

Real-Time Pilot Mental Workload Prediction Through the Fusion of Psychophysiological Signals

Matthew Masters, M.S.E.

Vollständiger Abdruck der von der
Fakultät für Luft- und Raumfahrttechnik
der Universität der Bundeswehr München
zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. rer. nat. Eric Jäggle
1. Berichterstatter: Univ.-Prof. Dr.-Ing. Axel Schulte
2. Berichterstatter: Univ.-Prof. Dr. Frederic Dehais

Die Dissertation wurde am 6. Juni 2023 bei der Universität der Bundeswehr München eingereicht und durch die Fakultät für Luft- und Raumfahrttechnik am 20. September 2023 angenommen. Die mündliche Prüfung fand am 16. Oktober 2023 statt.

Abstract

Military helicopter pilots and their aircraft form a unique system relied upon to be highly functioning. This work explores an aspect of the system largely ignored by the industrial developers of these systems, namely the mental workload of the pilots during flight. Supported by a systematic review of previously-published works, it is reasoned that mental workload is experienced uniquely by each individual and that it cannot be deduced through an analysis of the task load alone.

A technical solution is developed and tested for predicting pilot mental workload in real-time which processes and fuses psychophysiological data from multiple sources supporting a multi-modal assessment. Specifically, signals processed include functional near-infrared spectroscopy (fNIRS), electrocardiography (ECG), electrodermal activity (EDA), respiration, and eye-movement-related signals. The unique signal processing chains are presented including the algorithms for extracting workload-relevant features and methods implemented to ensure robust data acquisition and processing.

Experimentation of the system with ten operational military helicopter pilots and ten university students shows a moderate linear correlation between subjective and predicated mental workload (average Pearson's correlation coefficient of 0.36 ± 0.21). The individual feature with the strongest linear correlation to subjective mental workload is an instantaneous standard deviation of all deoxygenated hemoglobin channels recorded from the pre-frontal cortex. This discovery is significant as this feature has not been identified by previously-published works as being sensitive to mental workload.

The developed system (including an in-cockpit display) demonstrates a high level of transparency required for effective human-machine systems. At last, a gauge in the cockpit for the most important sub-system in the human-machine team – the human!

Kurzfassung

Militärhubschrauberpiloten und ihre Flugzeuge bilden ein einzigartiges System, das in hohem Maße funktionsfähig sein muss. Diese Arbeit untersucht einen Aspekt des Systems, der von den industriellen Entwicklern dieser Systeme weitgehend ignoriert wird, nämlich die mentale Arbeitsbelastung der Piloten während des Fluges. Gestützt auf eine systematische Durchsicht bereits veröffentlichter Arbeiten wird argumentiert, dass die mentale Arbeitsbelastung von jedem Individuum auf einzigartige Weise erlebt wird und nicht allein durch eine Analyse der Aufgabenbelastung abgeleitet werden kann.

Es wird eine technische Lösung zur Prädiktion der mentalen Arbeitsbelastung von Piloten in Echtzeit entwickelt und getestet, die psychophysiologische Daten aus verschiedenen Quellen verarbeitet und zusammenführt, um eine multimodale Bewertung zu ermöglichen. Zu den verarbeiteten Signalen gehören die funktionelle Nahinfrarotspektroskopie (fNIRS), die Elektrokardiographie (EKG), die elektrodermale Aktivität (EDA), die Atmung und augenbewegungsbezogene Signale. Die einzigartigen Signalverarbeitungsketten werden vorgestellt, einschließlich der Algorithmen zur Extraktion von belastungsrelevanten Merkmalen und der implementierten Methoden zur Gewährleistung einer robusten Datenerfassung und -verarbeitung.

Die Erprobung des Systems mit zehn Militärhubschrauberpiloten und zehn Universitätsstudenten zeigt eine mäßige lineare Korrelation zwischen der subjektiven und der prädiktierten mentalen Arbeitsbelastung (durchschnittlicher Pearson-Korrelationskoeffizient von $0,36 \pm 0,21$). Das individuelle Merkmal mit der stärksten linearen Korrelation zur subjektiven mentalen Arbeitsbelastung ist die momentane Standardabweichung aller desoxygenierten Hämoglobinkanäle, die vom präfrontalen Kortex aufgezeichnet wurden. Diese Entdeckung ist bedeutsam, da dieses Merkmal in früheren Arbeiten nicht als empfindlich für mentale Arbeitsbelastung identifiziert worden war.

Das entwickelte System (inklusive eines In-Cockpit-Displays) weist ein hohes Maß an Transparenz auf, das für effektive Mensch-Maschine-Systeme erforderlich ist. Endlich eine Anzeige im Cockpit für das wichtigste Subsystem im Mensch-Maschine-Team - den Mensch!

Acknowledgments

I express my sincere gratitude to my wife Bree for her never-wavering support and encouragement that I see this through. We did it! I also thank my parents, Mike and Tamara, for instilling within me a strong desire to learn and an appreciation for hard work. Mostly though, I thank them for their love.

I thank my “Doctor-father” (Doktorvater), Professor Axel Schulte, for providing me the opportunity and means to pursue this research as well as for his friendly yet critical guidance and mentorship. Whether approving hardware purchases in support of my work or supporting my attendance at conferences around the world, he always worked to ensure I had the means to succeed. Being a member of his team was a highlight of my life.

I owe much of my success in the lab to my teammates who surrounded and supported me daily. Specifically, recognition is due to Diana Donath, Carsten Meyer, Evgeni Pavlidis, Matthias Frey, Gunar Roth, Markus Zwick, and Dominik Künzel. All other colleagues and friends I likewise thank for the debugging help or simply for the rejuvenating conversations during lunch or on the soccer field.

Lastly, I thank and acknowledge the colleagues, students, and pilots who participated in my experiments over the years for graciously giving their time to support my work.

This dissertation is based in part on the previously published articles listed below. I have permission from my co-authors/publishers to use the works listed below in this dissertation.

- Masters, M., and Schulte, A. “Physiological Sensor Fusion for Real-Time Pilot Workload Prediction in a Helicopter Simulator.” *Proceedings of the AIAA SciTech 2022 Forum*, 2022. <https://doi.org/10.2514/6.2022-2344>.
- Masters, M., and Schulte, A. “Investigating the Utility of FNIRS to Assess Mental Workload in a Simulated Helicopter Environment.” *Proceedings of the 2020 IEEE*

International Conference on Human-Machine Systems (ICHMS), 2020.
<https://doi.org/10.1109/ICHMS49158.2020.9209549>.

- Mund, D., Pavlidis, E., Masters, M., and Schulte, A. “A Conceptual Augmentation of a Pilot Assistant System with Physiological Measures.” *Proceedings of the 3rd International Conference on Intelligent Human Systems Integration (IHSI)*, 2020, pp. 959–965. https://doi.org/10.1007/978-3-030-39512-4_146.
- Masters, M., Donath, D., and Schulte, A. “An Exploratory Analysis of Physiological Data Aiming to Support an Assistant System for Helicopter Crews.” *Proceedings of the 2nd International Conference on Intelligent Human Systems Integration (IHSI)*, 2019, pp. 744–750. https://doi.org/10.1007/978-3-030-11051-2_113.

Contents

Abstract	iii
Kurzfassung	iv
Acknowledgments.....	v
List of Tables	x
List of Figures	xi
List of Abbreviations	xiii
1 Introduction	1
1.1 Theory to Practice: Operationalizing Mental Workload	5
1.2 Real-Time Pilot Monitoring for Mental State Estimation: Current State of the Art.10	
1.3 Summary of Potential Physiological and Behavioral Signals for Evaluating Pilot Mental State	12
2 Real-Time Acquisition and Processing of Selected Physiological and Behavioral Signals.....	18
2.1 Functional Near-Infrared Spectroscopy (fNIRS)	19
2.1.1 fNIRS Data Acquisition and Pre-Processing	20
2.1.2 fNIRS Feature Extraction	23
2.1.3 fNIRS Sensitivity to Head Position	27
2.2 Eye Tracking	29
2.3 Respiration	32
2.4 Electrocardiography (ECG).....	34
2.4.1 ECG Data Acquisition	35
2.4.2 ECG Peak Detection	36
2.4.3 ECG Feature Extraction	38
2.5 Electrodermal Activity (EDA)	42
2.5.1 EDA Data Acquisition	43
2.5.2 EDA Feature Extraction.....	43

3	PhysHub: A Pilot Physiological Monitoring and Mental Workload Prediction Tool	46
3.1	Visual Inspection of System and Pilot State	48
3.2	Transparent Triggering System	49
3.3	Baseline Collection	50
3.4	In-Cockpit Pilot Interface to PhysHub	51
4	Experimental Testing of a Real-Time Multimodal Mental Workload Prediction System During Simulated Helicopter Flight	53
4.1	Introduction	53
4.2	Methods	55
4.2.1	Simulated Mission Design	57
4.2.2	Subjective Mental Workload Assessment	59
4.2.3	Physiological and Behavioral Data Acquisition and Baseline Measurement	62
4.2.4	Real-time Mental Workload Prediction through Supervised Learning	64
4.2.5	Post-Flight Questionnaire and Data Analysis	70
4.3	Results	71
4.4	Discussion	80
5	Conclusion	84
5.1	Summary of Contributions	85
5.1.1	Theoretical and Scientific Contributions	85
5.1.2	Practical Application Contributions	87
5.2	Future Work	88
5.2.1	Longitudinal Studies of Pilot Physiological Data and Subjective Mental Workload	88
5.2.2	Further Exploration of fNIRS Features	88
5.2.3	Suitability Studies of fNIRS in Real Flight Conditions	89
5.2.4	Integration into an Assistant System and Acceptability Studies	89

Bibliography	91
Appendices	97
Appendix A Output Variables Broadcast by the Proprietary SmartEye Pro Software ...	98
Appendix B Custom-Built ECG and EDA Processing GUI	99
Appendix C Study Participant Consent Form.....	100
Appendix D Pre- and Post-Experiment Questionnaire Results.....	101
Appendix E Selectable Pages of the Multi-Function Display	108
Appendix F Pre-Brief Slides Presented to Participants Before Mission Execution	110
Appendix G Conditions and Triggered Actions for Simulated Missions	111

List of Tables

Table 1.1	A sampling of published research aiming to assess mental workload.....	13
Table 1.2	Key findings/conclusions of published original research	14
Table 1.3	Key Findings/conclusions of published review articles.....	16
Table 1.4	Summary of Modality Utility.....	17
Table 2.1	Features extracted from fNIRS data.....	26
Table 2.2	Features extracted from the eye-tracking system.....	31
Table 2.3	Features extracted from the respiration system.....	33
Table 2.4	Features extracted from the ECG system.....	39
Table 2.5	Features extracted from the EDA system.....	43
Table 4.1	Participant Summary	55
Table 4.2	Mission Design - Elicitation of Various Mental States.....	58
Table 4.3	Summary of Extracted Features	63
Table 4.4	Complete mental workload correlation statistics for each participant.....	74
Table 4.5	Complete summary statistics of all mental workload correlation coefficients	75
Table 4.6.	Focused summary of the correlation analysis between subjective and predicted mental workload	75
Table 4.7	Correlation summary statistics between subjective mental workload and individual features.	77
Table A.0.1	Output variables from the proprietary SmartEye Pro software.....	98
Table A.0.2	Participant responses to the pre-experiment questionnaire	101
Table A.0.3	Participant responses to questions 1-4 of the post-experiment questionnaire...	102

List of Figures

Fig. 1.1 Multiple human-machine architectures depicting various relationship structures	3
Fig. 1.2 Plot showing a relationship between arousal and performance.....	6
Fig. 1.3 A model depicting the relationship between “task load” and “mental workload.”	7
Fig. 1.4 A diagram showing the four primary methods for measuring mental workload.....	8
Fig. 2.1 Optode layout of the 18-Optode forehead sensor pad from fNIR Devices	21
Fig. 2.2 A general depiction of the fNIRS signal acquisition and processing chain.	21
Fig. 2.3 Block diagram of fNIRS processing chain.	22
Fig. 2.4 fNIRS light intensity data being distorted by the eye tracking system	22
Fig. 2.5 A representative plot showing the effect of head tilt on fNIRS data.....	27
Fig. 2.6 A depiction of the IMU data acquisition and processing chain.....	28
Fig. 2.7 Various elements of the eye-tracking system.	29
Fig. 2.8 A depiction of the eye-tracking feature extraction processing chain.	30
Fig. 2.9 Representative plot of various eye-related features.....	31
Fig. 2.10 Custom-built double-strap respiration sensor.....	32
Fig. 2.11 Respiration processing chain diagram.....	33
Fig. 2.12 Time series plot of processed respiration data.....	34
Fig. 2.13 A depiction of the ECG feature extraction processing chain.	35
Fig. 2.14 Electrode placement yielding a Lead II ECG recording.....	35
Fig. 2.15 Representative plots of the ECG data sampling issue and its correction.	36
Fig. 2.16 Raw ECG recording with R-wave peaks highlighted.....	37
Fig. 2.17 Visualization of the “valid” ECG peak detection algorithm.	38
Fig. 2.18 Power Spectral Density of the IBI signal.	40
Fig. 2.19 Representative plot of ECG-extracted features	41
Fig. 2.20 EDA event profile following a sudden and surprising stimulus.....	42
Fig. 2.21 A depiction of the EDA feature extraction processing chain.	43
Fig. 2.22 Representative plot of EDA-extracted features	45
Fig. 3.1 The Phys Hub graphical user interface (GUI).....	48
Fig. 3.2 In-cockpit display of physiological data and predicted mental workload.....	51
Fig. 4.1 A photograph of the helicopter simulator used in this study.....	56
Fig. 4.2 High-level experiment design block diagram.....	57
Fig. 4.3 Maps of the two simulated helicopter missions designed to elicit varying levels of mental workload.	58

Fig. 4.4 Tool developed to gather a participant’s subjective mental workload post-flight.	61
Fig. 4.5 All sensors worn by each participant during simulated missions.....	62
Fig. 4.6 Extracted signals and features over the course of a simulated helicopter mission.....	64
Fig. 4.7 Correlation coefficient matrix showing the linear relationship between subjective mental workload and various physiological signals.....	67
Fig. 4.8 Procedure for fitting and applying a linear regression model to two consecutive simulated helicopter missions.....	68
Fig. 4.9 Default model weights used to predict mental workload for each study participant during the first of two simulated helicopter missions.	69
Fig. 4.10 Subjective and predicted mental workload over the course of simulated helicopter missions	73
Fig. 4.11 Grouped scatter plots of the correlation between subjective and predicted mental workload	76
Fig. 4.12 Paired box-and-whisker plots of the linear correlation between individual features and subjective mental workload.	78
Fig. 4.13 Time-synchronized plots of predicted mental workload centered on notifications of high workload.....	79

List of Abbreviations

Abbreviation	Definition
ANF	Adaptive IIR notch filter
ATC	Air Traffic Controller
CNN	Convolutional Neural Networks
DFP	Differential Pathlength Factor
ECG	Electrocardiogram or electrocardiography
EDA	Electrodermal activity
EEG	Electroencephalogram or electroencephalography
EOG	Electrooculogram or electrooculography
fNIRS	Functional near-infrared spectroscopy
HAT	Human-autonomy teaming
HHb	Deoxygenated hemoglobin
HMT	Human-machine teaming
HR	Heart rate
HRV	Heart rate variability
IBI	Inter-beat-intervals
IMU	Inertial measurement unit
LDA	Linear Discriminate Analysis
LF/HF	The ratio of low-frequency to high-frequency power
LSL	Lab Streaming Layer
mCHR	Modified Cooper Harper rating scale
NASA-TLX	NASA Task Load Index
O ₂ Hb	Oxygenated hemoglobin
PFC	Pre-frontal cortex
PNS	Parasympathetic nervous systems
RMSSD	Root mean square of successive heartbeat interval differences
SI	Saccadic intrusions
SNS	Sympathetic nervous systems
SVM	Support Vector Machine
SWAT	Subjective Workload Assessment Technique
TCP	Transmission Control Protocol
UPE	Unexplained physiological episode
WL	Waveform length

1 Introduction

The concepts of human-centered systems [1] and subsequently human-centered automation [2], [3] were born from the belief that humans possess unique capabilities that can be augmented by various tools, including automation, to enhance overall system effectiveness. One such system, requiring acute synergy between humans and machines is the aircraft. The highly complex system, dynamic environment, and severity of potential error necessitate a close coupling between the pilot and the aircraft's digital and mechanical systems. It is here in the aviation space that these concepts of human-centered systems and human-centered automation have largely been developed and refined. In addition to the highly-focused activity in the aviation space, these principles have been studied and applied in a wide range of fields to maximize system usability and overall performance.

A pioneering report compiled in 1951 by Dr. Paul M. Fitts presents many of the fundamental principles in the field still discussed today [4]. One such principle is that systems should be designed to account for the respective strengths and weaknesses of the human and machine sub-systems. Included in this report are lists of functions or capabilities he argued were either better suited for humans or machines. He offered, however, that any arrangement of functions between humans and machines “must, of course, be hedged with the statement that we cannot foresee what machines can be built to do in the future.” This is an important caveat as it accounts for future technological developments changing the distribution of functions between humans and machines. His assessment for example, that “sensory functions” such as identifying objects in a scene or hearing a faint noise are best performed by man is arguably no longer true due to advancements in sensing technologies since his writing. Regardless of the allocation, however, the principle maintains its validity – human-operated systems should be designed to account for the relative strengths and weaknesses of the human and the mechanical sub-systems.

Since the publication of Dr. Fitt’s capability allocation between humans and machines, the number of tasks for which human capabilities outperform those of machines has decreased. Subsequent research by Dr. Charles E. Billings at the National Aeronautics and Space Administration (NASA) largely reduced the tasks for which humans outperform machines to those related to high-level cognitive functions. Published in 1991, Billings suggested the “invaluable attributes” maintained by human operators are 1) their ability to detect a signal in a noisy environment 2) their ability to effectively reason in the face of uncertainty, and 3) their abilities of abstraction and conceptual organization [2]. He reasoned these invaluable attributes were precisely those essential to anyone wishing to pilot an aircraft. Regardless of how the unique abilities of humans are defined, it is clear one’s ability to utilize these and other similar cognitive abilities is not constant. Many factors influence one’s ability to perform these functions such as fatigue, fear, or mental workload [5]. Additionally, it is generally accepted that there are limits to human ability to process information and that information overload can lead to degraded performance in cognitive tasks (note that the “invaluable attributes” maintained by human operators are all cognitive abilities) [6]. This work aims to address an often-overseen aspect of human-machine systems – that of monitoring the human operator’s state to ensure they remain capable of fulfilling their uniquely-human roles. To harness the maximum utility of human pilots, they ought to be monitored and aided in a way that supports their ability to remain in a state of maximum cognitive function. For example, states of high anxiety or stress should be identified and controlled as these states are less conducive to effective cognitive functioning [7].

In many fields, technological advancements in areas such as sensing, computing, robotics, machine learning, and artificial intelligence have resulted in both software and hardware-based tools taking on a more teammate-like relationship with their users. In the field of military aviation, research and development efforts into human-autonomy teaming (HAT) and human-machine teaming (HMT) technologies have led to multiple national militaries pursuing “loyal wingman” programs in which one or more unmanned aircraft teams with a human-piloted aircraft to pursue a common objective. In this relationship, the human is not to simply use the machine as a tool, but rather, the two are to develop a cooperative relationship in the pursuit of a desired outcome. Using the nomenclature established by Schulte et al. in [8] to describe various human autonomy teaming architectures, Fig. 1.1 depicts multiple human-machine system architectures in which the relationship between the human and the

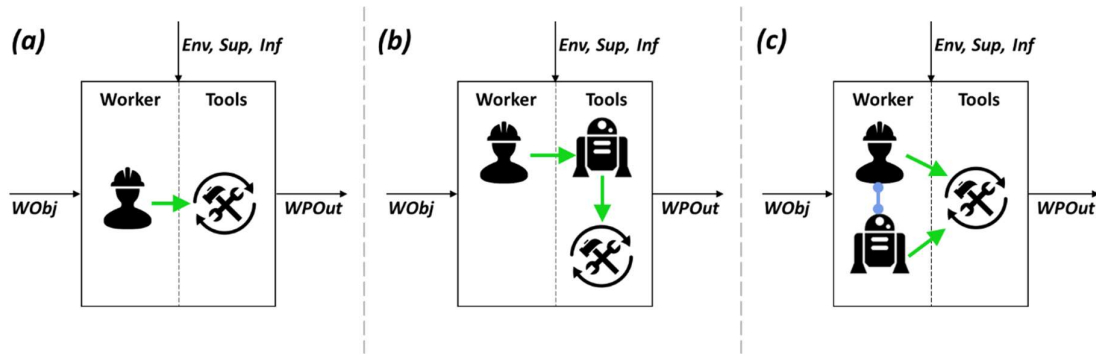


Fig. 1.1 Multiple human-machine architectures depicting various relationship structures between the human and technological systems where a green arrow represents a hierarchical relationship and a blue line represents a cooperative relationship. (a) depicts a hierarchical relationship between the human and basic tools. (b) introduces a “cognitive agent” capable of directing lower-level tools which is yet directed in a hierarchical relationship. (c) depicts a cognitive agent working cooperatively with the human to achieve the given objective. Adapted with permission from [8].

technological systems varies from a purely hierarchical relationship (a) to one of cooperation with a “cognitive agent” working together with the human to achieve the given objective (c).

