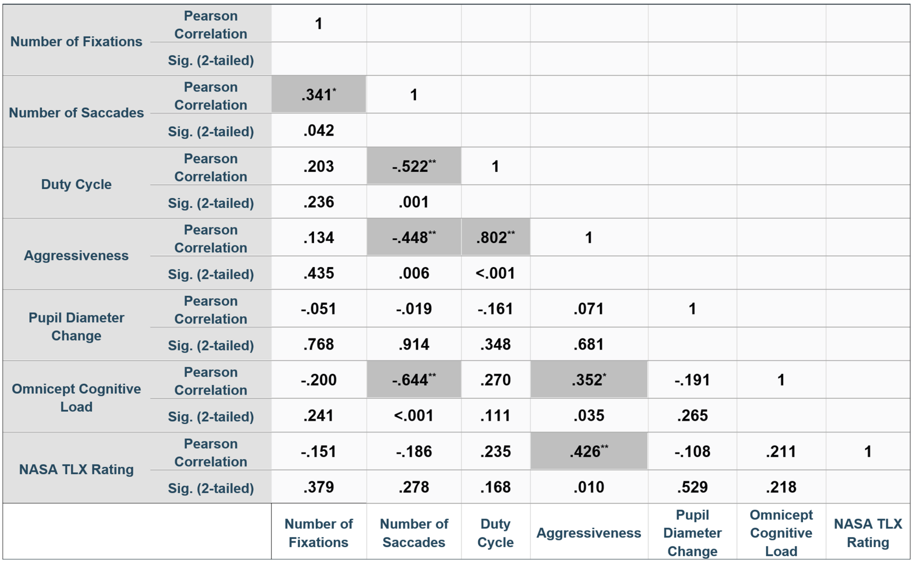
Number of Fixations

(g)

Number of Saccades

(h)

Fig. 3. Comparison of the measured feature sets: (a) Duty cycle, (b) Aggressiveness, (c) NASA TLX, (d) VRSQ scores, (e) Omnicept cognitive load, (f) Pupil diameter change, (g) Number of Fixations, and (h) Number of saccades to the six flying tasks, averaged across all six participants. Error bars represent 95% confidence intervals.



		Number of Fixations	Number of Saccades	Duty Cycle	Aggressiveness	Pupil Diameter Change	Omnicept Cognitive Load	NASA TLX Rating
Number of Fixations	Pearson Correlation	1						
	Sig. (2-tailed)							
Number of Saccades	Pearson Correlation	.341*	1					
	Sig. (2-tailed)	.042						
Duty Cycle	Pearson Correlation	.203	-.522**	1				
	Sig. (2-tailed)	.236	.001					
Aggressiveness	Pearson Correlation	.134	-.448**	.802**	1			
	Sig. (2-tailed)	.435	.006	<.001				
Pupil Diameter Change	Pearson Correlation	-.051	-.019	-.161	.071	1		
	Sig. (2-tailed)	.768	.914	.348	.681			
Omnicept Cognitive Load	Pearson Correlation	-.200	-.644**	.270	.352*	-.191	1	
	Sig. (2-tailed)	.241	<.001	.111	.035	.265		
NASA TLX Rating	Pearson Correlation	-.151	-.186	.235	.426**	-.108	.211	1
	Sig. (2-tailed)	.379	.278	.168	.010	.529	.218	

Fig. 4. Pearson correlation matrix of the feature sets, with significant correlations highlighted and starred (* corresponds to significance at the 0.05 level and ** corresponds to significance at the 0.01 level).

Lastly, we investigated the correlation between the MSSQ and the MSAQ scores to understand whether prior susceptibility to motion sickness influenced cybersickness. The results show no correlation between the two measures ($\rho = .008$, $p = .99$). Figure 5 portrays this correlation via a scatter plot of the scores provided by the six participants.