These system architectures can be used to analyze human-machine relationships in a wide variety of applications. This work focuses on one application of a human-machine system in which the machine could cooperatively support the human in the pursuit of a common objective – namely a piloted military aircraft. In the highly complex and rapidly changing environment of flight, especially true in the military setting, the “adaptive assistant system” would act as a co-pilot to the human pilot assisting in the planning and execution of a successful flight. The ultimate realization of such an adaptive assistant system supporting the human crew of an aircraft would be the actualization of R2-D2 from the movie *Star Wars*. How such systems ought to behave is a question of much research and ethical debate. One proposed structure for the design of cognitive automation given by Onken and Schulte [9] suggests an assistant system should take a tiered approach to assisting with the goal of minimizing assistance. First, if deemed necessary, the system should direct the human operator’s attention to a task. Next, if unsuccessful in its design, the system should simplify the task in some way enabling the human operator to accomplish the task. Finally, as a last resort, the tasks should be allocated to the assistant system and completed autonomously.

In addition to the fundamental principles guiding the behavior of the autonomous system, the particulars of its implementation are critically important for achieving the effective and necessary relationship between the pilot and the machine. For example, the timing and manner in which such a system provides feedback or intervention is critical to its effective realization. An inappropriately timed interruption can negatively impact an operator’s mental state and increase human error rates [10]. To inform the timing and manner in which feedback

or intervention is provided, significant contextual knowledge is required. Context-aware operator feedback and notification systems have been shown to improve coordination and performance between human-machine teams [11]. In the realm of military aviation, the “context” of the system at any given moment is highly complex with many clearly defined system states and others which are more ambiguous. Examples of states which are clearly defined are groundspeed, windspeed, angle of attack, the position of the landing gear, enemy position, etc. Even the binary state of whether or not a pilot is currently speaking with their copilot is clearly defined and can be determined. Other elements of the system’s “context” which are more ambiguous and less easily measured include the pilot’s current mental picture of the situation and their surroundings (i.e. their situational awareness), and their cognitive strain or mental workload. Despite the difficulty, it is hypothesized that obtaining even an approximate assessment of the pilot’s mental state would significantly enhance a system’s ability to provide relevant and useful context-sensitive assistance.

As anyone who has sat in the passenger seat of a car knows, comprehending the mental state of the driver is critical for providing productive assistance to the driver. However, even for the most highly trained psychologists, teachers, or parents, assessing the mental or cognitive state of another human being is a very complex and difficult task. Furthermore, acting on incorrect conclusions as to a person’s mental state often leads to confusion, frustration, and ultimately a less-than-optimally functioning system. This is true in a human-to-human relationship as it is in a human-to-machine relationship.

The challenges associated with assessing a pilot’s mental workload and acting upon that assessment are steep, yet the potential benefits are significant. A review article published in 2018 which assessed a random sampling of over 200 commercial air transport accidents and incidents from 2000 to 2016 reported that “human factors contribute to approximately 75% of aircraft accidents and incidents” and that situational awareness and non-adherence to procedures were the most significant factors contributing to these incidents [12]. If pilots could be supported by a never-sleeping co-pilot with super-human observation skills, many of these accidents and incidents could be avoided and the efficiency of the crew could be improved.

To move in this direction, this work pursues the idea of introducing psychophysiological measures into the cockpit’s human-machine interface providing the system with a view into the operator’s mental state, specifically their cognitive or mental workload. Psychophysiological measures refer to the subset of physiological signals that are

influenced by psychological or emotional states [13]. The integration of physiological measures and signals into computer systems and applications to provide valuable information and enable adaptive computer interfaces or systems is known as “physiological computing” [14]. Specifically, this work investigates the utility of various psychophysiological monitoring technologies and processes to infer the mental workload of an operator during simulated helicopter flight. It is anticipated that with an accurate model of the pilot’s mental workload, an adaptive assistant system could be developed which could optimize the human-machine team. Due to the extreme complexity of a person’s mental workload, a multi-modal approach is pursued in which various signals and features are extracted to shed light on this construct. By integrating many sensing technologies and methods, it is hypothesized that one can generate a more complete picture of the person’s mental workload than by any one method alone.

1.1 Theory to Practice: Operationalizing Mental Workload

If the goal of a human-machine system is to optimize performance, it is reasonable to question why some metric of system performance could not be assessed in real-time and used as feedback in the system to correct for errors. Certainly, this is an appropriate tactic when the task is simple and well-defined. If, for example, the task is for a pilot to maintain a specific speed, heading, and altitude, a simple control system can be built to detect and trigger corrections when deviations from the task objective are observed – and yes, such auto-pilot systems exist in abundance. However, in highly complex and dynamic situations where “performance” is ill-defined and correction is not possible via a simple control loop, an assistant system could support most by evaluating the state of the human operator and assisting when appropriate. In fact, in these situations where uncertainty is guaranteed and the cognitive abilities of the human operator are the critical enabler for mission success, it could be argued that the cognitive state of the operator is precisely the metric or state to be controlled.

Pioneering work by Yerkes and Dodson in 1918 with mice demonstrated a relationship between task difficulty, arousal (stimulus intensity), and learning rate [15]. With an easy task, where objects to be differentiated significantly differed in brightness, the speed of learning increased with increasing stimulation intensity. With medium and high-difficulty tasks, where the difference in brightness between objects was not as significant, the fastest learning occurred at some intermediate level of stimulation intensity whereas “the weak and strong stimuli were less favorable to the acquirement of the habit than the intermediate stimulus.” In these more difficult tasks, the majority of mice given a weak-intensity stimulus did not learn the intended

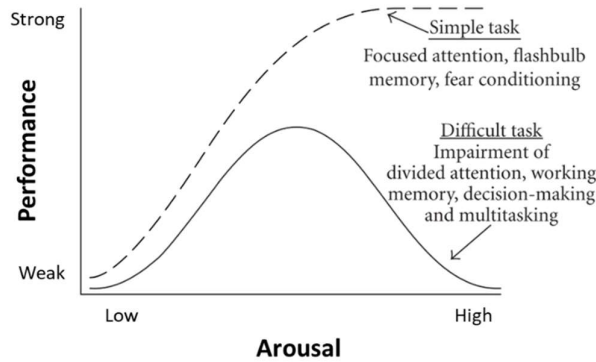


Fig. 1.2 Plot showing a relationship between arousal and performance which has come to be called the “Yerkes-Dodson-Law” originally presented in 1908 suggesting the quality of performance on difficult tasks (not simple tasks) is an inverted U-shaped function of arousal [15] (the plot is was published in 2007 in [16]). Although not synonymous, “arousal” has been largely replaced by “mental workload” in modern human factors and psychology vernacular.

behavior, all of the mice given a medium-intensity stimulus learned the behavior quickly, and the mice given a strong-intensity stimulus performed similarly to those who received the weak-intensity stimulus and did not learn the intended behavior. This relationship between “arousal” and performance is shown in Fig. 1.2 (published in 2007 in [16]). Although initially studied with mice and a particular visual learning task, this relationship between arousal and performance has become widely accepted for explaining human behavior in a variety of environments. In modern human factors and psychology literature, the term “arousal” has largely been replaced by the more general term “mental workload.” In general, when confronted with a difficult task, too little or too much mental workload experienced by the human operator yields sub-optimal performance. It then follows that a system that could reliably maintain an operator’s mental workload at some optimal level would aid in producing maximum performance. Even if not capable of maintaining an optimal level of mental workload, such systems could prevent overload and underload situations where poor performance is likely. Such systems are known as adaptive assistant systems or systems built with adaptive automation capabilities.

Before further discussing adaptive assistant systems, however, it is important to better understand mental workload. A model developed by German occupational science researchers Rohmert and Rutenfranz in 1975 known as the “Belastungs-Beanspruchungs-Modell” (translated as the “Stress-Strain-Model”) does well to establish a fundamental model of mental workload [17]. The English words “stress” and “strain” are not helpful in this discussion as they are often used interchangeably and with varying meanings. Rather, “Belastung” is translated as “task load” and “Beanspruchung” is translated as “mental workload.” A general depiction of this model is provided in Fig. 1.3. The model provides a relationship between the

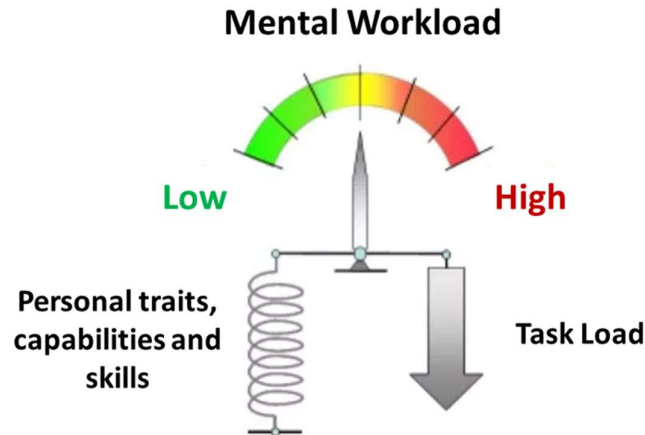


Fig. 1.3 A model depicting the relationship between “task load” and “mental workload.” This relationship was presented as the “Belastungs-Beanspruchungs-Modell” (translated as the “Stress-Strain-Model”) in [17].

task load, the unique traits of the operator, and the mental workload experienced by that person. As can be visualized in the figure, the same task load can induce varying levels of mental workload in a population of operators based on their unique traits, capabilities, skills, and other personal factors.

This model highlights multiple important aspects of mental workload. It suggests that mental workload is experienced uniquely by an individual and cannot be deduced through an analysis of the task load alone. The model depicts well the argument made by Durantin et al. in [18] that mental workload should be defined “in terms of the interaction between the task and the individual performing the task.” Mental workload may be defined as the subjective experience of a given task load.

Given the nebulous nature and complexity of mental workload, yet desirous to quantify the measure for various practical applications, researchers have developed or implemented a wide range of measurement techniques. Most of these measurement techniques can be categorized into four means of assessment: physiological measurements, behavioral actions, task performance, and subjective ratings. These four primary methods for measuring mental workload are depicted in Fig. 1.4. Each may provide a unique perspective into the true nature of subjective mental workload. Physiological measurements capture the objective response of the body to a given workload. Examples of these measurements include (but are not limited to): heart rate (HR), electroencephalography (EEG), electrodermal activity (EDA), respiration rate, body temperature, and pupil size. Behavioral measurements capture physical actions or movements of the body such as eye fixation location, grip strength, posture, linguistics (tone, vocabulary), and facial expression. Next, insight into one’s mental workload can be gleaned by

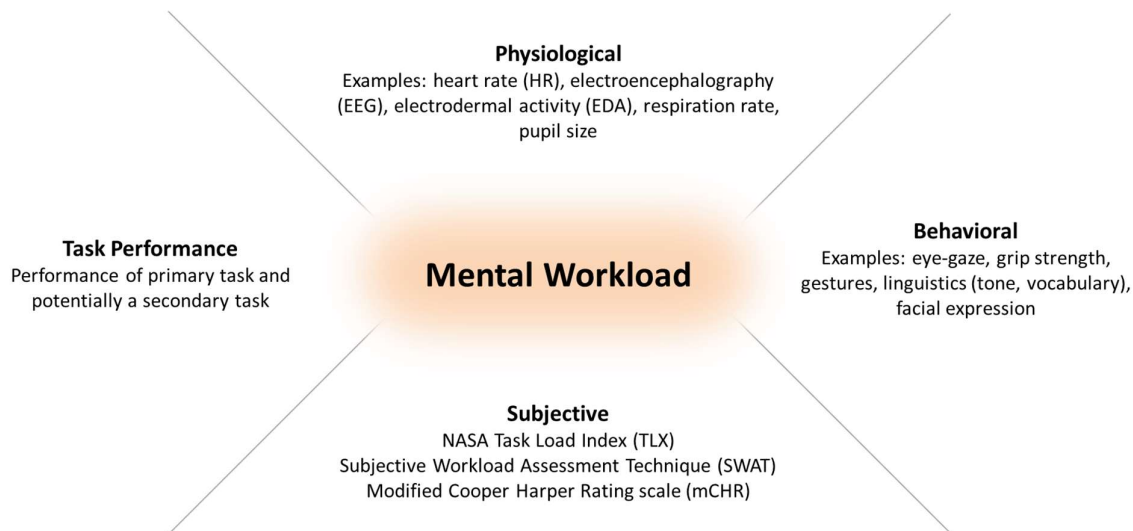


Fig. 1.4 A diagram showing the four primary methods for measuring mental workload. The figure is an adaptation of one published in [87]. The lack of a boundary around “Mental Workload” represents the nebulous nature of the construct.

monitoring the execution performance of a task. Generally, if task performance declines (while other factors such as commitment to the task remain constant), one can conclude an increase in mental workload. Obtaining a metric for real-time task performance is possible for simple tasks such as station-keeping while flying, but may not be feasible in a highly dynamic scenario with ill-defined performance parameters. If performance on the primary task cannot be continuously assessed, another strategy to assess mental workload is to evaluate the performance of a secondary task for which performance can be measured. Finally, subjective ratings can provide a unique perspective into a person’s mental workload. Multiple published methods for collecting subjective mental workload metrics exist including the NASA Task Load Index (TLX), Subjective Workload Assessment Technique (SWAT), and the Modified Cooper Harper Rating scale (mCHR). By virtue of the method, these questionnaire-based techniques cannot obtain a continuous metric throughout task execution. If implemented mid-task, they are disruptive and often counterproductive. Additionally, they can be overly complicated for the untrained subject.

Accepting the idea that mental workload is a result of the given task (and all relevant environmental factors) as well as the unique characteristics, thoughts, and perceptions of the individual performing that task, practitioners aiming to assess mental workload are confronted with a significant challenge. Previous work by Honecker et al. has investigated a task-based approach to assess the state of the pilot in the context of a helicopter simulator [19]–[21]. By mapping gaze fixation location to elements of the display as well as by monitoring the physical manipulation of the cockpit by the pilot, “evidences” could be generated suggesting the

execution of various tasks which were utilized in an online task recognition system. The utility of the resulting task prediction is highly reliant on a complete and accurate task model capturing all tasks which could possibly be performed by the user. Assuming an accurate task prediction (an accurate assessment of the task load), an operator's mental workload was predicted by accumulating subjectively-inferred and static mental resource demands of the currently-executed tasks. Additionally, the group assessed whether or not simultaneously-executed tasks conflicted with one another in a way that would compound the resulting mental workload. As with the reliance on a comprehensive task model, the workload prediction step in this approach relied on an accurate mapping of task demand to workload. Considering Fig. 1.3, this task-centered approach discounts the effect of the unique traits, skills, emotions, or experiences of the operator on the experienced mental workload. Arguably, some of these unique operator capabilities and attributes influencing mental workload (such as pilot training level) could be learned and used to create operator-specific task-to-workload prediction models as I and my colleagues have previously suggested (see [22]). Doing so, however, would likely fail to incorporate the wide variety of unpredictable and rapidly-changing conditions such as the pilot's emotional state which would likely make this an unfruitful endeavor.

Although the model depicted in Fig. 1.3 highlights appropriately the idea that mental workload is a function of more than task load alone and should encourage a practitioner to move beyond a purely task-based approach to mental workload prediction, the model does not capture the idea that the personal factors of the operator (represented by the spring) can vary with time. In other words, some personal factors contributing to mental workload can change significantly throughout task execution. Factors such as a person's commitment to a high level of performance (and hence their level of effort) for example may not be constant over time resulting in a varying level of mental workload throughout task execution given a constant task load. Veltman and Jansen in [23] support this reasoning suggesting the relation between task load, mental workload, and operator performance is unique for each individual and is time-variant. They suggest the relation can change due to talent, training, fatigue, external stressors, or other factors.

Rather than relying on fixed task and workload models, it is argued in this work that a more reliable method of determining mental workload is through the real-time monitoring of psychophysiological signals. The approach taken in this work to quantify a person's mental workload is not to assess task load nor the individual's traits, capabilities, and skills, but rather to infer the person's mental state through an analysis of various psychophysiological signals

which correlate with the person's mental workload. Thus, the strategy taken here is to assess mental workload by observing the resulting physiological state induced by a task load given the operator's unique traits, capabilities, and skills.

1.2 Real-Time Pilot Monitoring for Mental State Estimation: Current State of the Art

Real-time monitoring of pilots in the cockpit has been a growing line of research for decades. Although pursued for reasonable and significant purposes, the assessment of the pilot's physical state is not a concern of this work. Rather, this work is focused on assessing the pilot's mental state, specifically their mental workload, to optimize the human-machine team. Although a subject of research for decades, real-time pilot monitoring technologies remain largely confined to academic environments. This is especially true when the review is limited to real-time pilot mental state or mental workload assessment technologies. The following section will present the current state-of-the-art of this field in the academic, commercial, and military settings.

Many academic pursuits are documented in which sensing technologies are applied to simulated and actual flight scenarios to measure or assess pilot mental workload. The subsequent section, Section 1.3, provides a systematic review of many of these publications highlighting the sensing techniques employed to assess mental workload. As can be expected, most of the academic work in the field has been conducted in simulators of various fidelity, yet multiple have been conducted in real aircraft and during actual flight. Although these articles highlight unique challenges with actual flight, they consistently reaffirm the early finding reported in [24] that physiological responses to changes in mental workload are similar across simulated and actual flight [25]–[29].

Rather than reporting a continuous-valued metric of pilot mental workload over time, many of the recent publications classified or predicted discrete mental workload states (primarily “low” vs “high”) [25]–[28], [30], [31]. The reported classification accuracies of these efforts range from 67% ([27]) to 87% ([25]). These results demonstrate the fair, but imperfect ability to use physiological data to classify mental states. Perhaps it is the case, that existing published literature does not present a continuous-valued metric of pilot mental workload over time due to the difficulty in validating such a metric. One can experimentally set the relative difficulty of consecutive tasks and validate through a post-execution survey (e.g. NASA-TLX [32]) that the tasks were perceived as designed. It is more difficult, however,

to validate a task's difficulty continuously throughout task execution. Further details as to the academic state of this research are provided in Section 1.3.

Perhaps due to the high variability of results across these academic efforts, the commercial sector for supplying real-time pilot monitoring solutions remains weak. Designers and manufacturers of commercial aircraft (particularly the cockpit) largely do not consider the monitoring of the pilot into the design of the aircraft forcing research groups and commercial sensing companies to implement custom integration solutions. Despite the difficulty of integration, many stand-alone commercial sensing technologies for real-time pilot monitoring do exist. For example, the Canary Pilot Physiological Monitoring System by the Israeli company Elbit Systems is advertised as a non-invasive system fully integrated into a pilot's helmet capable of providing pilots with early warning of developing hypoxia and G-Force Induced Loss of Consciousness (G-LOC). Additionally, when integrated with the aircraft's flight controls, engages the autopilot engagement if the pilot loses consciousness. Wearable health monitors capable of measuring cardiovascular and respiratory features are also commercially available and could potentially be utilized for real-time pilot monitoring through custom integration solutions.

One ongoing project led by the U.S. Air Force seems to be bridging the gap between the laboratory and operational environments. The Integrated Cockpit Sensing (ICS) program aims to incorporate flight environment and pilot physiology data to improve pilot safety and performance [33]. Specifically, the program aims to assess pilot state through the measurement of respiratory and cardiovascular measurements and detect conditions conducive to unexplained physiological episodes (UPEs). According to the 2020 National Commission on Military Aviation Safety report, from 2013 to 2018, 718 UPEs were reported in fighter and trainer aircraft in which pilots across many airframes report unexplainable hypoxia-like symptoms [34]. The ICS program began in June 2020 and will complete its initial development and testing and November 2023. In a personal interview on 30 November 2022, the ICS Program Manager disclosed that the program aims to develop five prototype systems which will be distributed across three target customers: Air Education and Training Command (AETC), the 412th Operations Group within Air Force Materiel Command (AFMC), and Air Combat Command (ACC). The system is intended to be flown on AETC's T-6 Texan II trainer aircraft, ACC's F-16 Fighting Falcon fighter aircraft, and support the various testing activities of the 412th Operation Group and its flight test squadrons. Although the primary function of the system will be to enable post-flight analysis of aircraft and pilot state data, it is anticipated

that the system will provide real-time notifications to the pilot of various states (presumably those related to hypoxia such as low blood saturation). Determining how and when these notifications will be presented to the pilot remains an ongoing effort. The ICS sensing suite includes the Canary Pilot Physiological Monitoring System by Elbit Systems discussed previously.

In summary, a significant amount of research has been conducted over the last few decades showing strong correlations between physiological measures and various mental workload states, yet whether in commercial industry or the military environment, no fielded system currently exists to enable the real-time mental workload prediction of pilots. Such a system could have significant utility for training purposes, for optimizing the safety and performance of pilots and their crews, and for the development of pilot-assistant systems capable of providing context-sensitive assistance.

1.3 Summary of Potential Physiological and Behavioral Signals for Evaluating Pilot Mental State

As presented previously, the four primary methods for assessing mental workload are through physiological measurements, an assessment of behavioral actions, evaluating task performance, and through subjective ratings. This section presents a systematic review of previously-published original academic research which employed these methods in various ways as well as a summary of multiple review articles on the topic. Table 1.1 presents a summary of original research sorted by publication year showing the environment in which the study was conducted (simulated flight, real flight, or other), the number of participants included in the study, and which measurements were collected to evaluate mental workload. Table 1.2 highlights the main findings of these selected articles as it relates to the evaluation of mental workload. Table 1.3 provides the conclusions reached by published review articles assessing the utility of various psychophysiological signals to measure mental workload. Finally, summarizing the findings of the original research articles as well as the review articles, Table 1.4 provides a summary of sensing modality utility across multiple physiological and behavioral collection methods in supporting a real-time mental workload assessment system in a cockpit. Together, the data and analysis presented in these tables provide a valuable picture of the current state of academic research on this topic.

Table 1.1 A sampling of published research aiming to assess mental workload through physiological, behavioral, and subjective measures. Terms: HR-heart rate, HRV-heart rate variability, EDA-electrodermal activity, EEG-Electroencephalography.

Citation	Year	Type	Participant Count	HR	HRV	Blood Pressure	Blink Features	Eye Movement	Pupil Size	Saccade Features	Fixation Features	EDA	EEG	fNIRS	Respiration	Saliva cortisol	Subjective
Roscoe et al. [24]	1988	Sim	9	x													x
Hankins et al. [35]	1998	Real	15	x			x						x				x
Miyake et al. [36]	2001	Other*	12		x							x					x
Wilson [37]	2002	Real	10	x	x		x					x	x				x
Veltman [29]	2002	Sim & Real	20	x	x	x	x								x	x	x
Lee et al. [38]	2003	Sim	10	x													x
Nickel et al. [39]	2003	Computer	14		x												
Di Nocera et al. [40]	2007	Sim	10								x						x
Dehais et al. [41]	2008	Real	6						x		x						
Kikukawa et al. [42]	2008	Real	4											x			
Girouard et al. [43]	2009	Packman	9											x			x
Luigi et al. [44]	2010	Sim (drive)	18							x							x
Power et al. [45]	2010	Other**	10											x			
Dahlstrom et al. [46]	2011	Real	7	x			x	x					x				x
Tokuda et al. [47]	2011	N-Back	14						x	x							
Durantín et al. [18]	2014	Computer	12		x									x			
Herff et al. [48]	2014	N-Back	10											x			
Derosière et al. [49]	2014	Computer	7											x			
Gateau et al. [50]	2015	Sim	19											x			
Dehais et al. [51]	2015	Sim	7							x							
Causse et al. [52]	2016	Computer	24						x				x				
Mansikka et al. [53]	2016	Sim	26	x	x												
Aghajani et al. [54]	2017	N-Back	17										x	x			
Causse et al. [30]	2017	Sim	26											x			x
Hidalgo-Muñoz et al. [55]	2018	Sim	21	x	x												
Scannella et al. [31]	2018	Real	11	x	x					x	x						x
Dehais et al. [25]	2018	Sim & Real	4										x	x			
Gateau et al. [26]	2018	Sim & Real	28											x			
Verdière et al. [27]	2018	Sim	12											x			

Citation	Year	Type	Participant Count	HR	HRV	Blood Pressure	Blink Features	Eye Movement	Pupil Size	Saccade Features	Fixation Features	EDA	EEG	fNIRS	Respiration	Saliva cortisol	Subjective
Dehais et al. [28]	2019	Real	22										x				
Alaimo et al. [56]	2020	Sim	23		x												x
Hebbar et al. [57]	2021	Sim	12						x		x		x				

*Puzzle, tracking, and numerical logic

**Mental arithmetic and music imagery

Table 1.2 Key findings/conclusions of published original research aiming to assess mental workload through physiological, behavioral, and subjective measures. Abbreviations: EDA-electrodermal activity, EEG-Electroencephalography, fNIRS-functional near-infrared spectroscopy, HR-heart rate, HRV-heart rate variability.

Citation	Key Findings/Conclusions
Roscoe et al. [24]	Failures in simulated flight induce a similar HR response as failures in real flight.
Hankins et al. [35]	Multiple psychophysiological measures provide a “comprehensive picture” of the mental demands of flight. It may be possible to develop systems that provide on-line monitoring of mental workload that can provide feedback to the pilot and aircraft systems. Blink rate is sensitive to visual demand. EEG theta band power is sensitive to mental workload. HR is not specific to demand resource.
Miyake et al. [36]	Feelings of one’s performance may influence assessments of subjective mental workload yet the correlation between such feelings and the physiological responses during the task may be low.
Wilson [37]	Cardiac and electrodermal measures are highly correlated and exhibit changes in response to flight demands. HRV is less sensitive than HR. EEG alpha and delta bands show significant changes to varying demands. Blink rate decreases during visually demanding flight.
Veltman [29]	HR, HRV, and respiratory rate behaved similarly across simulated and real flight. A combination of blink parameters provides more information about mental workload than blink rate alone. Cortisol was not affected by simulated flight yet increased greatly following real flight suggesting that cortisol is not affected by mental effort.
Lee et al. [38]	HR correlates with subjective mental workload ratings (NASA-TLX).
Nickel et al. [39]	HRV is an indicator of time pressure or emotional strain - not mental workload.
Di Nocera et al. [40]	Eye fixation distribution is more variable during takeoff and landing than during level flight. Subjective ratings (NASA-TLX) correlate to high and low workload phases.
Dehais et al. [41]	Shorter fixation time on instruments and fixations on fewer instruments in high workload conditions. Average pupil diameter is larger during higher workload conditions.
Kikukawa et al. [42]	fNIRS (especially O2Hb measurements), provides a sensitive method for the monitoring of cognitive demand in helicopter pilots.
Girouard et al. [43]	Prefrontal cortex oxygenated hemoglobin decreased during activity as compared to a rest state. 94% classification accuracy between play and rest. 61% classification accuracy between easy and hard. Large inter-subject variability.
Luigi et al. [44]	Saccadic peak velocity correlates strongly with mental workload while saccade amplitude and duration do not. Saccadic peak velocity decreased as the mental workload increased.
Power et al. [45]	Prefrontal cortex fNIRS classification of mental arithmetic vs musical imagery yielded 77.2% accuracy.
Dahlstrom et al. [46]	Increased HR in high mental workload conditions during low-G flight. Blink rate and eye movement do not correlate with mental workload. EEG proved difficult to analyze due to muscle artifacts and yielded no correlation with mental workload.
Tokuda et al. [47]	More saccadic intrusions (SI) with increased mental workload. SI count is more informative than pupil diameter changes which were very minimal.
Durantini et al. [18]	Unable to discriminate between the easiest and hardest conditions using fNIRS and HRV. Prefrontal cortex oxygenated hemoglobin increases with task difficulty until the most difficult task when it decreases significantly.
Herff et al. [48]	fNIRS is used to classify task difficulty in N-back tests with a 2-class accuracy of 78% and a 4-class accuracy of 45%.

Citation	Key Findings/Conclusions
Derosière et al. [49]	fNIRS is used to classify pilot attentional states with a 2-class accuracy of 65-95% (low vs high).
Gateau et al. [50]	fNIRS is used to classify "working memory" states (easy vs hard ATC instructions) with a 2-class accuracy of 80%.
Dehais et al. [51]	The ratio of short saccades to long saccades correlates with periods of "mental conflict."
Causse et al. [52]	High "working memory load" correlates with increased pupil size and lower EEG P600 amplitude.
Mansikka et al. [53]	HR and HRV varied significantly across tasks where performance remained steady suggesting
Aghajani et al. [54]	EEG and fNIRS are used to classify N-back difficulty with 2-class accuracy of 90.9% (EEG only 85.9%, fNIRS only 74.8%). A window size of 20 seconds yields the highest accuracy.
Causse et al. [30]	Increased task difficulty resulted in an increased O2Hb and decreased HHb (for a task with two difficulty levels).
Hidalgo-Muñoz et al. [55]	Increased cognitive task difficulty as modulated by a secondary task results in an increased HR and decreased HRV.
Scannella et al. [31]	Saccade rate and HR are used to classify flight phases (takeoff, downwind, land) with 75% accuracy. The classifier was trained on one flight and applied to a second. Saccade rate helps discriminate between tasks that HR cannot.
Dehais et al. [25]	fNIRS and EEG are used to classify the first two traffic patterns ("low fatigue") from the second two patterns ("high fatigue") achieving 87% accuracy in simulated and real flight.
Gateau et al. [26]	fNIRS features used to classify ATC commands (simple or difficult) achieving 77% and 78% accuracy for simulated and real flight respectively. O2Hb concentration increased under a more difficult task load. A larger increase in HbO2 concentration in real flight than in the simulator.
Verdière et al. [27]	fNIRS connectivity features used to classify automated vs manual landings achieving 67% accuracy.
Dehais et al. [28]	Dry-electrode EEG system measured higher P300 amplitude and higher alpha and theta power in the low load condition than in the high load condition. Classification of the two states using EEG features achieved 70%.
Alaimo et al. [56]	Subjective assessment of mental workload often does not match objective measurements. For only 11 of the 23 pilots did subjective scoring (NASA-TLX) match recorded objective measurements (HRV features). The relationship between mental workload, biometric data, and performance indexes are characterized by intricate patterns of nonlinear relationships. Mental workload cannot be evaluated by subjective measures alone.
Hebbar et al. [57]	Pupil, eye movement, and EEG features are sensitive to changes in task difficulty. EEG beta and theta band power increase with increasing task difficulty.

The studies highlighted in Table 1.1 and Table 1.2 each contribute to an improved understanding of the sensitivity and diagnostic potential of these various physiological, behavioral, and subjective measurements as they relate to mental workload. It is important to recognize for example, that some measures, like heart rate, are quite sensitive to mental workload, yet lack the diagnostic potential to differentiate between the cause of the workload while others, such as eye-movement features are less sensitive to mental workload, but are specific to changes in visual demands [35]. Likewise, these studies highlight what was summarized in [37] that "the complexity of flying requires that the pilot use numerous cognitive processes, and determining the pilot's mental workload requires more than one measure. Any one measure should not be expected to give full insight into the multifaceted nature of piloting."

Table 1.3 Key Findings/conclusions of published review articles assessing the utility of various psychophysiological signals to measure mental workload.

Citation	Year	Key Findings/Conclusions
Roscoe [80]	1992	Lengthy presentation of a small number of real flight tests. Extensive section on the historical testing of HRV showing an increased mental load correlates with a decreased HRV. Suggest a high HRV may indicate the onset of under arousal or reduced vigilance. Review of respiration work since 1963. Respiration rate is sensitive to mental workload, but is often distorted by speech.
Jorna [81]	1993	HR and HRV well suited for mental state inference. The low frequency component of HRV is reduced under mentally taxing conditions.
Togo et al. [88]	2009	An increase in mental workload (“work stress” or “job strain”) was associated with decreases in the HF component of HRV and an increase in the LF/HF ratio.
Borghini et al. [60]	2014	An increased task load results in an increased EEG theta band power, a decreased EEG alpha band power, a decreased blink rate, a decreased blink duration, and an increased HR.
Charles et al. [58]	2019	No single physiological signal can provide a single true measure of mental workload (there is no “silver bullet”). Physiological signals measure the “experience” of the person. An increase in mental workload is associated with an increased HR, a reduced HRV (studies differ on which band reduces most significantly), an increased respiration rate, an increased blood pressure, a decreased P300 amplitude, a decreased EEG Alpha-band power, and an increased EEG Theta-band power. Blink frequency and duration decreases under high visual workload. EDA is sensitive to sudden changes in mental workload.
Tao et al. [89]	2019	Cardiovascular measures, eye movement measures, EEG measures, respiration measures, skin conductance, and neuroendocrine measures were assessed in 91 studies identifying 78 physiological measures. Of the 403 instances in which a particular physiological measure was used to assess mental workload, 292 (72%) showed a statistically significant relation to mental workload. Recommended the use of HR and HRV measures in future work. Concluded that most physiological measures can discriminate changes in mental workload, but they are not universally valid in all task scenarios.

Multiple review articles have been published similarly evaluating previously-published original research on the topic of assessing pilot mental workload in the cockpit. Table 1.3 summarizes the key findings and conclusions of these review articles.

The “review of reviews” presented in Table 1.3 further suggests that the scientific community has coalesced in its understanding that multiple physiological measurements are sensitive to mental workload, yet there is no “silver bullet” ([58]) for measuring a pilot’s mental workload. This conclusion further substantiates the claim made previously that a more robust assessment of a pilot’s mental workload is obtained through the measurement of multiple signals rather than relying on only one source.

Citing the original research given in Table 1.1 and Table 1.2, as well as the review articles in Table 1.3 which represent existing literature on the topic, Table 1.4 provides a summary of the utility for the sensing modalities considered in Table 1.1. Although an incomplete list of all possible sensing modalities, it represents the majority of modalities seriously considered for this use. Thus, Table 1.4 provides a summary of the general conclusions that can be made from the existing literature on the utility of these particular sensing modalities.

Table 1.4 Summary of Modality Utility in Supporting a Real-Time Mental Workload Assessment System in a Cockpit

Modality	Summary of Utility
HR	Increased mental workload → Increased HR [24], [35], [37], [46], [55], [58], [60]
HRV	Increased mental workload → Decreased HRV [37], [55], [58], [81]. HRV is sensitive to time-pressure or emotional strain, not mental workload [39]. Conflicting reports of the sensitivity of LF/HF to MWL ([81] and [88])
Blood pressure	Increased mental workload → Increased blood pressure [58]. Blood pressure is difficult to collect in real flight [29]
Blink features	Increased visual demand → Decreased blink frequency [35], [37], [58], [60]. No correlation with mental workload [46]
Eye movement	No correlation with mental workload [46]
Pupil size	Increased mental workload → Larger pupil diameter [41]. Very weak correlation [47]
Saccade features	Increased mental workload → Saccadic peak velocity decreased [44]. More saccadic intrusions with increased MWL [47].
Fixation features	Increased mental workload → Shorter fixation time, fewer instruments fixated upon [41]
EDA	Increased mental workload → Increased EDA events [37]. EDA is sensitive to sudden changes in mental workload [58]
EEG	Increased mental workload → Decreased Alpha band (8–12 Hz) [28], [37], [58], [60], decreased Theta band [28], increased Theta band [58], [60], decreased P300 amplitude [28], [58].
fNIRS	Increased mental workload → Decreased O2Hb [43], increased O2Hb [26], [30], [42], increased O2Hb then decreased O2Hb [18]. Two-state mental state classification yields fairly high classification results [26], [27], [43], [45]. Sensitivity g-forces may limit utility in real-flight conditions [90].
Respiration	Increased mental workload → Increased respiration rate [29], [58]
Saliva cortisol	Cortisol unaffected by simulated flight (only real flight) thus likely unaffected by mental workload [29]
Subjective	HR correlated with NASA-TLX scores [38]. Subjective assessment of mental workload often does not match objective measurements [56]. Nonlinear relationship between workload, physiological data, and performance indexes [56]. Mental workload cannot be evaluated by subjective measures alone [56]. Subjective assessment of mental workload is likely influenced by one's performance [36].

2 Real-Time Acquisition and Processing of Selected Physiological and Behavioral Signals

The following sections describe the signal acquisition processes implemented for individual modalities and the feature-extraction methods employed in an attempt to obtain signals relevant to the evaluation of a person's mental state. This section is not a comprehensive presentation of all possible psychophysiological signal processing chains, but rather it is a presentation of the processing chains developed for those modalities selected for further experimentation and integration into the testing environment. Additionally, each sub-section not only provides a general overview of the sensing technology but also presents the unique processing methods developed and employed in this work. The presentation of this analysis is ordered according to the location from which it is collected on the human body, from head to foot. Functional Near Infrared Spectroscopy (fNIRS) is recorded from the head, eye movement is analyzed through eye-tracking technologies, respiration is analyzed through a chest-mounted stretch sensor, an electrocardiogram (ECG) is obtained through electrodes on the chest, and finally, electrodermal activity (EDA) is monitored and analyzed through electrodes on the foot.

fNIRS was selected for further experimentation due to its unique proximity to the object of investigation – namely the brain. Compared to the other non-invasive brain-activity-monitoring tool electroencephalography (EEG), fNIRS was assessed to be less sensitive to head movement and other movement-related artifacts which are anticipated in the cockpit. Previously-published work has confirmed the difficulty of processing EEG in actual flight [46]. Additionally, fNIRS has a higher spatial resolution than EEG at approximately 1 cm² depending on the sensor geometry [59]. This resolution enables the observation of specific structures of the brain believed to be responsible for high-level cognitive processes including focused attention, namely the pre-frontal cortex [59]. Finally, its application in the cockpit environment could be more easily envisioned as it requires less expertise in placement and setup than typical EEG systems.

Eye movement and other ocular-related features were included because an eye-tracking system had previously been integrated into the laboratory's testing environment and multiple studies had shown a correlation between various features and mental workload [41], [47]. Cockpit-mounted eye-tracking systems also have the advantage of being completely non-invasive to the user. It is noted, however, that eye-related features have not been shown to be robust markers of mental workload in the cockpit scenario (see Section 1.3).

Respiration was chosen for experimentation due to the consistent finding in previously-published works that it was sensitive to mental workload [29], [58]. It was also postulated that the signal could be acquired robustly from a sitting pilot and would not be strongly influenced by other factors such as physical exertion.

ECG was included to enable the extraction of heart rate and other cardiovascular-related features which have consistently been shown to be sensitive to mental workload [24], [35], [37], [46], [55], [58], [60]. Of all psychophysiological signals extracted to assess mental workload, these features are most commonly researched in published scientific works. It was postulated these signals could be used to assess the general validity of the other extracted signals.

EDA was selected for its robustness in detecting sudden changes in mental workload [37], [58]. A person's surprise or anxiety felt in response to a visual or auditory stimulus is often clearly manifest in recorded EDA within 2-5 seconds. This transitory element of an EDA signal is known called a skin conductance response (SCR). However, because GSR is sensitive to many external factors including temperature, humidity, and time of day, slower-responding elements of the signal are likely not suitable for mental workload determination in an aircraft cockpit. Of all physiological signal signals measured, EDA has the unique ability to capture transient responses occurring within only a few seconds.

2.1 Functional Near-Infrared Spectroscopy (fNIRS)

The following introduction on fNIRS theory and applicability to the fields of operator mental workload estimation and adaptive assistant systems is taken largely from my previously-published work [61].

Functional near-infrared spectroscopy (fNIRS) is a neuroimaging technique, similar to fMRI, that measures changes in blood oxygenation in the superficial layer of the cortex (to a depth of approximately 1-2 cm) [62]–[65]. Because blood oxygenation changes are due in part

to the neuronal activity of the local brain tissue, this method can be used to non-invasively probe the activity of the brain in real-time.

Many research groups have shown the ability to discriminate between various mental states using fNIRS data. Operator attentional states among seven subjects have been distinguished with an accuracy of 65-95% (low vs. high) [49]. Levels of cognitive fatigue among four participants could be discriminated with an accuracy of 87% both in a flight simulator and in actual flight (low vs. high) [25]. Working memory across 19 pilots was classified at an 80% accuracy (low vs. high load) [50]. Game difficulty level across 9 subjects was determined with accuracies of 94% (play vs. rest) and 61% (easy vs. hard) [43]. Lastly, whether a person was performing mental arithmetic or imagining an emotional musical arrangement was classified with an average accuracy of 77% [45].

Although this is only a very limited sampling of the work being done in this field, it is clear that fNIRS-based mental state classification is being applied to a wide range of problem sets. It is apparent, however, that the majority of the work involving mental state classification is concerned with discriminating between two or three well-defined and unique states (e.g., rest vs. non-rest) in laboratory settings. Additionally, many studies report significant inter-subject variability. Additionally, despite the limited complexity of the two and even three-class classification problem, a wide range of accuracies are reported – including those below chance levels.

2.1.1 fNIRS Data Acquisition and Pre-Processing

Multiple commercial fNIRS acquisition systems exist ranging in price from \$12,000 to more than \$99,000. The following systems were evaluated for use in this work: the Brite-24 system from Artinis (\$30,000), the NirsSport2 8x8 from NIRx (\$55,000), and the fNIR203C from fNIR Devices (\$12,000). Prices reported are those quoted from the respective companies in 2019. Notably, the variations in hardware and software between these devices may likely contribute to the variability in results reported by published studies as previously noted (see Section 1.3). Inconsistent product features across these devices include the number of channels, sensor design, sensor comfort, sensor placement, emitted and collected IR frequencies, emitter-detector separation, the existence of short channels, the ability to vary emitter power and detector gains, and the ability to manipulate what is considered the “baseline” when computing change in hemoglobin.