5 Discussion

5.1 Experimental Findings

The main results of this study were that the PIW measures and some of the physiological measures (number of saccades measure and Omnicept's cognitive load measure) were significantly affected by the tasks having varying levels of complexity. The PIW plots in Fig. 2(b), when compared to the general layout of a PIW plot in Fig. 2(a), provide an exceptional overview of the complexity involved in the six tasks used in this study. A higher duty cycle implies that more time was required to bring the aircraft under control, whereas higher aggressiveness refers to random quick movements of the aircraft's control inceptor. In our study, the tasks 180-degree Turns and the Barrel Roll Maneuver induce the highest pilot gain. The results also show that aggressiveness had a large effect size, which further strengthens our argument to use it as a ground truth measure.

Moreover, It can be inferred from Fig. 3 that except for the number of saccades, all other feature sets display an almost similar trend across the six tasks;

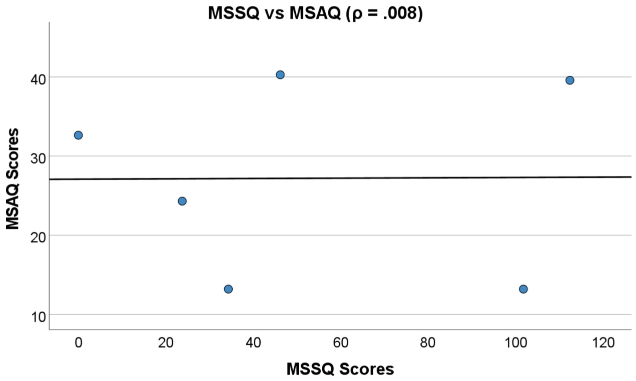


Fig. 5. A scatter plot depicting the correlation between the MSSQ and the MSAQ scores.

participants found tasks to be increasing in complexity until task 3. Furthermore, they found the Barrel Roll Maneuver to be the most cognitively demanding task according to their NASA-TLX ratings, which are also corroborated by the uptick in the physiological responses and the PIW measures.

However, the Loop Maneuver task (task 4) shows a drop in all measures except the number of saccades which exhibit a negative correlation. The Loop Maneuver task was the most challenging for participants since if not done correctly, the aircraft would stall (as it requires flying against gravity before flying towards it) and stumble into an uncontrolled spiral, at which point the task had to be restarted. This was the case with five out of six participants from our observations. We infer that this could have led to practice effects among the participants, and hence the decline in the measures. Another reason could be a compensatory mechanism invoked by increased novelty, complexity, or difficulty of the task [31,62], hence exhibiting a lower physiological response. Another alternative possibility is that pausing and restarting the task skewed the dependent measures being collected, and we note that future work should be done to further explore these measurements.

The lower aggressiveness can be attributed to participants being more steady with the flight stick during the loop, which is necessary for a more even and precise circle. Similarly, as participants held the stick at a stable position, this period was not accounted for in the duty cycle calculations and is a limitation of the equation. The duty cycle formula does account for situations where the stick is in its maximum position with the reasoning that pilots would move it further if possible. However, in our study, participants rarely used the stick in its maximum position as this would cause high G's on the pilot, and the simulation would blackout until the stick was restored to a suitable position.

Meanwhile, the number of fixations measure displayed similar variations as the duty cycle measure, consistent with previous findings [1,3,7,10,11], although except for task 6. The task required participants to fixate on the head-up display

(HUD) to report the flight's information, and they often complained about the information being hard to read due to the low contrast between the sky and the numbers displayed on the transparent HUD. Our reasoning for the increase in fixations in this task is that participants had to strenuously focus on the HUD to report accurate information.

Though past research has shown that the number of saccades and saccade rates are positively correlated with cognitive load [3,7,44], we cannot corroborate their findings as we encounter a statistically significant negative correlation between the PIW and Omnicept's cognitive load measures. The mean number of saccades also exhibits an uptick in task 4, opposing all other physiological and PIW measures. The effect size also indicates that the number of saccades had a large effect. The argument of whether saccades are a dependable positive indicator of mental effort needs to be re-evaluated.