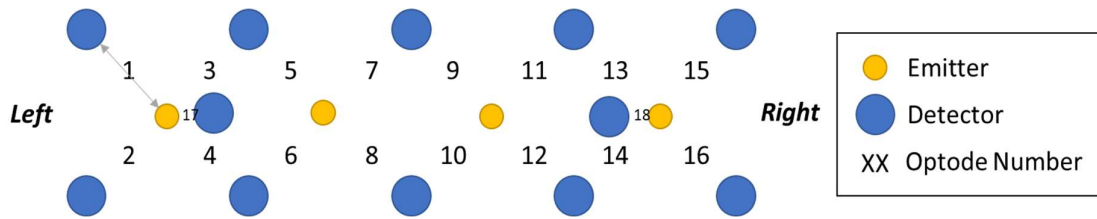


Fig. 2.1 Optode layout of the 18-Optode forehead sensor pad from fNIR Devices as if viewed from behind the wearer (channel 1 is on the wearer's left). Optodes 17 and 18 have a 10 mm emitter-detector separation of while optodes 1 through 16 have a 25 mm separation.

With price and comfort weighed heavily in the evaluation of which system to purchase, the fNIR203C headband from fNIR Devices was selected and used throughout the majority of this research. The device emits 730 nm and 850 nm wavelength light and has 16 channels with 25 mm optode spacing and two “short channels” with 10 mm optode spacing. The optodes are mounted in a headband to measure from the pre-frontal cortex (PFC). The layout of emitters and detectors for this device is depicted in Fig. 2.1.

A depiction of the fNIRS signal acquisition and processing chain is provided in Fig. 2.2. Raw light intensity for the 730 nm, 850 nm, and “ambient light” signals was sampled at 10 Hz from the 18-channel system and passed via Transmission Control Protocol (TCP) to custom software written in Python for processing. A block diagram showing the processing steps taken after raw signal acquisition is given in Fig. 2.3.

The first step in the signal processing pipeline following raw signal acquisition is removing the interference induced in the fNIRS data by the eye-tracking system used in the simulator. The interaction of these two systems resulted in significant noise on each of the three raw signals in each of the 18 channels. This interference can be seen in Fig. 2.4. Approximately every 19.5 seconds, each signal exhibits a dramatic rise and fall of intensity lasting

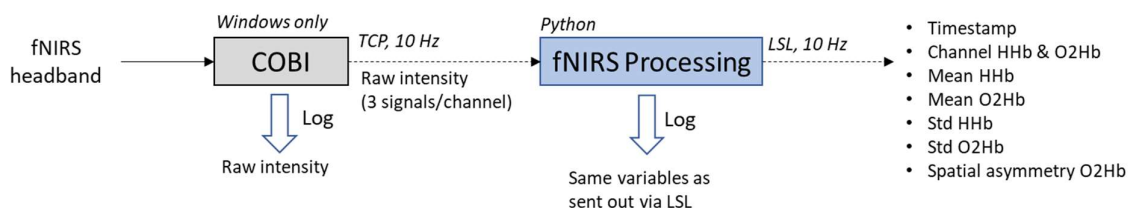


Fig. 2.2 A general depiction of the fNIRS signal acquisition and processing chain. Raw light intensity was sampled at 10 Hz by the fNIR203 system from fNIR Devices and sent from the proprietary COBI software via Transmission Control Protocol (TCP) and received and processed in custom software written in Python. There raw light intensity for the 730 nm, 850 nm channels was converted into change in oxygenated and deoxygenated hemoglobin and features of these signals were extracted and transmitted via Lab Streaming Layer (LSL) for subsequent incorporation in multi-modal mental workload prediction (see Chapter 3).

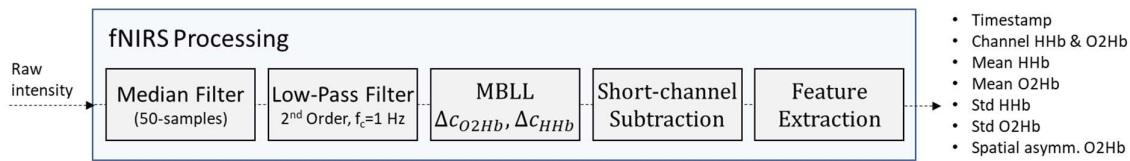


Fig. 2.3 Block diagram of fNIRS processing chain. A median filter is applied to remove noise induced by the eye-tracking system. A low-pass filter is applied to remove physiological noise (cardiac activity, respiration, etc.). The Modified Beer-Lambert Law (MBLL) is used to calculate change in oxygenated and deoxygenated hemoglobin concentration. “Short-channel subtraction” is performed to remove surface-layer contributions to concentration change calculations. Features are extracted as shown.

approximately two seconds. After simulated testing, it is believed this interference is caused by the periodic alignment of the fNIRS detection frequency (10 Hz) (and its associated collection duration) and the eye-tracking IR pulse emission frequency (approximately 60 Hz). Simulating the eye-tracking system pulsing IR at 60.05128 Hz (rather than exactly 60 Hz) reproduces the observed interference period of 19.5 seconds.

To account for this interference, a rolling 50-sample (5-second) median filter was applied to the raw light intensity data resulting in an approximate 2.5 s delay. The effect of this filter on the raw data can be seen in the bottom plot of Fig. 2.4. A filter of any shorter duration was incapable of robustly removing the interference from the eye-tracking system.

Following the application of the median filter, a 2nd order low-pass filter with a cutoff frequency of 1 Hz is applied resulting in an additional 0.2 s delay. This filter is applied to remove unwanted signals that arise in fNIRS data from physiological processes such as cardiac activity and respiration.

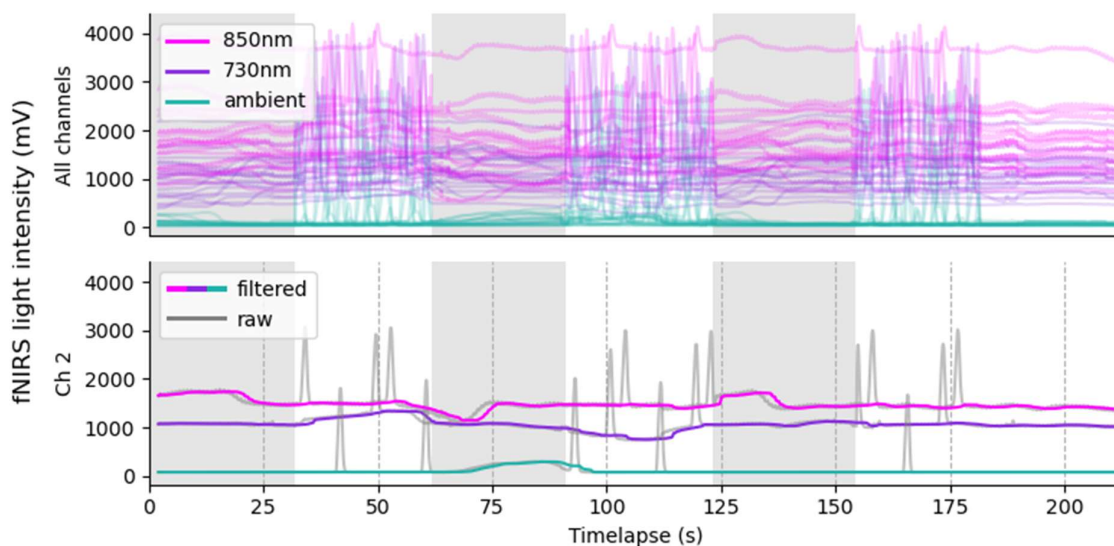


Fig. 2.4 fNIRS light intensity data being distorted by the eye tracking system mounted in the simulator. Areas in grey represent periods during which the eye-tracking system was off. It is noted that each of the three signals on each of the 18 fNIRS channels is distorted approximately every 19.5 seconds.

2.1.2 fNIRS Feature Extraction

Following these pre-processing steps, the change in oxygenated hemoglobin concentration (Δc_{O_2Hb}) and change in deoxygenated hemoglobin concentration (Δc_{HHb}) relative to the median concentration over the previous two minutes were calculated using the Modified Beer-Lambert Law (MBLL) [66]. This law relates the absorbance of a turbid medium (exhibits absorption and scattering) to the medium's absorption coefficient, the concentration of the absorbing species, and the path length of the light through the medium and is given by equation (2.1).

$$A = \log\left(\frac{I_s}{I_d}\right) \approx \varepsilon \cdot c \cdot DFP \cdot \rho + G \quad (2.1)$$

Where:

A is the absorbance or Optical Density of the medium,

I_s is the light intensity at the source,

I_d is the light intensity at the detector,

ε is the molar extinction or absorption coefficient of the medium,

c is the concentration of the absorbing molecule,

DFP is the Differential Pathlength Factor,

ρ is the source-detector separation distance,

and

G is a constant attenuation factor attributable to the scattering properties of the medium.

Assuming all variables remain constant except the concentration of the absorbing molecule, the MBLL can be manipulated to account for a change in absorbance as shown in equation (2.2).

$$\Delta A = \log\left(\frac{I_s}{I_{d_t}}\right) - \log\left(\frac{I_s}{I_{d_{t=0}}}\right) \approx \Delta c \cdot \varepsilon \cdot DFP \cdot \rho \quad (2.2)$$

Which can be reduced to

$$\Delta A = \log\left(\frac{I_{d_{t=0}}}{I_{d_t}}\right) \approx \Delta c \cdot \varepsilon \cdot DFP \cdot \rho \quad (2.3)$$

Where:

ΔA is a change in absorbance or Optical Density of the medium,

$I_{d_{t=0}}$ is the light intensity at the detector at time $t = 0$,

I_{d_t} is the light intensity at the detector at time t ,

and

Δc is a change in the concentration of the absorbing molecule.

Finally, if the medium contains more than one light-absorbing molecule, the contribution of each molecule to the overall change in absorbance is modeled by (2.4).

$$\Delta A = \log\left(\frac{I_{d_{t=0}}}{I_{d_t}}\right) \approx DFP \cdot \rho \cdot \sum_n \Delta c_n \cdot \varepsilon_n \quad (2.4)$$

Where:

n designates a particular molecule in the medium.

Because the absorbance of human tissue is largely attributable to the concentration of oxygenated and deoxygenated hemoglobin in red blood cells, for the application of fNIRS, equation (2.4) can be written as

$$\Delta A = \log\left(\frac{I_{d_{t=0}}}{I_{d_t}}\right) \approx DFP \cdot \rho \cdot (\Delta c_{O_2Hb} \cdot \varepsilon_{O_2Hb} + \Delta c_{HHb} \cdot \varepsilon_{HHb}) \quad (2.5)$$

If absorbance changes are measured at two or more wavelengths, equation (2.5) can be used for each wavelength to approximate these concentration changes using the resulting system of linear equations. As presented previously, the frequencies emitted by the fNIRS system used in this work are 730 nm and 850 nm resulting in the following system of equations:

$$\log\left(\frac{I_{d_{t=0}, \lambda_{730}}}{I_{d_t, \lambda_{730}}}\right) \approx DFP \cdot \rho \cdot (\Delta c_{O_2Hb} \cdot \varepsilon_{O_2Hb, \lambda_{730}} + \Delta c_{HHb} \cdot \varepsilon_{HHb, \lambda_{730}}) \quad (2.6)$$

$$\log\left(\frac{I_{d_{t=0}, \lambda_{850}}}{I_{d_t, \lambda_{850}}}\right) \approx DFP \cdot \rho \cdot (\Delta c_{O_2Hb} \cdot \varepsilon_{O_2Hb, \lambda_{850}} + \Delta c_{HHb} \cdot \varepsilon_{HHb, \lambda_{850}}) \quad (2.7)$$

Equations (2.6) and (2.7) can be solved for Δc_{O_2Hb} and Δc_{HHb} resulting in equations (2.8) and (2.9).

$$\Delta c_{O_2Hb} = \frac{\varepsilon_{HHb, \lambda_{730}} * \log\left(\frac{I_{d_{t=0}, \lambda_{850}}}{I_{d_t, \lambda_{850}}}\right) - \varepsilon_{HHb, \lambda_{850}} * \log\left(\frac{I_{d_{t=0}, \lambda_{730}}}{I_{d_t, \lambda_{730}}}\right)}{(\varepsilon_{HHb, \lambda_{730}} * \varepsilon_{O_2Hb, \lambda_{850}} - \varepsilon_{HHb, \lambda_{850}} * \varepsilon_{O_2Hb, \lambda_{730}}) * \rho * DPF} \quad (2.8)$$

$$\Delta c_{HHb} = \frac{\varepsilon_{O_2Hb, \lambda_{850}} * \log\left(\frac{I_{d_{t=0}, \lambda_{730}}}{I_{d_t, \lambda_{730}}}\right) - \varepsilon_{O_2Hb, \lambda_{730}} * \log\left(\frac{I_{d_{t=0}, \lambda_{850}}}{I_{d_t, \lambda_{850}}}\right)}{(\varepsilon_{HHb, \lambda_{730}} * \varepsilon_{O_2Hb, \lambda_{850}} - \varepsilon_{HHb, \lambda_{850}} * \varepsilon_{O_2Hb, \lambda_{730}}) * \rho * DPF} \quad (2.9)$$

Where:

$$\varepsilon_{HHb, \lambda_{730}} = 1.1022,$$

$$\varepsilon_{O_2Hb, \lambda_{730}} = 0.39,$$

$$\varepsilon_{HHb, \lambda_{850}} = 0.69132,$$

$$\varepsilon_{O_2Hb, \lambda_{850}} = 1.058,$$

$$\rho = 0.025,$$

and

$$DPF = 1.38.$$

It is noted that the DPF used here (1.38) was selected to yield an output consistent with the offline output from the commercial software to aid in the validation of the real-time implementation. Although lower than typically used by others who report using values between three and six [67], varying this constant would only affect the magnitude of hemodynamic changes which are not relevant to subsequent processing as all data is normalized prior to use in mental workload prediction processes.

The decision to calculate the change in concentration with respect to the median of each channel over the previous two minutes is significant and was made deliberately. Calculating the change in hemoglobin concentration with a different reference will yield vastly different results. Although it could be argued that calculating the change in hemoglobin from some rest state observed at the beginning of a session might be reasonable, doing so introduces significant error due to the drift in the signal over time. This drift may be due to slowly changing light conditions, movement of the emitters or detectors, or other factors. By calculating the difference over the last two minutes, it is anticipated that observed hemodynamic responses would reflect neuronal activity changes in the brain occurring within this time-scale which are precisely those events of particular interest in this research.

Table 2.1 Features extracted from fNIRS data.

Feature	Description
μ_{HHb}	Mean change in deoxygenated hemoglobin across all 16 fNIRS channels
μ_{O2Hb}	Mean change in oxygenated hemoglobin across all 16 fNIRS channels
σ_{HHb}	Instantaneous standard deviation across all 16 deoxygenated hemoglobin signals
σ_{O2Hb}	Instantaneous standard deviation across all 16 oxygenated hemoglobin signals
<i>Spatial Asymmetry O2Hb</i>	Difference between right and left hemisphere mean change in oxygenated hemoglobin

As depicted in Fig. 2.3, following the application of the MBLL to calculate hemoglobin changes, short-channel subtraction was then performed. This was done by subtracting the right and left short-channel signals from the other signals on their respective hemisphere. In other words, the signal from the short-channel located on the right hemisphere was subtracted from each of the other signals collected from the right hemisphere and vice versa for the signals collected from the left hemisphere.

Finally, features were extracted from the processed fNIRS signals as noted in Table 2.1 and as defined in equations (2.10) through (2.14). Each feature was calculated as an instantaneous measurement without the application of a sliding window due to the nature of the processed fNIRS signals already representing a window of time. In addition to the mean and standard deviation features, one spatial asymmetry measure was extracted as noted by “*Spatial Asymmetry O2Hb*”. This feature provides the difference between the mean O2Hb of the right and left hemispheres which has been suggested by other work to correlate with approach-related motivational tendencies [68].

$$\mu_{O2Hb} = \frac{1}{N} \sum_{i=1}^N O2Hb_i \quad (2.10)$$

$$\mu_{HHb} = \frac{1}{N} \sum_{i=1}^N HHb_i \quad (2.11)$$

$$\sigma_{O2Hb} = \sqrt{\frac{1}{N} \sum_{i=1}^N O2Hb_i - \mu_{O2Hb}} \quad (2.12)$$

$$\sigma_{HHb} = \sqrt{\frac{1}{N} \sum_{i=1}^N HHb_i - \mu_{HHb}} \quad (2.13)$$

Spatial Asymmetry O2Hb

$$= \frac{1}{N_{right}} \sum_{i=1}^N O2Hb_{i|O2Hb_i=right} - \frac{1}{N_{left}} \sum_{i=1}^N O2Hb_{i|O2Hb_i=left} \quad (2.14)$$

The filtered change in hemoglobin signals and these extracted features were subsequently broadcasted on the network via the open-source network streaming middleware Lab Streaming Layer¹ (LSL) for subsequent retrieval and use (see Chapter 3).

2.1.3 fNIRS Sensitivity to Head Position

During preliminary testing, it was found that significant changes in hemoglobin concentration were attributable to changes in the participant's head position, specifically its pitch. Fig. 2.5 shows the dramatic effect a large downward tilt of the head (negative pitch angle) can have on the recorded change in hemoglobin concentrations. To enable the potential correction of this influence on fNIRS signals, an inertial measurement unit (IMU) was used in conjunction with an Arduino-based microcontroller to record head roll, pitch, and yaw. A depiction of the processing chain and a photograph of the custom-built sensor is provided in Fig. 2.6. The sensor, housed in a 3D printed case, was built using an MPU-6050 IMU (~5€) and was controlled using a Teensy 4.0 microcontroller (~23€) running Arduino firmware. The well-documented “gimble lock” issue was overcome by mounting the IMU such that the pitch axis would remain between -90 and 90 degrees of rotation during standard wear.

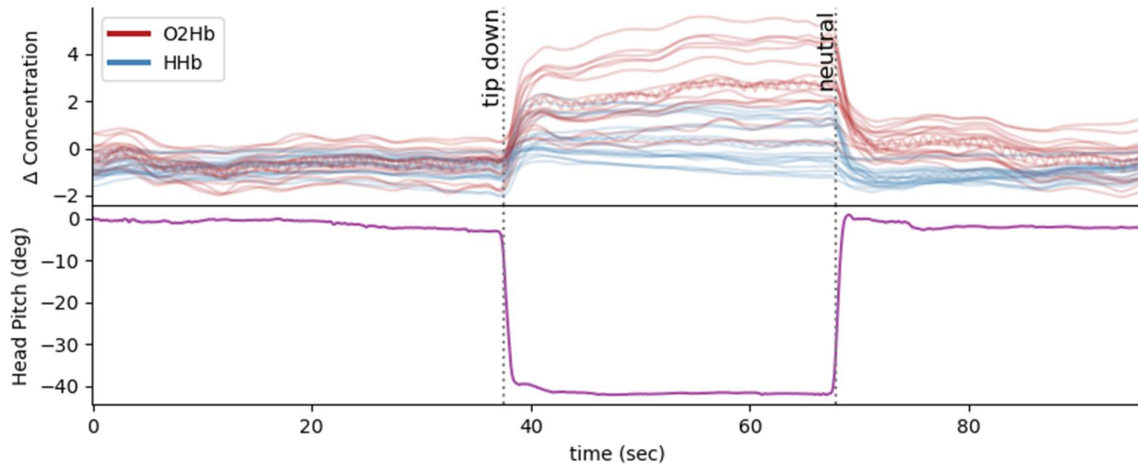


Fig. 2.5 A representative plot showing the effect of head tilt on fNIRS data. Pitching the head forward causes a large increase in the measured change in hemoglobin concentrations.

¹ Lab Streaming Layer (LSL) is an overlay network for real-time exchange of time series between applications. The Python interface is “pyls” and its source can be found here: <https://github.com/labstreaminglayer/liblsl-Python>

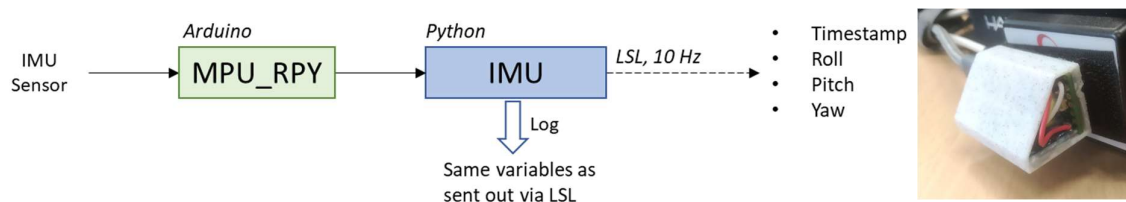


Fig. 2.6 A depiction of the IMU data acquisition and processing chain. Raw data was collected from the IMU sensor using an Arduino-based microcontroller (right). Roll, pitch, and yaw values were filtered in a custom python script and transmitted over the network using at 10 Hz for retrieval and use in subsequent processes. To the right is a photograph of the custom-built IMU sensor attached to the fNIRS headband.