The mean pupil diameter change measure shows a similar trend to the duty cycle measure, albeit having larger variability. This trend agrees with past work establishing pupil diameter change as a reliable workload indicator [6,15,29]. Regardless, no significant correlation was detected between the pupil diameter change measure and the PIW measures. The Omnicept cognitive load measure, on the other hand, exhibited a large effect size ($r = .71$) and demonstrated a significant correlation with the aggressiveness measure, although not a strong one. We conjecture that it may serve as a reliable workload estimate, though it should not be the only estimate.

Lastly, the VRSQ scores display a nearly similar trend to the number of saccades measure, which would mean that saccades could be a positive indicator of cybersickness and should be explored further. Comparing MSSQ and MSAQ scores showed no correlation between them, hinting that prior susceptibility to motion sickness may not influence VR cybersickness. However, no significance was detected in this correlation, and the sample size was limited to only six - one of the participants also claimed to feel nauseous but was unsure if it was from the simulation or food poisoning; hence necessitating further research.

5.2 Study Limitations

Although we demonstrated a significant negative correlation of saccades with cognitive load, saccades are swift movements of the eye to shift the central foveal vision, and detecting them accurately requires an eye tracker with a high sampling rate (≈1000 Hz) [33]. A limitation in our study is the low sampling rate (120 Hz) of the in-built eye tracker [23], which may impact its ability to detect saccades. Nonetheless, there does not yet exist a VR HMD with an eye-tracking frequency greater 200 Hz. Babu et al. [3], however, used the Tobii Pro Glasses 2 with 100 Hz sampling frequency to establish a positive trend between the number of saccades and workload. Additionally, the I-VT algorithm used to classify eye movements in this study, by its nature, is more targeted towards classifying fixations than saccades [42].

Furthermore, we experienced a failure to capture the heart rate data for all participants, had a limited sample size, and our simulation testbed contained an issue for Task 4. Due to an unknown reason, heart rate data was only captured for three of the six participants. Upon further inspection of the three data files, one was unusable due to severe noise in the signal, so these data were not included in this study. After some troubleshooting, we found that the noise was likely due to the tightness of the headset, but we could not find a reason for the complete failure to capture. Our sample was limited mostly in part to a small subject pool available at the time of data collection. The issue with Task 4 was an uncontrollable spiraling and required restart triggered by a failed attempt at the loop. The testbed is being investigated to find a solution that prevents these outcomes for future testbed usage.

6 Conclusion and Future Work

In the present study, we validated the use of unobtrusive sensors in estimating cognitive load through VR simulations. Our study supports previous findings demonstrating that ocular parameters such as pupil diameter change and the number of fixations pose a sensitive and reliable discriminant of task complexity in regards to cognitive load. We also contradict prior research by demonstrating a significant negative correlation between the number of saccades and cognitive load. In further research, it would be useful to conduct the study with a more representative sample such as military pilots. Using virtual gates during maneuvers would assist participants in performing them more accurately and diligently. Consequently, future research should also examine fixations and saccades through other less explored classification algorithms such as the Dispersion-Threshold Identification algorithm [47], which classifies fixations and saccades based on their clusterization.

Acknowledgements. We would like to acknowledge the U.S. Navy for supporting this research via contract number N68335-18-C0133.