Due to the late identification of this relationship in the experimental preparations, these IMU measurements were not incorporated into the real-time mental workload prediction algorithm presented in subsequent sections of this work. As presented in greater detail in Section 4.3 however, an offline analysis following simulated flight scenarios showed only a weak average correlation between head pitch and change in oxygenated hemoglobin (-0.08 ± 0.16). The lack of a strong negative correlation between head tilt and change in oxygenated hemoglobin in the pre-frontal cortex shown through this analysis suggests the influence of head tilt on recorded fNIRS signals is not a significant confounding factor in the designed experiment. This is potentially due to the limited range of head tilt required during the execution of the flight scenarios. It was found that the average standard deviation of head tilt across all participants was only 5.77 degrees. Thus approximately 95% of the head tilt angles fell within approximately 20 degrees. The head tilt movements made during the experiment are thus much less extreme than that depicted in Fig. 2.5. A stronger negative correlation may have been observed if the task required more substantial vertical movements of the head. The small degree of head tilt in this scenario may explain the minimal impact this movement had on the recorded fNIRS signal.

In addition to this analysis suggesting head tilt does not significantly impact recorded fNIRS signals in the cockpit setting, the only identifiable published work on this topic similarly concluded that head tilt is likely not detrimental to the utility of uncorrected-fNIRS signals in this setting [42]. In this work, it was determined that the extracted fNIRS signals returned to baseline levels within three seconds of the operator's head returning to its pre-tilt position. Thus, it is argued that by calculating the change in hemoglobin over a longer duration of time, signal variation due to transient head tilt can be mitigated.

Despite these findings, future work could explore a more robust correction for head tilt in the processing of fNIRS signals.

2.2 Eye Tracking

“The eyes are a window to the soul.” This old proverb suggesting that the eyes reveal clues to a hidden state within a being has been rigorously tested in many scientific settings. One early theory regarding the connection between the eyes and mind is the Eye-Mind Hypothesis (EMH) which posits “there is no appreciable lag between what is fixated and what is processed” [69]. Although a strong relationship has been shown to exist between fixation and cognitive processing, the theory has been challenged in various ways and is generally not accepted without multiple caveats. For example, it is well studied and understood that attention can be directed towards things or ideas not fixated upon (known as “covert attention”) [70]. Although a perfect association between the eyes and mind cannot be made, strong correlations certainly exist and can be leveraged in this application.

As summarized in Section 1.3, many have experimented with extracting various features of eye-related movement or activity to deduce the mental workload of a subject. Blink frequency has been shown to generally decrease with higher visual demand ([35], [37], [58], [60]), pupil size has been shown to increase with an increased mental workload ([41], [47]), saccade peak velocity has been shown to decrease with increased workload [44], and fewer and shorter fixations have been observed with increased workload [41]. Due to these and other promising results, eye-tracking features were extracted in this work to aid in real-time mental workload prediction.

In conjunction with previous experimentation within the laboratory, a Smart Eye Pro eye-tracking system was integrated into the helicopter simulator. Various elements of this



Fig. 2.7 Various elements of the eye-tracking system. Sub-figure (a) shows the IR transmitter (left) and camera (right) used in the Smart Eye Pro eye tracking system. Sub-figure (b) shows the real-time pupil and iris detection used in the SmartEye Pro system. Sub-figure (c) is a photograph of the helicopter simulator with the three eye-tracking cameras on the right cockpit circled in red.

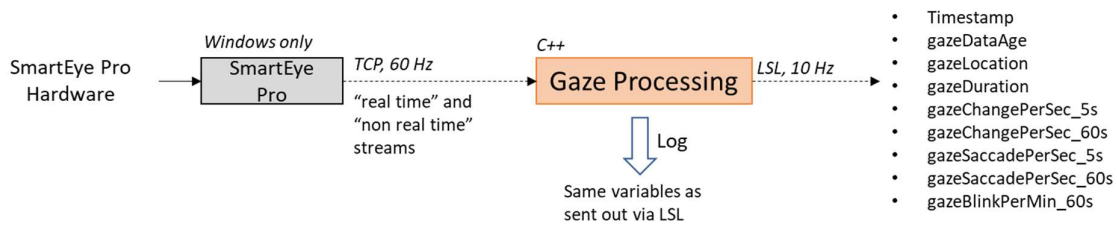


Fig. 2.8 A depiction of the eye-tracking feature extraction processing chain.

system are shown in Fig. 2.7. Sub-figure (a) shows the IR transmitter (left) and camera (right) used in the Smart Eye Pro eye tracking system. Sub-figure (b) shows the real-time pupil and iris detection used in the SmartEye Pro system. Finally, sub-figure (c) is a photograph of the helicopter simulator with the three eye-tracking cameras on the right cockpit circled in red. A depiction of the eye-tracking signal acquisition and processing chain implemented in this work is provided in Fig. 2.8.

Data from the SmartEye Pro eye-tracking system is received into custom data processing software written in C++ denoted in Fig. 2.8 as “Gaze Processing.” This program receives the “real-time” and “non-real-time” output of the proprietary SmartEye software at 60 Hz and extracts the desired features listed in Table 2.2. The full list of variables broadcast from SmartEye is provided in Appendix A. The first signal listed in Table 2.2, “gazeDataAge,” represents the time in seconds since the arrival of the last valid sample from the SmartEye system. At times, the system loses track of the user’s eyes and hence cannot update its gaze prediction. Rather than withholding transmission of data as might be expected, the proprietary software continues transmission of the last-valid sample until tracking is re-established. Hence, to inform downstream processing of the loss of tracking, the duration since the arrival of the last valid sample is extracted. Transmitting this signal as opposed to simply withholding transmission until data is available enables the differentiation between loss of eye tracking and a more significant failure in the proprietary eye-tracking system. As presented hereafter in Section 3.2, warning messages and other pilot-assisting procedures can be effectively triggered by gaze data. Having an accurate representation of the data’s currency is essential to such triggers.

Table 2.2 Features extracted from the eye-tracking system.

Feature	Description
gazeDataAge	Time in seconds since arrival of last valid eye-tracking sample from SmartEye
gazeLocation	Binary output reporting whether the pilot is looking within or out of the cockpit
gazeDuration	Duration in seconds that the pilot maintains their gaze either within or out of the cockpit
gazeBlinkPerMin_60s	Number of eye blinks in the last minute
gazeChangePerSec_5s	Frequency of fixation change from above to within cockpit in the last 5 seconds
gazeChangePerSec_60s	Frequency of fixation change from above to within cockpit in the last 60 seconds
gazeSaccadePerSec_5s	Frequency of saccades in the last 5 seconds
gazeSaccadePerSec_60s	Frequency of saccades in the last 60 seconds

The remaining features are calculated by summing the number of respective events detected by the SmartEye software over a given window of time. Blinks and saccadic movements are not extracted from the raw video data as part of this work. Rather, these events are utilized as delivered by the proprietary SmartEye software. Multiple features are extracted using both a 5-second and 60-second window in an attempt to capture unique information at these different time scales. For example, the 60-second saccade rate is extracted by summing the number of saccadic movements over the previous 60 seconds while the 5-second saccade rate is extracted by summing the movements over the previous 5 seconds.

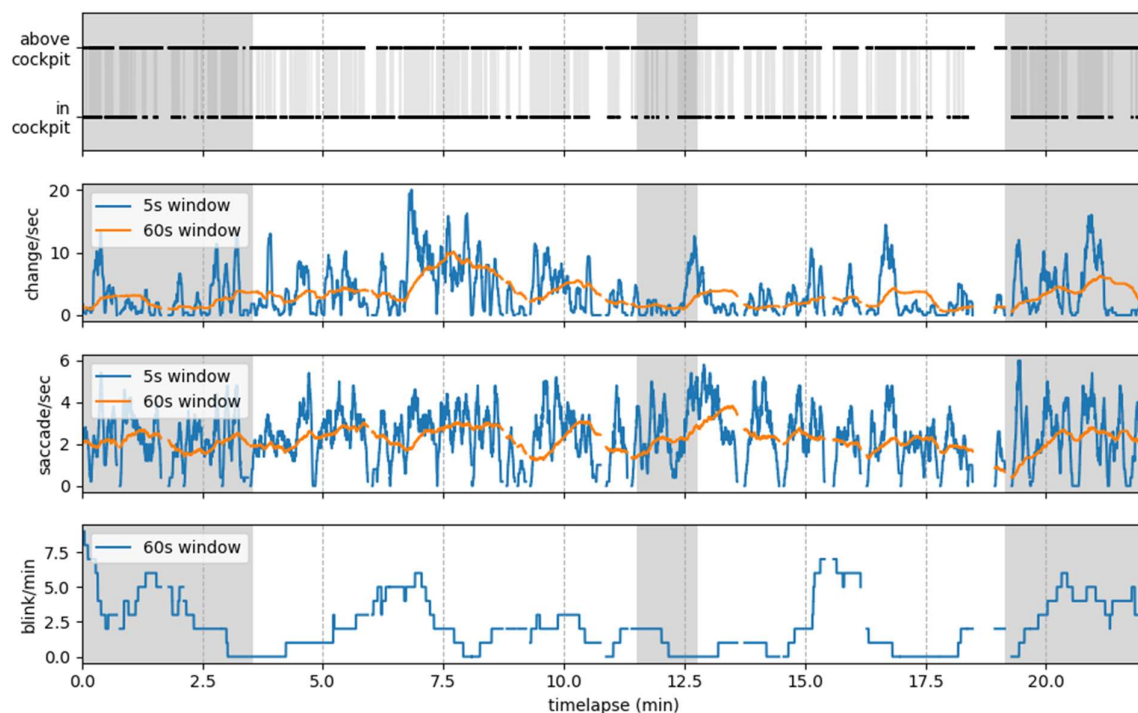


Fig. 2.9 Representative plot of various eye-related features over the course of approximately 22 minutes in a helicopter simulator. Regions in grey represent periods during which the helicopter was on the ground.

The frequency at which the pilot switches their fixation between objects internal to the cockpit and those external to the cockpit is also extracted. This was performed utilizing the returned fixation “object” from the SmartEye system having been previously calibrated to the geometry of the simulator. This feature is extracted with the intent of capturing the cognitive load associated with task-switching.

Table 2.2 provides the complete list of the extracted features and their descriptions. These features are subsequently broadcast via LSL at 10 Hz for subsequent incorporation in multi-modal mental workload prediction (see Chapter 3). Fig. 2.9 provides a representative plot of these eye-related features over approximately 22 minutes of flight in a helicopter simulator.

Respiration

A custom respiration sensor was built using a conductive rubber cord, a voltage divider circuit, and a microcontroller. Fig. 2.10 subfigures (a) and (b) show the initial and final hardware implementation respectively. In the final implementation, stretch sensors mounted in straps around the stomach and chest are integrated such that a person’s breathing is captured whether the expansion of the abdomen is localized to the stomach or the chest. The combination of signals from the stomach and chest sensors can be seen in subfigure (c).

The signal processing chain of the respiration data is depicted in Fig. 2.11. Raw stretch intensity from both the chest and stomach sensors is captured via an Arduino-based microcontroller at approximately 500 Hz. Light Emitting Diodes (LEDs) mounted in the microcontroller housing are lit to assist in the proper placement of the sensors. If the sensor is stretched too tightly such that the signal saturates and breath information cannot be captured, the LED is lit red. When the signal resides in a range effective for gathering breath fluctuations, the LED is lit green and increases in brightness until saturation.

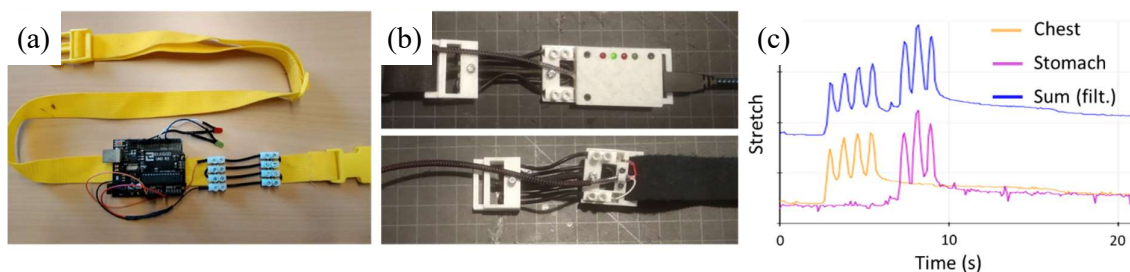


Fig. 2.10 Custom-built double-strap respiration sensor. Subfigure (a) shows the first version of the sensor using a standard Arduino-based microcontroller for easy prototyping. Subfigure (b) shows a the latest version using a microcontroller housed in a 3D-printed housing with two stretch sensors integrated into the sensor. Subfigure (c) shows data collected by the sensor in which data from the chest and stomach sensors are combined and filtered in to a single stretch value.

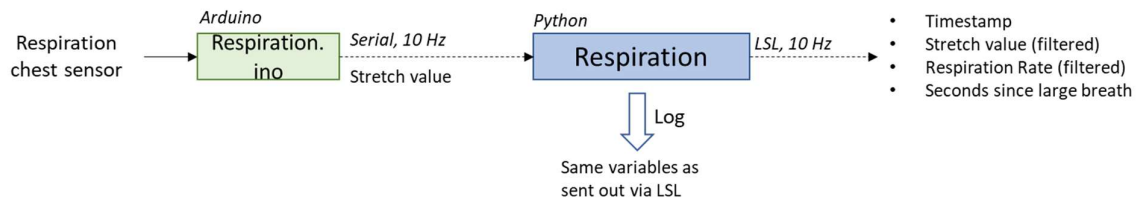


Fig. 2.11 Respiration processing chain diagram. Stretch data from chest and stomach stretch sensors are used to determine respiration rate and to detect abnormally large breaths. The smoothed stretch value, respiration rate, and the time in seconds since the last large breath are transmitted via LSL for incorporation in down-stream processes.

After the collection of raw stretch data from each sensor, the data streams are added to each other and smoothed by averaging over the last 100 samples (0.2 seconds). This single filtered respiration signal is then transmitted via a serial connection to a host PC at 10 Hz. Processing continued via custom-developed software written in Python on the host computer. There, a high pass filter is applied to remove the DC offset after which respiration rate was calculated and unusually large breaths were detected. These features are summarized in Table 2.3. Respiration rate was calculated using an adaptive IIR notch filter (ANF) as presented in [71] and [72]. Additionally, a 3rd-order low-pass Butterworth filter with a cutoff frequency of 0.2 Hz was applied to smooth the output. Respiration rate was calculated using the ANF method as opposed to more traditional frequency estimation techniques due to the findings published in [73] that the ANF method “could not only estimate [respiration rate] more quickly and more accurately than the conventional methods, but also is most suitable for online RR monitoring systems, as it does not use any overlapping moving windows that require increased computational costs.” One such conventional method for computing the frequency of a signal is through the periodogram technique (as implemented in [74] for example) which utilizes the fast Fourier transform (FFT) to implement the discrete Fourier transform (DFT). As seen in Fig. 2.12, the periodogram approach is limited in its resolution by the data window size and the DFT length which limitation can prevent the detection of small changes in respiration rate.

Table 2.3 Features extracted from the respiration system.

Feature	Description
RespRate	Breaths per minute
SecSinceAbnormal	Time in seconds since the last abnormally large breath

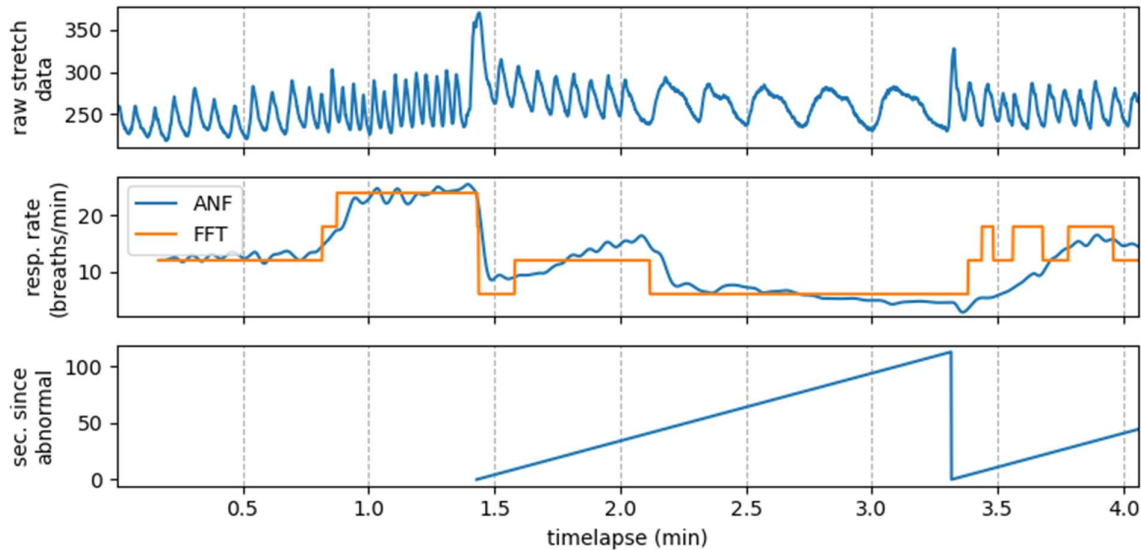


Fig. 2.12 Time series plot of processed respiration data. The upper-most plot shows raw stretch data as collected by the dual-strap respiration sensor. The middle plot shows respiration rate calculated using both the Fast Fourier Transform (FFT) method and the adaptive IIR notch filter (ANF) method. The bottom-most plot shows the time since an abnormally large breath. Before the first large breath at approximately 1.5 minutes, this value is undefined and hence not plotted.

Unusually large breaths were detected using a custom algorithm of the following structure:

1. Compute the difference in stretch over the last 5 seconds (max-min),
2. Compute the median of this difference over the last 60 seconds,
3. Flag as “large breath” if the current difference is greater than 2 times this median difference,
4. Calculate the time since the last “large breath” was detected.

These processed signals and features (smoothed stretch value, respiration rate, and seconds since the last large breath) are transmitted via LSL for downstream processing. Fig. 2.12 provides a representation of these data. Also included in the plot is respiration rate as calculated using the commonly-implemented FFT method.

2.4 Electrocardiography (ECG)

Recording of the electrocardiogram (ECG) enables the extraction of some of the most important psychophysiological features related to stress and mental workload. Specifically, one’s heart rate and heart-rate-variability have been shown by many independent researchers in a variety of contexts to be sensitive to mental workload (see section 1.3 and specifically Table 1.4 for references).

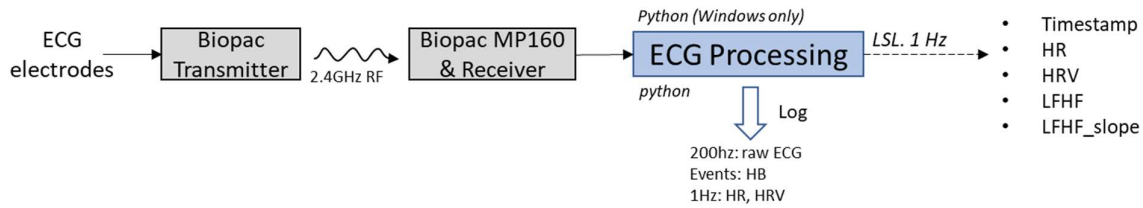


Fig. 2.13 A depiction of the ECG feature extraction processing chain. Raw ECG is collected from the participant's chest via a BIOPAC physiological data acquisition system which transmits the signal wirelessly to a receiver. The raw data is then read and processed by a custom-written ECG processing and feature extraction program. Features relevant to mental workload are extracted and broadcast via LSL at 1 Hz for subsequent use.

2.4.1 ECG Data Acquisition

A depiction of the ECG signal acquisition and processing chain implemented in this work is provided in Fig. 2.13. A single-channel ECG signal is collected at 200 Hz using the BIOPAC MP160 data acquisition system. ECG data are collected using single-use disposable Ag/AgCl pre-gelled electrodes arranged on the participant's chest arranged to sample the Lead II vector as depicted in Fig. 2.14. Lead II is sampled due to the large R-wave amplitude observed on this lead during testing which facilitates a robust peak detection. The participant's skin is cleaned using alcohol wipes and slightly abraded using a roughened sponge-like material prior to electrode placement.

Inspection of the acquired raw ECG data acquired by the BIOPAC system showed a data acquisition issue that had to be overcome to facilitate accurate peak detection and

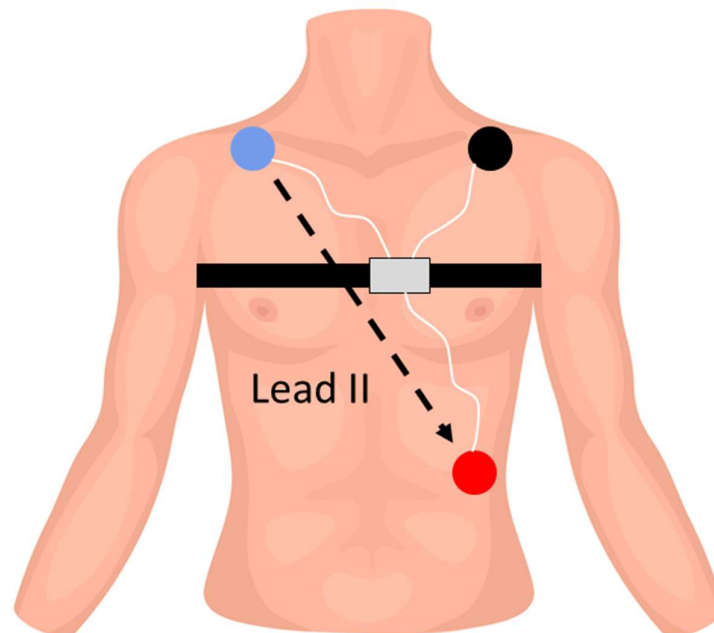


Fig. 2.14 Electrode placement yielding a Lead II ECG recording. The gray block is the wireless transmitter to which each of the three electrodes is connected. The transmitter is mounted to a flexible strap worn around the chest (which also houses one of the two respiration sensors).

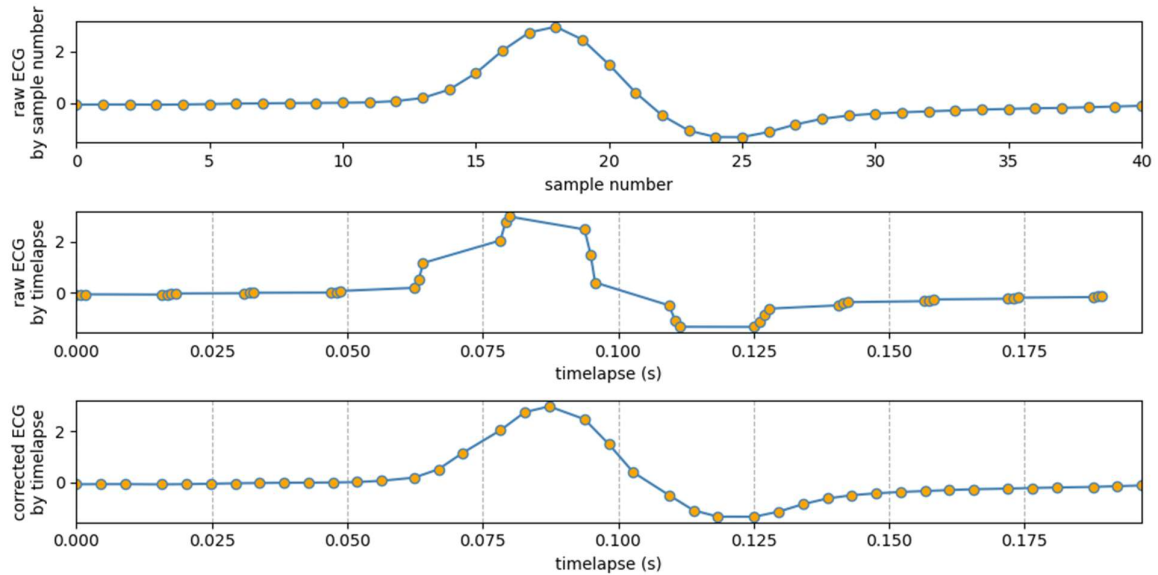


Fig. 2.15 Representative plots of the ECG data sampling issue and its correction. The upper plot shows the even sampling of the ECG signal as plotted by sample number. The middle plot shows the inconsistent retrieval of data samples from the BIOPAC system. The bottom plot shows the corrected waveform after inducing a mandatory delay between samples.

subsequent feature extraction. It was found that the sampling from the ECG electrodes was consistent (200 Hz), but the rate at which samples were made accessible for retrieval by the host computer was not. This can be seen in Fig. 2.15 where it is noted that the R-wave is smooth when plotted by sample number, but distorted when plotted by timestamp of data receipt. It is unclear whether this issue is unique to the particular BIOPAC hardware used in the lab or was a more general software or networking issue. Only through the use of a function in BIOPAC’s proprietary “dynamic link library” (DLL) can the sampled data be read into memory. To account for this issue, raw ECG data is read into memory using the DLL-provided function as fast as samples were made available and placed into a first-in-first-out (FIFO) buffer or “queue,” while at the same time, a second multi-processing thread reads from this queue and if the time between samples is less than 80% of the expected period according to the set sampling rate, a delay was imposed increasing the period to 80% of the expected period. Thus, a minimum of $0.80 * \frac{1 \text{ second}}{200 \text{ samples}} = 0.004$ seconds was imposed between samples. These data were then sent forward for subsequent processing. The result of this data acquisition strategy is to smooth the ECG waveform naturally without imposing a time-delay thus allowing for accurate identification of R-wave peaks used for subsequent feature extraction processes.

2.4.2 ECG Peak Detection

In this work, ECG feature extraction including heart rate (HR), heart rate variability (HRV), and other HRV-related features is reliant on an accurate and robust detection of R-

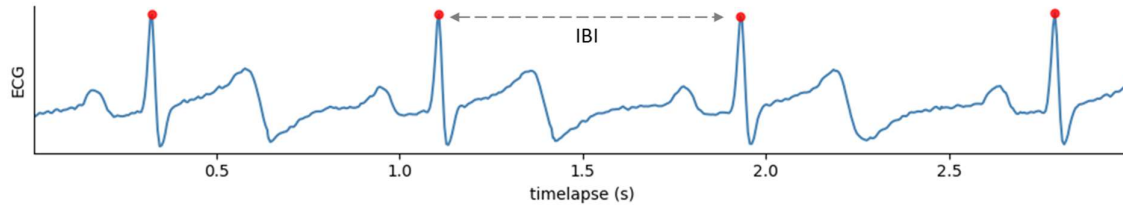


Fig. 2.16 Raw ECG recording with R-wave peaks highlighted. An inter-beat-interval (IBI) is the time between consecutive R-wave peaks.

wave peaks. With a robust detection of peaks, the intervals between consecutive peaks, or inter-beat-intervals (IBI), can be derived and the aforementioned features can be determined. Fig. 2.16 shows the relationship between R-wave peaks in a raw ECG signal and an IBI. In pursuit of this goal, a custom peak detector was developed using a series of conditions on the ECG signal, the amplitude of the detected peak compared to previously-detected peaks, and the duration of the IBI compared to preceding IBIs. This algorithm was tested and developed on a large number of datasets collected from over 20 individuals and proved highly effective at detecting valid R-wave peaks while rejecting noise. Other open-source algorithms were found to be less robust to noise and/or took more time to re-acquire peak detection after loss. This novel peak detection algorithm and the determination of “valid” peaks is performed using the following procedure (and can be visualized in Fig. 2.17):

1. A local maximum or “peak” is identified. A peak is identified when an ECG local maximum is found to exist above a 3-second rolling 98th percentile threshold updated every second. Due to the high variability of ECG data and its susceptibility to noise, the small rolling window served to ensure the threshold remained current and useful.
2. The amplitude of the peak is compared with the amplitude of those detected in the last 5 seconds. If it falls below 75% or above 125% the mean amplitude of the peaks in this window, the detected peak is labeled as having an amplitude that is “too different” from the others and is not considered further as a potential “valid” peak. If “valid” peaks have been identified in the last 5 seconds, this comparison is made with the mean amplitude of these peaks. If no “valid” peaks have been identified in the last 5 seconds, the comparison is made with all observed peaks during this time.
3. The IBI is calculated as the time in milliseconds since the last peak.
4. A determination is made whether or not the IBI falls between 400 ms and 1500 ms which corresponds to a maximum and minimum heart rate of 150 and 40 beats per minute (bpm) respectively. If the IBI is outside this range, the peak is labeled “IBI out of range.” If within the range, the peak is considered further as a potential “valid” peak.

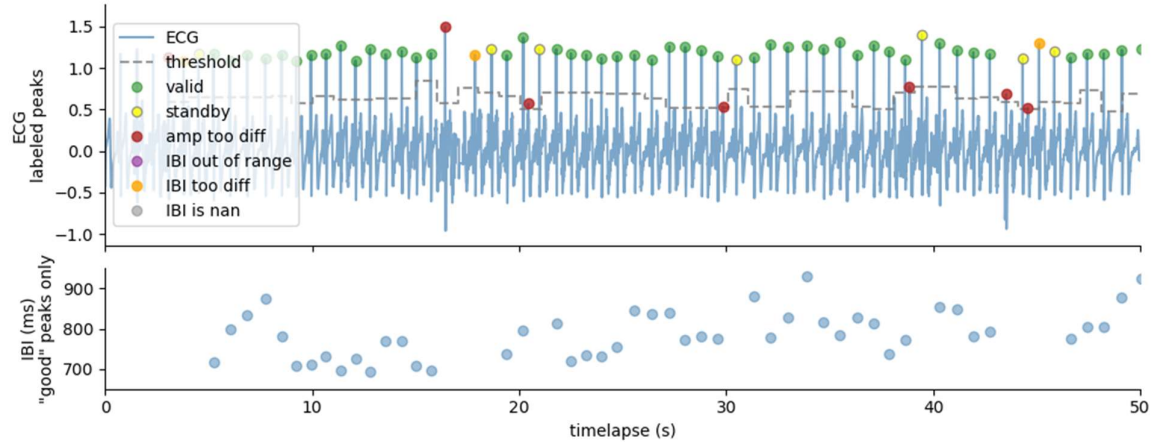


Fig. 2.17 Visualization of the “valid” ECG peak detection algorithm. The upper plot shows the ECG signal, the threshold used to identify local maxima, and the automated labeling of peaks. The lower plot gives the IBI of “valid” peaks which signal is used in subsequent feature extraction methods. The dashed grey line shows the rolling threshold above which local maximum are identified which begins the peak detection process. As visualized in this dataset, increased noise in the ECG signal lowers the threshold thus capturing local maxima not corresponding to R-Wave peaks making filtering essential for subsequent feature extraction relying on an accurate IBI signal. Seen at approximately 16 seconds, a peak is labeled as having an amplitude that is “too different” than the preceding peaks and is rejected as a “valid” peak. Next, whether from a sampling error or a skipped heartbeat, the peak detected next is found to have a very large IBI (approximately two times the previous IBIs) and hence is labeled as having an IBI that is “too different” from the preceding (valid) IBI’s. Following this peak, the next meets all the criteria except that its predecessor was not labeled “valid” or “standby” so it is labeled “standby.” The following meets all the criteria and its predecessor was “standby” so it is labeled “valid” and an IBI for this peak is carried forward into subsequent feature extraction processes.

5. The IBI is compared with IBIs of the last 5 seconds. If it falls below 75% or above 160% of the mean IBI of those in the last 5 seconds, the peak is labeled as having an IBI that is “too different” from the others. If within the range, the peak is considered further as a potential “valid” peak. As with the amplitude comparison, this comparison is made with “valid” IBIs if they exist; otherwise, it is made with all IBIs regardless of label.
6. Finally, if the current peak meets the above-stated criteria but the previously-detected peak did not, the current peak is labeled “standby.” If the current peak meets the above-mentioned criteria and the previously-detected peak was labeled “standby” or “valid,” then the current peak is labeled “valid.” This check ensured some stability in the system before declaring a “valid” peak.

2.4.3 ECG Feature Extraction

As is true with all feature engineering endeavors, great care must be taken to generate features of the most value given the signal from which a feature is being extracted and the environment in which the system is to be deployed. In the case of ECG feature extraction, one must balance the stability of the feature with the desired time resolution. For example, when extracting heart rate for use in a pilot monitoring application, one could obtain a highly stable

Table 2.4 Features extracted from the ECG system.

Feature	Description
HR	Heart rate calculated as the mean inter-beat-interval (IBI) over the last 20 seconds
RMSSD	Root mean square of successive RR interval differences over the last 30 seconds (a HRV feature)
LF/HF	Low-frequency to high-frequency power ratio of the IBI over last 120 seconds (a HRV feature)
LF/HF_slope	Change in LF/HF over one second (a HRV feature)

heart rate using a sliding window of 10 minutes. This long window, however, would prevent one from observing changes to a pilot's heart rate at shorter time-scales such as might happen when an unexpected emergency light is seen but is understood and resolved after only a few seconds. On the other hand, using a very short window for heart rate estimation results in a highly variable heart rate signal susceptible to data acquisition noise or other non-psychophysiological factors. Thus, for each of the ECG-extracted features, windows of varying duration were examined and a subjective determination was made as to the appropriate window size. As implemented in other sections of this work (EDA feature extraction and eye tracking feature extraction), features can be extracted with multiple window lengths in the hopes of capturing unique information from each, but this was not done in the case of the ECG-extracted features.

Table 2.4 provides a summary of the features extracted from the ECG in this work. We can take advantage of the fact that “a healthy heart is not a metronome” [75]. At any given time, one's HR is the result of the interplay between the neural activity of the sympathetic (SNS) and parasympathetic nervous systems (PNS) where sympathetic neural activity increases HR and parasympathetic neural activity decreases it. First and foremost, heart-rate is extracted as the mean IBI (of “valid” peaks) over the last 20 seconds. Windows of varying duration were examined and it was subjectively determined that a window of 20 seconds was responsive to stimuli within approximately 10 seconds which aligns with the timescale of mental workload variations expected in the environment. Additionally, it was found that a window of this duration yielded a heart rate signal which was not overly sensitive to IBI variability induced by noise or other non-psychophysiological factors.

In addition to heart rate, heart rate variability (HRV) features in both the time and frequency domains are also extracted as they have likewise been shown to be sensitive to mental workload (see Section 1.3). In the time domain, the HRV feature extracted is the Root mean square of successive heart beat interval differences (RMSSD) over a sliding window of 30 seconds. The RMSSD reflects the beat-to-beat variance in HR and is the primary time-domain measure used to estimate the vagally mediated changes (parasympathetic nervous

system) reflected in HRV [76]. Although, HRV features are typically extracted at timescales of tens of minutes or even hours, previously published works have proposed short-term periods of 60, 30, and even 10 seconds [76].

The RMSSD is calculated using equation (2.15) and is measured in milliseconds.

$$RMSSD = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} (IBI_{i+1} - IBI_i)^2} \quad (2.15)$$

Where:

IBI is the inter-beat-interval between successive R-wave peaks

and

n is the number of samples in the window of question.

In addition to the time-domain feature RMSSD, a feature in the frequency domain is also extracted namely the ratio of low-frequency to high-frequency power (LF/HF). As explained in [76], “the assumptions underlying the LF/HF ratio is that LF power may be generated by the SNS while HF power is produced by the PNS.” Thus, a low LF/HF ratio may reflect parasympathetic dominance which occurs when relaxed and engaging in “tend-and-befriend behaviors.” On the other hand, a high LF/HF ratio may indicate sympathetic dominance, which occurs when stressed and engaging in “fight-or-flight” behaviors. It is cautioned however against putting too much weight behind this rationale, especially in short-

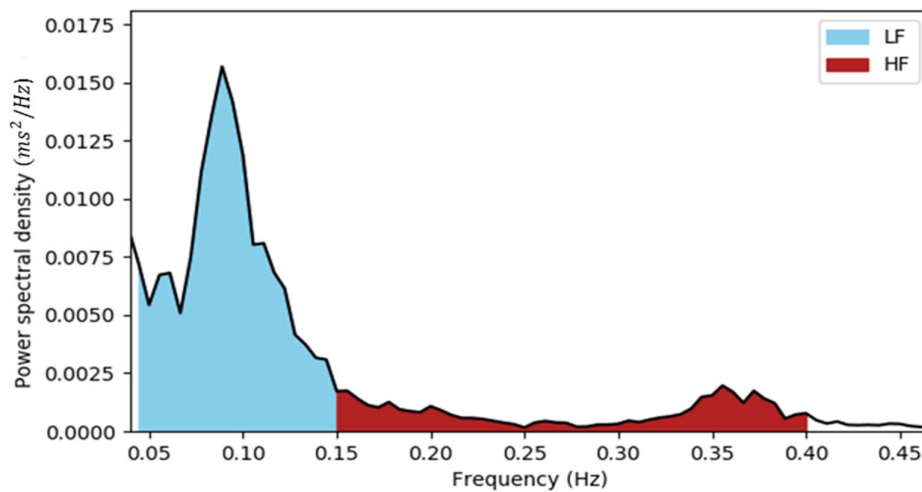


Fig. 2.18 Power Spectral Density of the IBI signal. Absolute power for the low (0.04 and 0.15 Hz) and high frequency (0.15 and 0.4 Hz) bands are calculated by approximating the area under the curve of the PSD within these respective bands. Finally, LF/HF is calculated as the ratio between these two values.

term (<5 min) recordings due to the complexity of the system not fully accounted for by this simplistic model.

In this work, LF/HF is calculated every second by estimating the power spectral density of IBI over a window of the previous two minutes using Welch's method [77]. As with the extraction of HR, this window size was selected after a subjective evaluation of the signal over a wide range of window sizes. Absolute power in the low frequency (0.04 and 0.15 Hz) and high frequency (0.15 and 0.4 Hz) ranges were calculated by approximating the area under the curve of the power spectral density of these ranges. These ranges can be seen in Fig. 2.18. The LF/HF signal is then smoothed using a 2nd order low pass Butterworth filter with a cutoff frequency of 0.5 Hz. Additionally, the instantaneous rate of change, or slope, of LF/HF was extracted due to its potential to signal LF/HF.

Fig. 2.19 shows all ECG-extracted features over the course of a simulated helicopter flight.

The aforementioned ECG processing chain is made accessible to an experimenter through a graphical user interface (GUI) depicted in Appendix B. The GUI provides the following capabilities to the experimenter:

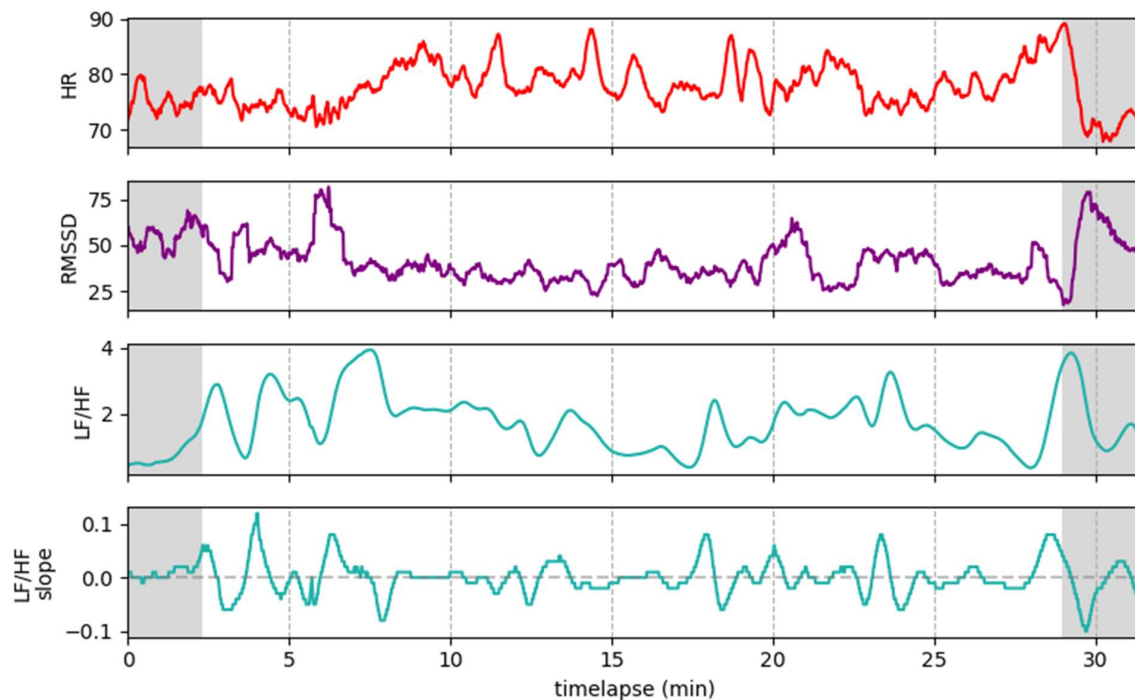


Fig. 2.19 Representative plot of ECG-extracted features over the course of approximately 31 minutes in a helicopter simulator. Regions in grey represent periods during which the helicopter was on the ground. Periods of high mental workload can be inferred from an analysis of the ECG-extracted features. The few minutes leading up to landing for example, show an increased HR, a decreased RMSSD, an increasing LF/HF, and a positive LF/HF slope, all of which have been shown to correlate with an increased mental workload.