References

1. Arjun, S., Rajshekar Reddy, G., Mukhopadhyay, A., Vinod, S., Biswas, P.: Evaluating visual variables in a virtual reality environment. In: 34th British HCI Conference 34, pp. 11–22 (2021)
2. Babu, M.D.: Investigation of pilot inceptor workload and workload buildup technique through simulator and in-flight studies. Int. J. Aerosp. Psychol. **32**(2–3), 65–94 (2022)
3. Babu, M.D., JeevithaShree, D., Prabhakar, G., Saluja, K.P.S., Pashilkar, A., Biswas, P.: Estimating pilots' cognitive load from ocular parameters through simulation and in-flight studies. J. Eye Mov. Res. **12**(3) (2019)
4. Beatty, J., Lucero-Wagoner, B., et al.: The pupillary system. In: Handbook of Psychophysiology, 2nd edn, pp. 142–162 (2000)
5. Casner, S.M., Gore, B.F.: Measuring and evaluating workload: a primer. NASA Technical Memorandum, NASA–TM 2010-216395 (2010)
6. Chen, S., Epps, J.: Automatic classification of eye activity for cognitive load measurement with emotion interference. Comput. Methods Programs Biomed. **110**(2), 111–124 (2013)

7. Coral, M.P.: Analyzing cognitive workload through eye-related measurements: a meta-analysis. Ph.D. thesis, Wright State University (2016)
8. Cowley, B.U., et al.: The psychophysiology primer: a guide to methods and a broad review with a focus on human-computer interaction. Found. Trends Hum. Comput. Interact. **9**(3–4), 151–308 (2016)
9. Davis, S., Nesbitt, K., Nalivaiko, E.: A systematic review of cybersickness. In: Proceedings of the 2014 Conference on Interactive Entertainment, IE 2014, pp. 1–9. Association for Computing Machinery, New York (2014). https://doi.org/10.1145/2677758.2677780
10. Di Nocera, F., Camilli, M., Terenzi, M.: Using the distribution of eye fixations to assess pilots' mental workload. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 50, pp. 63–65. SAGE Publications, Los Angeles (2006)
11. Feng, C., Wanyan, X., Yang, K., Zhuang, D., Wu, X.: A comprehensive prediction and evaluation method of pilot workload. Technol. Health Care **26**(S1), 65–78 (2018)
12. Forte, G., Favieri, F., Casagrande, M.: Heart rate variability and cognitive function: a systematic review. Front. Neurosci. **13**, 710 (2019)
13. Gianaros, P.J., Muth, E.R., Mordkoff, J.T., Levine, M.E., Stern, R.M.: A questionnaire for the assessment of the multiple dimensions of motion sickness. Aviat. Space Environ. Med. **72**(2), 115 (2001)
14. Golding, J.F.: Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. Brain Res. Bull. **47**(5), 507–516 (1998)
15. Grant, T., Dhruv, K., Eloy, L., Hayne, L., Durkee, K., Hirshfield, L.: A neurophysiological sensor suite for real-time prediction of pilot workload in operational settings. In: Stephanidis, C., et al. (eds.) HCII 2020. LNCS, vol. 12425, pp. 60–77. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-60128-7_5
16. Gray, W.: Boundary-escape tracking: a new conception of hazardous PIO. Technical report, Air Force Flight Test Center, Edwards AFB, CA (2004)
17. Gray, W.: Boundary avoidance tracking: a new pilot tracking model. In: AIAA Atmospheric Flight Mechanics Conference and Exhibit, p. 5810 (2005)
18. Gray, W.: A generalized handling qualities flight test technique utilizing boundary avoidance tracking. In: 2008 US Air Force T&E Days, p. 1648. Aerospace Research Central (2008)
19. Gray, W.: A boundary avoidance tracking flight test technique for performance and workload assessment. In: Proceedings of the 38th Symposium of Society of Experimental Test Pilots, San Diego (2007)
20. Hart, S.G., Staveland, L.E.: Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. **52**, 139–183 (1988)
21. Hebbar, P.A., Bhattacharya, K., Prabhakar, G., Pashilkar, A.A., Biswas, P.: Correlation between physiological and performance-based metrics to estimate pilots' cognitive workload. Front. Psychol. **12**, 954 (2021)
22. HP Development Company: Eye Tracking Calibration (2022). https://developers.hp.com/omnicept/eye-tracking-calibration. Accessed 23 Feb 2022
23. HP Development Company: Fundamentals (2022). https://developers.hp.com/omnicept/docs/fundamentals?language=es-un. Accessed 23 Feb 2022
24. HP Development Company: HP Reverb G2 Omnicept Edition (2022). https://www.hp.com/us-en/vr/reverb-g2-vr-headset-omnicept-edition.html. Accessed 23 Feb 2022