- The collection and processing of ECG (and EDA) data from two participants simultaneously.
- The playback of previously-recorded ECG and EDA data as if being received in real-time.
- The viewing of raw ECG (and EDA) and extracted features in plotting windows.
- The logging of raw ECG (and EDA) and extracted features.
- The transmission of ECG (and EDA) features via Lab Streaming Layer.

2.5 Electrodermal Activity (EDA)

Electrodermal Activity (EDA), or the recording of the galvanic skin response is a measure of skin conductance. When the pores on the surface of the skin expand and sweat is excreted, the skin's conductance increases. As presented in Section 1.3, EDA has been shown to be sensitive to changes in mental workload. Specifically, it has been shown that an increased mental workload leads to an increased frequency of EDA events where an EDA event is the rapid rising and falling of EDA over a few seconds [37]. Additionally, and more significantly relevant to this work, EDA has been shown to be particularly sensitive to sudden changes in mental workload [58]. Unlike some of the other signals and features extracted, EDA provides sensitivity to events occurring at small time-scales. For example, a startling event, such as the realization that one was within seconds of collision with the terrain but which could be quickly resolved may not result in a change of heart rate, yet it almost certainly would result in a dramatic EDA event. This situation has been observed on countless occasions in the simulator and a representative time series of this event is provided in Fig. 2.20.

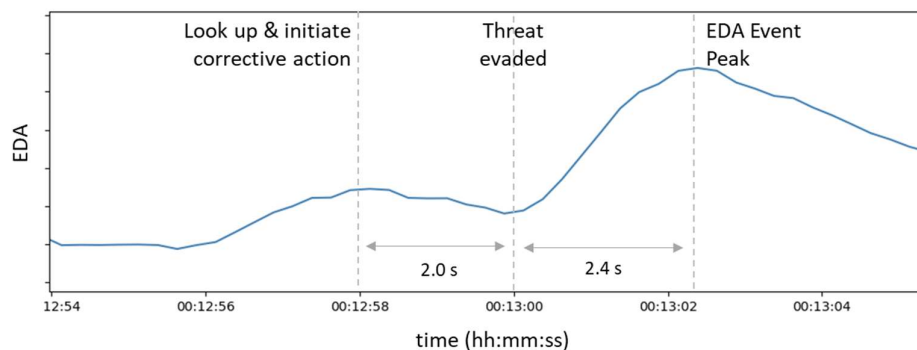


Fig. 2.20 EDA event profile following a sudden and surprising stimulus. Shown here is an EDA event captured in response to a pilot observing the need to redirect their aircraft to avoid collision with mountainous terrain. It is noted the peak of the EDA event occurs approximately 4.4 seconds after the stimulus (even after corrective action had been taken to avoid collision).

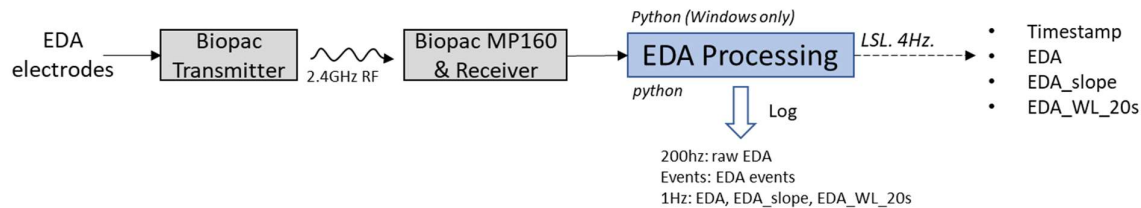


Fig. 2.21 A depiction of the EDA feature extraction processing chain. Raw EDA is collected from the participant’s foot via a BIOPAC physiological data acquisition system which transmits the signal wirelessly to a receiver. The raw data is then read and processed by a custom-written EDA processing and feature extraction program. Multiple features are extracted and broadcast via LSL at 1 Hz for subsequent use.

2.5.1 EDA Data Acquisition

A depiction of the EDA signal acquisition and processing chain implemented in this work is provided in Fig. 2.21. Single-channel EDA is collected at 200 Hz using the BIOPAC MP160 data acquisition system. EDA data are collected using single-use disposable Ag/AgCl electrodes pre-gelled with isotonic gel secured to the medial arch of the participant’s right foot with approximately 8 cm spacing between electrodes. The raw sampled data, measured in microsiemens (μS) is transmitted wirelessly from the transmitter worn on the participant’s ankle to the receiver and MP160. The same tool developed to receive and process ECG data is also used to receive and process EDA data.

2.5.2 EDA Feature Extraction

Features extracted from EDA in this work are listed in Table 2.5. The signals were extracted with the aim of obtaining information sensitive to events over various durations. Raw EDA is smoothed using a rolling average filter of 50 samples. Thus, sampled raw at 200 Hz, the mean of every 50 samples is calculated and passed forward at 4 Hz as the first EDA “feature” for subsequent processing. This slightly-smoothed EDA signal is potentially sensitive to long-term (low frequency) changes in EDA, yet it retains its sensitivity to high-frequency EDA events. Next, the instantaneous rate of change, or slope, of the smoothed EDA signal is also calculated as the difference in EDA magnitude of consecutive samples. This feature extracts information only at the current moment so is sensitive to current and transient (2-5 second) events. Finally, the time-domain waveform length (WL) of EDA over the last 20 seconds is also calculated as the sum of the magnitude of differences between consecutive

Table 2.5 Features extracted from the EDA system.

Feature	Description
EDA	Smoothed electrodermal activity
EDA_slope	Change in EDA over one second
EDA_WL_20s	Time-domain waveform length of EDA over last 20 seconds

samples in this window. This feature is intuitively the cumulative length of the waveform over the segment and is calculated with equation (2.16).

$$WL = \sum_{i=1}^{n-1} |EDA_{i+1} - EDA_i| \quad (2.16)$$

Where:

WL is the extracted waveform length feature,

EDA is the raw instantaneous EDA signal,

and

n is the number of samples in the feature extraction window.

This feature was extracted with the aim of obtaining a metric that accumulated the effects of small, transient, EDA events. In this way, the feature would be sensitive to EDA activity over a longer time-scale than the previous “EDA slope” feature which has no temporal memory.

Although not used in subsequent mental workload estimation processes due to its discrete (and not continual) nature, EDA event peaks are also extracted. EDA event peaks are found by identifying local maxima in the slope of the EDA that exceed a given threshold. In this work, EDA peaks were extracted when the slope exceeded 0.02 $\mu\text{S}/\text{second}$. As presented hereafter in Section 3.2, the system described in that section can react to these detected events in a manner defined by the operator. Potential reactions to these events include auditory notification to the pilot, the co-pilot, or both.

Fig. 2.22 shows the EDA-extracted features over the course of a simulated helicopter flight. It is evident that the three features each highlight different aspects of the signal and may potentially inform an assessment of mental workload across different time scales.

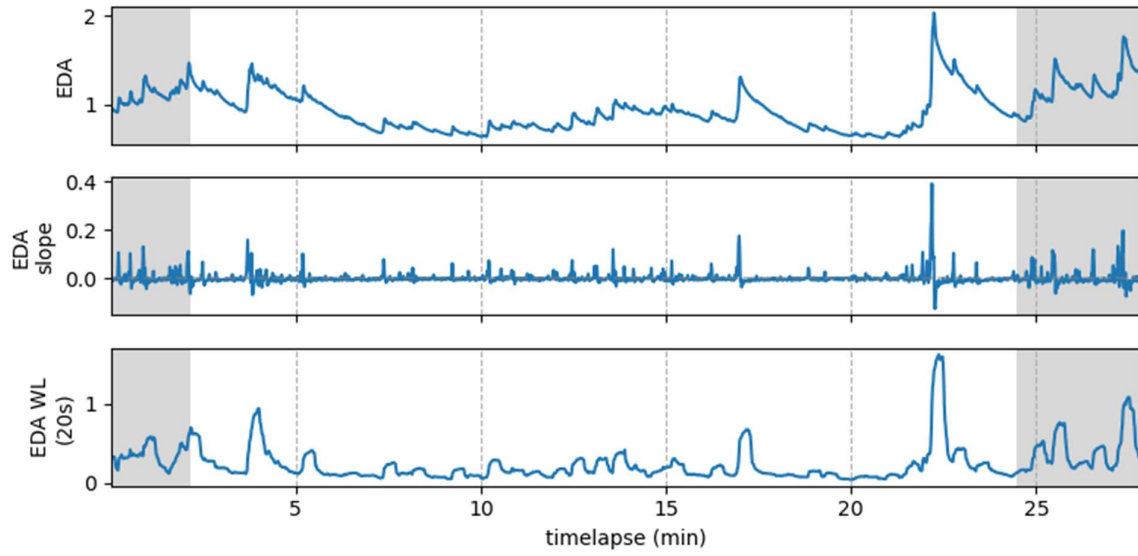


Fig. 2.22 Representative plot of EDA-extracted features over the course of approximately 28 minutes in a helicopter simulator. Regions in grey represent periods during which the helicopter was on the ground.

3 PhysHub: A Pilot Physiological Monitoring and Mental Workload Prediction Tool

To enable experimentation in the flight simulator utilizing the various physiological and behavioral signals described in the previous section, a central control tool needed to be designed and developed. Through an iterative design process, both functional and non-functional requirements were established. Functional requirements are those which specify what the system should be able to do while non-functional requirements are those which describe the attributes or characteristics of the system and specify the constraints for how the system should be built [78]. Functional requirements for the system included:

- The tool must present the status of the individual signal-processing sub-systems and data streams.
- The tool must continuously plot the incoming data streams to enable rapid identification of sub-system issues and provide the experimenter with a consolidated overview of the incoming data.
- The tool must enable the synchronized collection and logging of all incoming data streams. This synchronized logging shall enable the collection of “baseline” measurements of all incoming data streams prior to simulated flight and the storage of sub-system data during flight for offline processing and analysis.
- The tool must fuse incoming data streams into a single-valued metric representative of the pilot’s instantaneous mental workload.
- The tool must enable the creation and manipulation of conditional states which, when met, initiate or “trigger” subsequent notifications or actions.

Non-functional requirements for the system included:

- The tool should be intuitive and transparent to the experimenter without requiring a comprehensive understanding of the algorithms or processes applied in the background.
- The tool should be viewable and controllable through a graphical user interface (GUI).

- The tool must interface with the existing simulator control software to ensure the logging of physiological data is initiated and terminated in unison with the other systems.
- The triggering sub-system should likewise be transparent to both the experimenter and the pilot.

The tool created to satisfy these requirements is known as “PhysHub” and a screenshot of the tool while collecting and processing physiological data from a participant flying a helicopter simulator is provided in Fig. 3.1. Through significant iterative development, PhysHub meets or exceeds all of the design requirements set forth above. Most fundamentally, however, it serves as the centralized control system for fusing the many physiological and behavioral measurements into a single-valued metric representing the operator’s mental workload. Although employed in a simulated aircraft cockpit, the tool is agnostic to the particular application environment and could serve a similar purpose across a wide variety of environments. The tool itself is not intended for use or display to the operator, but rather to an external experimenter. As presented in Section 3.4 however, an in-cockpit interface was designed and implemented to facilitate transparency between the pilot and the tool.

The tool was written using Object Oriented Programming (OOP) in Python using a multitude of data-processing, multi-processing, and multi-threading libraries including NumPy,² pandas,³ pylsl,⁴ threading,⁵ and multiprocessing.⁶ The GUI was designed using PyQt5.⁷ The high-level software architecture of the tool consists of a PyQt5 QMainWindow GUI class containing methods and attributes for proper GUI functionality as well as an instance of a “PhysHub” class through which all the data retrieval and processing is performed. The GUI class contains 46 methods while the PhysHub class contains 31 methods. To facilitate a responsive, continually-updating GUI as well as to enable the simultaneous reception and processing of multiple datastreams, the tool relies extensively on the multiprocessing library to spawn and terminate subprocesses that run in parallel to the main parent process. For the subprocesses which share objects, such as those which monitor and pull data from the incoming Lab Streaming Layer (LSL) streams transmitted by the processing tools described in Section

² <https://numpy.org/doc/stable/>

³ <https://pandas.pydata.org/>

⁴ <https://github.com/sccn/labstreaminglayer/>

⁵ <https://docs.python.org/3/library/threading.html>

⁶ <https://docs.python.org/3/library/multiprocessing.html>

⁷ <https://wiki.python.org/moin/PyQt>

2, server processes are created (using the library’s “Manager” method), which “hold Python objects and allow other processes to manipulate them using proxies.”⁸

The following sub-sections describe the multiple elements of the tool and how the pre-established design requirements were met.

3.1 Visual Inspection of System and Pilot State

The most basic function provided by PhysHub is its receipt and display of system state information. Seen on the upper left-hand side of Fig. 3.1, the status of individual data streams is noted. In the figure, the green text “yes” is seen below each stream name denoting that the LSL streams broadcast from each of the sub-systems (HR/HRV, EDA, fNIRS, Respiration, IMU, Gaze Statistics, and Collision) are all visible to PhysHub and that data can be retrieved. The red button in the upper-left with the text “Stop Pulling Data” would pause the acquisition of data from these streams.

Below the system state labels, data streams from the individual sub-systems are visualized continuously in real-time over the past 60 seconds with the latest-received data shown to the right. A live view of the interface would show the signals slowly scrolling to the left. For those streams with multiple features, the user can select which feature to plot via the selector button to the left of the plot. For example, through the selector button, it is possible to

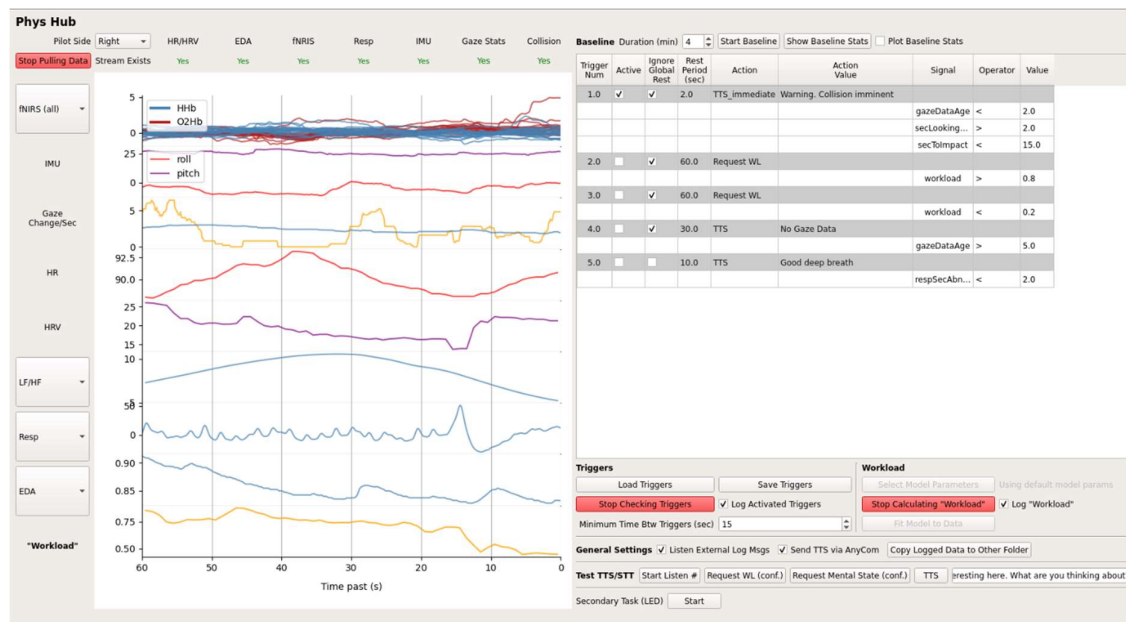


Fig. 3.1 The Phys Hub graphical user interface (GUI) providing a view of the incoming signals and data streams, and enabling the experimenter to collect baseline measurements, fuse the incoming data into a single-valued mental workload metric, set and manipulate “triggers” for interacting with the pilot, the co-pilot, and other systems.

⁸ <https://docs.python.org/3/library/multiprocessing.html#multiprocessing.Manager>

switch between “Resp” and “Resp Rate” to see raw respiration data or the extracted respiration rate. The visualization of these real-time streams allows the experimenter to assess the validity of the incoming streams. Additionally, plotted on the bottom-most sub-plot is the single-valued mental workload metric derived through the fusion of the many individual physiological signals as discussed in detail hereafter in section 4. Through a brief scan of these many signals, the experimenter can identify sub-system errors and glean information about the pilot’s state by noting the trends in the signals.

3.2 Transparent Triggering System

A triggering system was developed which allows the experimenter to set conditions upon which desired actions can be performed. For example, the experimenter can cause a custom notification message to be spoken to the pilot when their heart rate exceeds a particular threshold. The triggering system was developed with the goal of being completely transparent as to when and why a particular action was performed. It has been designed in a way that the custom set of conditions and actions can be imported from and exported to excel files for readability and manipulation outside the tool. The utility of this system is made clear by understanding the first trigger listed on the right-hand side of Fig. 3.1. This trigger contains the following three conditions:

- $GazeDataAge < 2$. Thus, to be evaluated as to be true, the data from the eye tracking system has to be current within 2 seconds.
- $SecLookingDown > 2$. Thus, to be evaluated as to be true, the pilot must be looking down in the cockpit for more than 2 seconds.
- $SecToImpact < 15$. Thus, to be evaluated as true, the helicopter must be within 15 seconds of impact given no course correction.

When each of these conditions evaluates to true, an immediate text-to-speech message is transmitted to the pilot with the text “Warning. Collision imminent.” Thus, this trigger prevents the collision warning message from being delivered when the pilot is looking up out of the cockpit and is arguably aware of their situation. Conditions can easily be edited or added to upon experimenter or pilot request.

Additionally, the system was built with limited speech-to-text functionality making it possible for a trigger to cause a question to be posed to the pilot and for the system to receive an answer from the pilot. This feature was utilized when testing the utility of an automated mental workload verification system which would request the pilot’s subjective mental

workload at pre-determined intervals or upon the meeting of pre-defined signal thresholds. The table in Fig. 3.1 shows two such triggers in which the pilot's subjective mental workload level is requested when the predicted mental workload value is above 0.8 or below 0.2 (on a scale of 0 to 1).

In this work, other than to provide collision warning notifications, the triggering system was used occasionally to notify the pilot of their perceived mental workload level. In this case, a trigger was set to execute when the predicted mental workload value exceeded an established threshold (e.g., 0.8 on a scale of 0 to 1). When this threshold was exceeded, the pilot would be notified with the message "Your workload appears high, consider taking a deep breath and radio for help if needed."

Such a system could be studied and expanded upon with more robustness in the future to assess the utility of an assistant system that interacts with the pilot via text-to-speech.

3.3 Baseline Collection

PhysHub enables the collection of baseline measurements from all sub-systems other than the eye-tracking system. This baseline measurement is utilized in subsequent feature scaling operations as part of the mental workload estimation process. As seen in Fig. 3.1, the experimenter can set the duration of the baseline measurement, start and manually stop (if necessary) baseline data collection, view the baseline statistics, and plot the signals recorded during the baseline collection for validation of data integrity.

Obtaining an accurate baseline measurement of psychophysiological signals in an experimental setting is difficult. Factors such as being in a new place, being anxious about an unknown experience, and perhaps desiring to please the experimenter may all lead to a physiological state unique from a true "baseline." In an effort not to exacerbate this situation, baseline measurements were collected outside the cockpit. It was postulated that an in-cockpit collection would lead to an artificially stressed state. Because the collection was done outside the cockpit, baseline eye-tracking features could not be collected as the eye-tracking hardware was mounted within the simulator. Even if it was available outside the simulator, the utility of eye-tracking baseline metrics collected before flight would be of questionable utility. The features extracted from the eye-tracking signals are highly specific to the environment in which they are collected (e.g., saccade movements are likely very different when looking around a classroom than when flying a helicopter).

3.4 In-Cockpit Pilot Interface to PhysHub

It is surprising that among the plethora of in-cockpit displays available to a pilot to assess the state of their system, there is no known commercial system in use today that presents information on the mental or even physical state of the pilot, co-pilot, or crew during flight. Taking a step in this direction, a prototype in-cockpit display was created to present the pilot's and co-pilot's physiological data and predicted mental workload level in real-time and is depicted in Fig. 3.2. This touch-screen display provides functionality that begins to incorporate real-time monitoring of pilots into the human-machine system of the cockpit. Through this display, the pilot and co-pilot can view their own respective data, but perhaps more significantly, they can view the data of the other. This information may aid each in quickly assessing the mental state of the other – a critical task left today to be performed completely without technological aid. In a survey of ten operational German military helicopter pilots conducted in conjunction with this work, only one objected to their physiological data being viewable to either their co-pilot or ground-station personnel, all others welcomed the idea and considered it a practical aid to enhance crew coordination.