25. Huttunen, K., Keränen, H., Väyrynen, E., Pääkkönen, R., Leino, T.: Effect of cognitive load on speech prosody in aviation: evidence from military simulator flights. Appl. Ergon. **42**(2), 348–357 (2011)
26. Itoh, Y., Hayashi, Y., Tsukui, I., Saito, S.: The ergonomic evaluation of eye movement and mental workload in aircraft pilots. Ergonomics **33**(6), 719–732 (1990)
27. Kim, H.K., Park, J., Choi, Y., Choe, M.: Virtual reality sickness questionnaire (VRSQ): motion sickness measurement index in a virtual reality environment. Appl. Ergon. **69**, 66–73 (2018)
28. Kim, Y.Y., Kim, H.J., Kim, E.N., Ko, H.D., Kim, H.T.: Characteristic changes in the physiological components of cybersickness. Psychophysiology **42**(5), 616–625 (2005)
29. Krejtz, K., Duchowski, A.T., Niedzielska, A., Biele, C., Krejtz, I.: Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. PLoS ONE **13**(9), e0203629 (2018)
30. Lancaster, J.A., Casali, J.G.: Investigating pilot performance using mixed-modality simulated data link. Hum. Factors **50**(2), 183–193 (2008)
31. Latham, G.P., Seijts, G., Crim, D.: The effects of learning goal difficulty level and cognitive ability on performance. Can. J. Behav. Sci. **40**(4), 220 (2008)
32. LaViola, J.J.: A discussion of cybersickness in virtual environments. SIGCHI Bull. **32**(1), 47–56 (2000). https://doi.org/10.1145/333329.333344
33. Leube, A., Rifai, K., Rifai, K.: Sampling rate influences saccade detection in mobile eye tracking of a reading task. J. Eye Mov. Res. **10**(3) (2017)
34. Lockheed Martin Corporation: Lockheed Martin Prepar3D (2022). https://www.prepar3d.com. Accessed 23 Feb 2022
35. Lohani, M., Payne, B.R., Strayer, D.L.: A review of psychophysiological measures to assess cognitive states in real-world driving. Front. Hum. Neurosci. **13**, 57 (2019)
36. Mathur, A., Gehrmann, J., Atchison, D.A.: Pupil shape as viewed along the horizontal visual field. J. Vis. **13**(6), 3–3 (2013)
37. McCraty, R., Shaffer, F.: Heart rate variability: new perspectives on physiological mechanisms, assessment of self-regulatory capacity, and health risk. Glob. Adv. Health Med. **4**(1), 46–61 (2015)
38. Mehler, B., Reimer, B., Wang, Y.: A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload. In: Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, pp. 590–597 (2011)
39. Mohanavelu, K., et al.: Cognitive workload analysis of fighter aircraft pilots in flight simulator environment. Defence Sci. J. **70**(2), 131–139 (2020)
40. Mulder, L.J.M.: Assessment of cardiovascular reactivity by means of spectral analysis. Ph.D. thesis, Rijksuniversiteit (1988)
41. Niewind, I.: Pilot gain and the workload buildup flight test technique: about the influence of natural pilot gain on the achievable pilot gain range. Institute report IB 111-2012/69. German Aerospace Centre (2012)
42. Olsen, A.: The Tobii I-VT fixation filter. Tobii Technology, p. 21 (2012)
43. Olsen, A., Matos, R.: Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. In: Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA 2012, pp. 317–320. Association for Computing Machinery, New York (2012). https://doi.org/10.1145/2168556.2168625

44. Prabhakar, G., Mukhopadhyay, A., Murthy, L., Modiksha, M., Sachin, D., Biswas, P.: Cognitive load estimation using ocular parameters in automotive. Transp. Eng. **2**, 100008 (2020)
45. Roscoe, A.H.: Assessing pilot workload. Why measure heart rate, HRV and respiration? Biol. Psychol. **34**(2–3), 259–287 (1992)
46. Roscoe, A.H., Ellis, G.A.: A subjective rating scale for assessing pilot workload in flight: a decade of practical use. Technical report, Royal Aerospace Establishment Farnborough, United Kingdom (1990)
47. Salvucci, D.D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, pp. 71–78 (2000)
48. Schmidt, E.A., Schrauf, M., Simon, M., Fritzsche, M., Buchner, A., Kincses, W.E.: Drivers' misjudgement of vigilance state during prolonged monotonous daytime driving. Accid. Anal. Prev. **41**(5), 1087–1093 (2009)
49. Shaffer, F., Ginsberg, J.P.: An overview of heart rate variability metrics and norms. Front. Public Health **5**, 258 (2017)
50. Shaffer, F., McCraty, R., Zerr, C.L.: A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. Front. Psychol. **5**, 1040 (2014)
51. Shakouri, M., Ikuma, L.H., Aghazadeh, F., Nahmens, I.: Analysis of the sensitivity of heart rate variability and subjective workload measures in a driving simulator: the case of highway work zones. Int. J. Ind. Ergon. **66**, 136–145 (2018)
52. Siegel, E., et al.: HP Omnicept cognitive load database (HPO-CLD)-developing a multimodal inference engine for detecting real-time mental workload in VR. Technical report, HP Labs, Palo Alto (2021). https://developers.hp.com/omnicept
53. Spyker, D., Stackhouse, S., Khalafalla, A., McLane, R.: Development of techniques for measuring pilot workload. Technical report, NASA (1971)
54. Thayer, J.F., Lane, R.D.: Claude Bernard and the heart-brain connection: further elaboration of a model of neurovisceral integration. Neurosci. Biobehav. Rev. **33**(2), 81–88 (2009)
55. Thrustmaster: HOTAS WARTHOG (2022). https://www.thrustmaster.com/products/hotas-warthog/. Accessed 23 Feb 2022
56. Thrustmaster: T.Flight Rudder Pedals (2022). https://www.thrustmaster.com/en-us/products/t-flight-rudder-pedals/. Accessed 23 Feb 2022
57. Valve Corporation: SteamVR on Steam (2022). https://store.steampowered.com/app/250820/SteamVR/. Accessed 23 Feb 2022
58. Villafaina, S., Fuentes-García, J.P., Gusi, N., Tornero-Aguilera, J.F., Clemente-Suárez, V.J.: Psychophysiological response of military pilots in different combat flight maneuvers in a flight simulator. Physiol. Behav. **238**, 113483 (2021)
59. Wang, Z., Yang, L., Ding, J.: Application of heart rate variability in evaluation of mental workload. Zhonghua lao dong wei sheng zhi ye bing za zhi= Zhonghua laodong weisheng zhiyebing zazhi= Chin. J. Ind. Hyg. Occup. Dis. **23**(3), 182–184 (2005)
60. Wikipedia contributors: Aerobatic maneuver – Wikipedia, the free encyclopedia (2021). https://en.wikipedia.org/w/index.php?title=Aerobatic_maneuver&oldid=1044302151. Accessed 23 Feb 2022
61. Wikipedia contributors: Microsaccade – Wikipedia, the free encyclopedia (2022). https://en.wikipedia.org/w/index.php?title=Microsaccade&oldid=1064617483. Accessed 23 Feb 2022
62. Wong, L., et al.: Performance on the robotics on-board trainer (ROBoT-r) spaceflight simulation during acute sleep deprivation. Front. Neurosci. **14**, 697 (2020)