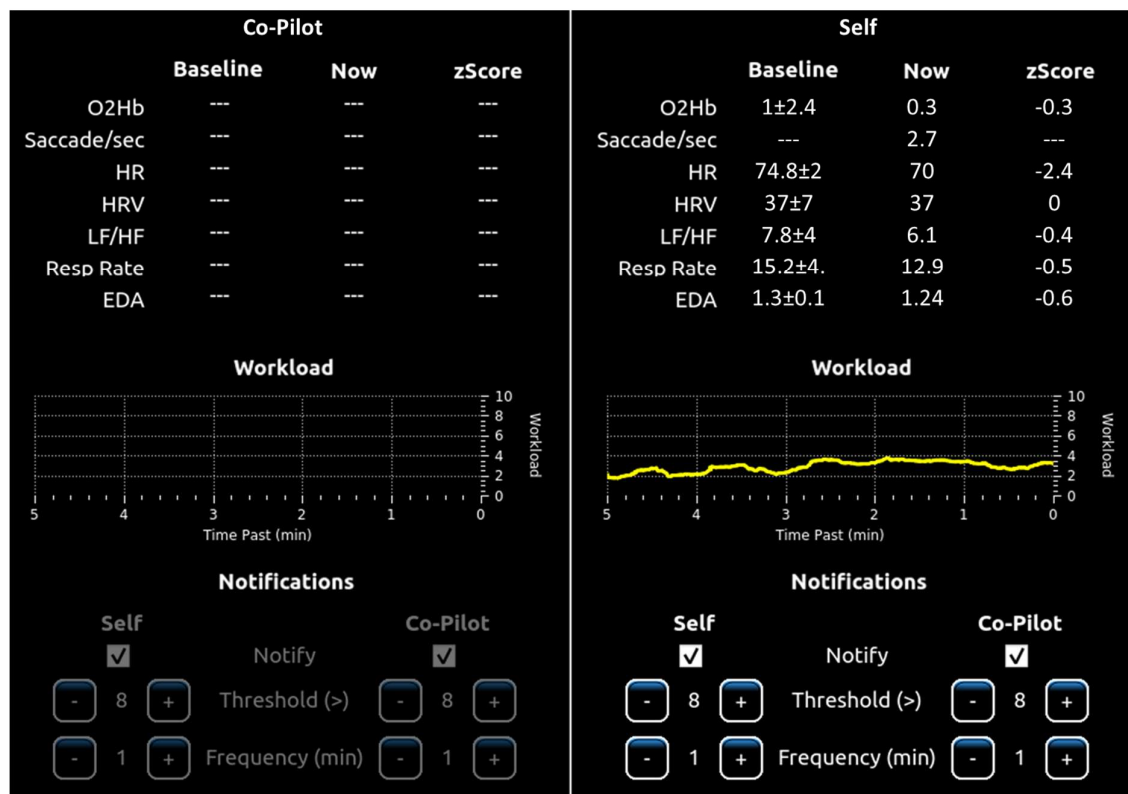


Fig. 3.2 In-cockpit display of physiological data and predicted mental workload. Through this display, pilots can see physiological states and predicted mental workload for themselves and their co-pilot. They can also set the parameters for receiving notifications of their own workload level and for that of their co-pilot. This screenshot was taken with only a single pilot in the cockpit thus is only populated for “self.”

Various physiological data and the predicted mental workload from PhysHub are visualized in the prototype display. To aid in rapid state assessment, a “z-score” showing how many standard deviations above or below the mean the current value is from the baseline mean is provided for all physiological data for which a baseline measurement was taken. Although reviewing this column alone can aid in quickly identifying unusual states, these summary statistics are nevertheless likely overly complex and unnecessary in a future implementation. Future optimization of the display could more prominently highlight the current predicted mental workload level. To encourage acceptance and support ease of use, this could be done through the use of an analog gauge which gauges are ubiquitous in the cockpit.

In the prototype display, following the presentation of individual physiological metrics, a timelapse plot of the predicted mental workload over the last five minutes is provided. This is shown to aid in assessing the temporal characteristics of the data and support an evaluation of the current state as compared to the past.

Finally, as seen at the bottom of the prototype display, the pilots can set parameters for receiving notifications of their own mental workload level and for that of their co-pilot. In the situation depicted, the pilot (“self”) has selected to receive notifications for themselves and for their co-pilot (note the check boxes are both checked). Notifications are set to trigger when the predicted mental workload exceeds a value of 8 (out of 10) with a frequency of no more than one notification per minute.

These options are provided to allow the pilots maximum control over the system. Future experimentation and usability studies are required to determine the utility of the pilot-directed notification system. Controlling various aspects of the automation from the cockpit provides a custom experience for the pilots and one that ensures ultimate human supervisory control.

4 Experimental Testing of a Real-Time Multimodal Mental Workload Prediction System During Simulated Helicopter Flight

This section significantly expands upon my published work “Physiological Sensor Fusion for Real-Time Pilot Workload Prediction in a Helicopter Simulator” published in the Proceedings of the AIAA SciTech 2022 Forum [79].

4.1 Introduction

As summarized and discussed in Section 1, numerous studies have been conducted to evaluate pilot mental state using psychophysiological data during actual or simulated flight yielding a wide range of results. Generally, however, the results converge suggesting that some metric of human mental workload, stress, arousal, or otherwise can be predicted (at least weakly) using one or more psychophysiological signals. The early review article by Roscoe in [80] concluded that heart rate, heart rate variability, and respiration rate are three variables that can be used in simulated and real flight to assess a pilot's arousal. The review article by Jorna in [81] similarly concluded that cardiovascular measures such as heart rate and heart rate variability can help to identify different mental states and observe dynamic responses to mental workload changes in both simulated and actual flight. Additionally, a review article on the topic considered other psychophysiological data including electroencephalography (EEG), electrooculography (EOG), eye blinks, and cardiovascular measures and found correlations among these signals with induced task load [60]. Most recently, Charles and Nixon, in their review article, concluded that there was not a single particular measure that most effectively characterizes mental workload, but rather they suggest that physiological measures “capture the experience of the user” which may prove useful in developing a functional system [58]. Non-review articles presenting original research pursuing a multi-modal approach to assessing mental state during real and simulated flight are summarized in Section 1.3.

This section presents original research aiming to predict mental workload in-real time during simulated helicopter flight. Participants in this study include fully-trained military helicopter pilots as well as un-trained university students. This study applies the novel methods for acquiring and processing various signals as presented in Section 2 of this work. It also utilizes the physiological monitoring and mental workload prediction tool PhysHub presented in Section 3.

The primary aim of this study is to apply and evaluate a real-time mental workload prediction algorithm during simulated helicopter flight. This is done by fitting or training a model to predict subjective mental workload from one mission and applying that model in real-time to data from a second mission yielding a continuous prediction of mental workload. It is hoped that the output of this system could have potential utility to an automated pilot assistant system. The system's effectiveness is evaluated primarily by assessing the linear correlation between the predicted mental workload and the participant's subjective mental workload provided post-mission execution. Secondary research questions include:

- Which features extracted from multiple physiological and behavioral signals are most useful in predicting a pilot's subjective mental workload?
- How frequently and significantly will participants change their subjective mental workload assessment after viewing a system-predicted mental workload?
- How would study participants respond to receiving notifications of perceived high mental workload
- How would the system be subjectively perceived by the active-duty military helicopter pilots and the university students who participated in the study and what is their readiness to engage with such a system in real-flight operations?

This work builds upon the author's previous work investigating the utility of ECG-derived signals, EDA, and various eye-related features in predicting the subjective mental workload of pilots in a helicopter simulator [82]. The authors have additionally published work investigating the utility of fNIRS in pursuit of this objective [61]. The complexity of data collection and synchronization was significantly greater in this work than in these previous studies. Data originating from multiple sensors and sources are collected at different rates and unique processing is applied to each signal enabling the extraction of relevant features. These processed signals are then time-synchronized and merged with a common sampling rate enabling the application of supervised machine learning algorithms which predict the mental

workload of the study participants in real-time. The prediction is then used to update an in-cockpit display and is used to trigger notifications to the pilot.

In some instances, when the prediction of mental workload exceeds a set threshold, study participants are provided real-time feedback on their perceived high mental workload. This feedback is provided via a text-to-speech interface which verbally informs the participant of their current perceived mental state and suggests they make an effort to relieve perceived tension. The pilots are also encouraged to observe their predicted mental workload level along with other physiological metrics through a display integrated into their glass-cockpit. The current implementation of this display can be seen in Section 3.4. It is hypothesized that by providing pilots (or others such as ground station personnel) with real-time pilot-state information, they will be supported in remaining in a state suited to fulfill the uniquely human roles they are to perform. This support may come in the form of adaptive automation [83], [84], but is also anticipated to come through improved crew communication as well as through self-regulation strategies of the pilots themselves.

4.2 Methods

Ten operational helicopter pilots of the German Air Force and ten university students with various levels of experience with simulated helicopter flight participated in this experiment. After a thorough explanation of each sensing modality and the experimental protocol, each participant provided voluntary and informed consent for participation in the

Table 4.1 Participant Summary

<i>ID</i>	<i>Age</i>	<i>Gender</i>	<i>Flight Experience</i>
Pilot 1	43	M	1900 hrs
Pilot 2	29	M	170 hrs
Pilot 3	51	M	3000 hrs
Pilot 4	43	M	3000 hrs
Pilot 5	43	M	2200 hrs
Pilot 6	51	M	3600 hrs
Pilot 7	44	M	2500 hrs
Pilot 8	48	M	4000 hrs
Pilot 9	25	M	Test Flight Engineer. Many hours helicopter simulator
Pilot 10	44	M	2000 hrs
Student 1	21	F	2 hrs lab simulator
Student 2	22	F	Flight screening tests, 15 min lab simulator
Student 3	23	M	120 hrs glider and motor aircraft
Student 4	25	M	100 hrs home computer simulator, 1 hr lab simulator
Student 5	22	M	No flight experience (real or sim)
Student 6	22	M	200 hrs glider and motor aircraft, 1 hr lab simulator, flight screening test
Student 7	23	M	600 hrs glider, home computer simulator, 30 min lab sim
Student 8	24	M	Flight screening tests, 4 hrs helicopter simulator
Student 9	20	M	Flight screening tests
Student 10	25	M	Flight screening tests, home computer simulator

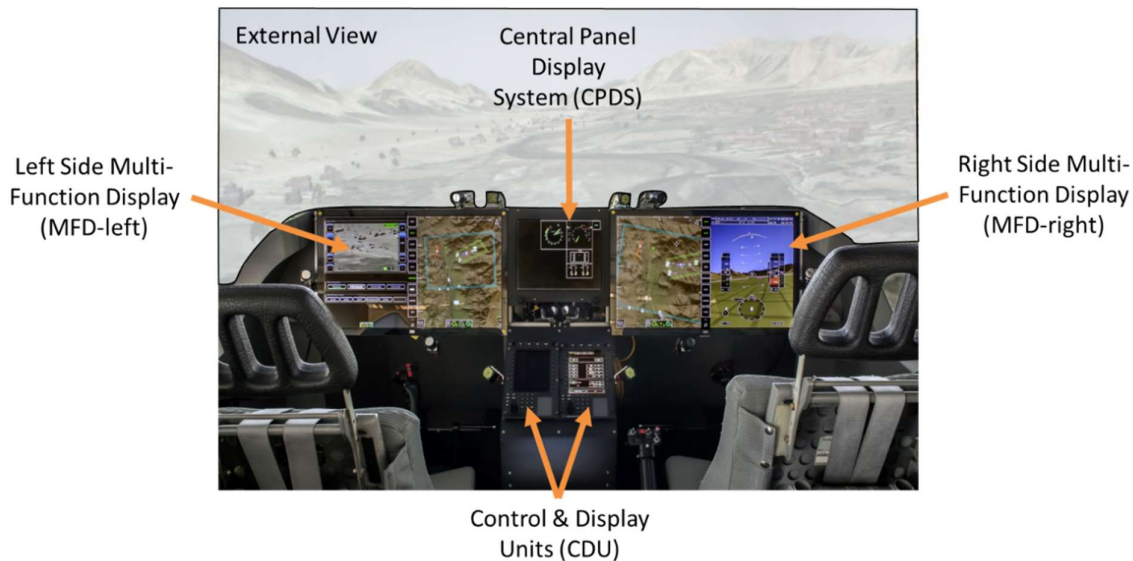


Fig. 4.1 A photograph of the helicopter simulator used in this study. Although the simulator has positions for a pilot and a co-pilot, study participants executed the simulated missions alone in the cockpit as a single pilot.

study. The form used to provide this consent (the “Einverständniserklärung”) is provided as Appendix C. Each participant completed the experiment with no prior knowledge of the experiment design. Table 4.1 provides a summary of the participant demographics including a short description of their flight experience. The experiment was conducted in a research-focused helicopter simulator equipped with touch-screen multi-function displays and three projector-based external views. Although the cockpit can seat both a pilot and a co-pilot, the experiment was conducted individually with a single pilot in the cockpit (the study participant). A photograph of the simulator’s cockpit is given in Fig. 4.1. Visible in the photograph are the right and left Multi-Function Displays (MFDs), the Central Panel Display System (CPDS), the Control and Display Units (CDUs), as well as the flight control elements including the pedals, the cyclic, and the collective.

Utilizing a “page selector” on the touch-sensitive MFD, the pilot can access multiple elements of information including standard IFR/VFR displays, dynamically-updating maps, system settings, preflight and landing checklists, and the in-cockpit PhysHub pilot interface as presented in Section 3.4. Views of these MFD “pages” are given in Appendix E.

A high-level overview of the experiment design is depicted in Fig. 4.2. Following simulator familiarization, training, and setup, two consecutive simulated training missions are flown each lasting approximately 20 to 30 minutes during which physiological and behavioral signals are recorded. The details of these missions and the tasks required of the participants are provided hereafter in Section 4.2.1. Immediately following each simulated mission,

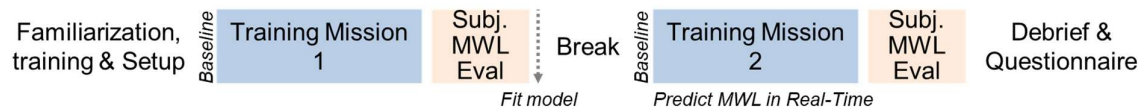


Fig. 4.2 High-level experiment design block diagram. Following simulator familiarization, training, and setup, the experiment consists of two simulated missions each followed by a subjective evaluation of mental workload. A prediction model is trained on data from the first mission to predict the mental workload of the pilot in real-time during the second mission.

participants are guided in providing an assessment of their subjective mental workload over the course of the mission. The details of how this assessment was made are given hereafter in Section 4.2.2. Following the first mission, parameters are learned for a model aimed at predicting the participants subjective mental workload given the various recorded physiological and behavioral signals. This model is then applied in real-time during the second mission to predict the participant’s mental workload. Details regarding the model selection, training, and real-time implementation are provided in Section 4.2.3. Following the second subjective mental workload evaluation, sensors are removed and participants complete a post-experiment questionnaire. In total, participation in the experiment spans approximately three hours. Following the completion of the experiment, offline analysis is conducted to assess the degree to which the participants’ subjective mental workload assessment correlated with the predicted mental workload measurement provided by the model.

4.2.1 Simulated Mission Design

Two simulated military helicopter training missions were designed to elicit varying degrees of mental workload throughout each mission. The missions were designed using the military training simulation software Virtual Battlespace 3 (VBS 3) by Bohemia Interactive Simulations.⁹ The helicopter chosen for simulation was the Sikorsky S-76c and its dynamics were simulated using X-Plane.¹⁰ The use of this particular aircraft was selected by previous researchers and integrated into the laboratory’s simulator. The use of VBS enabled the manipulation of multiple elements of the simulated environment including ground vehicles, aircraft, ground personnel, point sources of smoke, and weather. Maps of the two missions are shown in Fig. 4.3.

⁹ <https://bisimulations.com/products>

¹⁰ <https://www.x-plane.com/>

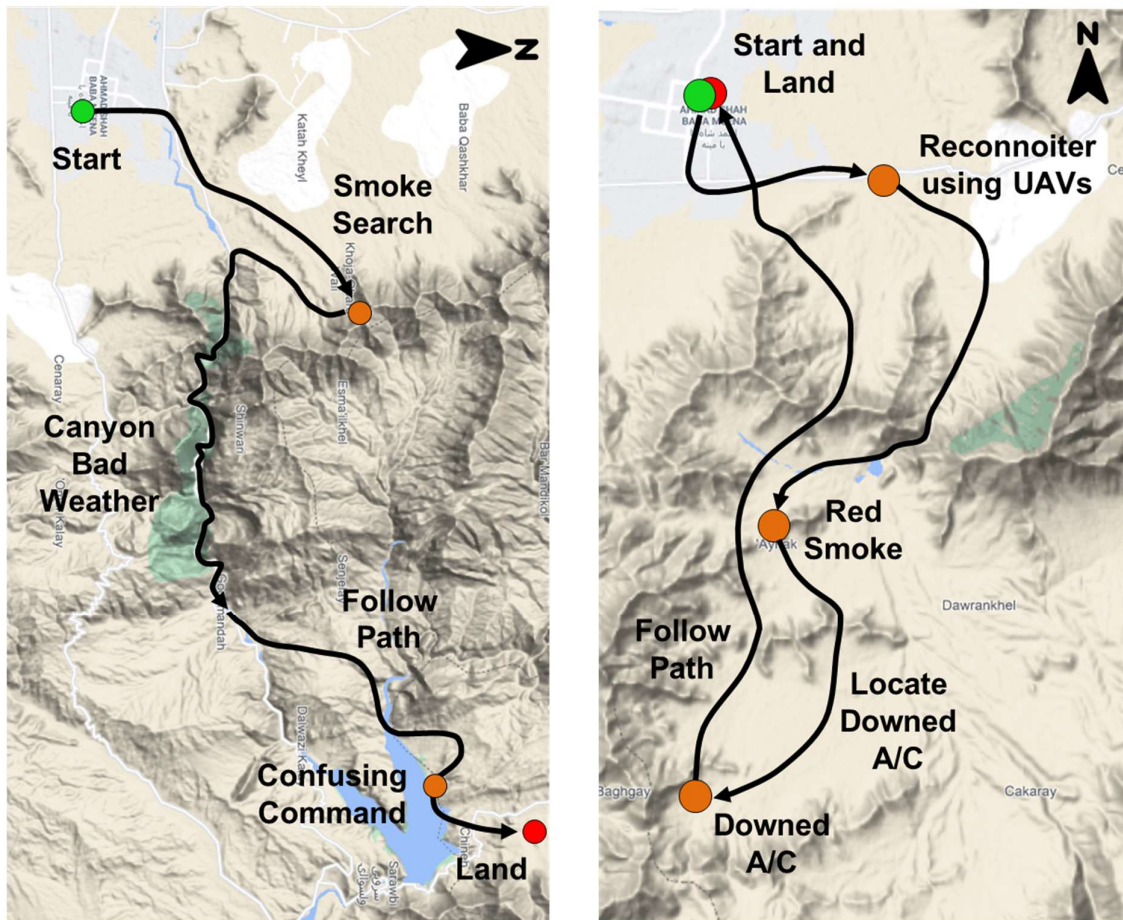


Fig. 4.3 Maps of the two simulated helicopter missions designed to elicit varying levels of mental workload. Mission 1 is shown to the left while Mission 2 is shown to the right.

As previously stated, the two missions were each designed to elicit mental workload states across the arousal continuum from relaxed, to focused, to overwhelmed. Additionally, each was designed to briefly induce a feeling of confusion in the participants. How the elicitation of these states was designed into each mission is summarized in Table 4.2. Relaxed states were designed into each mission during rest periods on the ground and in the air during unconstrained free flight. Focused states were induced through visual search and constrained manual flight tasks. Situations of high mental workload were elicited through low-level canyon

Table 4.2 Mission Design - Elicitation of Various Mental States

Intended Elicited Mental State	Mission 1	Mission 2
<i>Relaxed (non-flight)</i>	On ground post-mission	On ground post-mission
<i>Relaxed (simple unconstrained flight)</i>	Unconstrained flight to specified area	Unconstrained flight to final landing site
<i>Focused Visual Search</i>	Search for colored smoke	Search for downed aircraft
<i>Focused Manual Control</i>	Hover	Follow path while maintaining set altitude
<i>Overwhelmed</i>	Narrow canyon with poor weather	Maintain set altitude while using UAVs to reconnoiter for target
<i>Confused</i>	Inexecutable command	Long and complex command

flight in poor weather conditions and through the tasking of unmanned aerial vehicles (UAVs) from the cockpit's touchscreen Multi-Function Display (MFD) while maintaining a given altitude.

Following a high-level pre-brief of each mission (provided in its original form in Appendix F), study participants were led through the execution of each mission by means of verbal direction given by the experimenter at pre-defined stages of each mission. In general, the participants were directed to particular waypoints identifiable on their MFD where, upon arrival, they received additional instruction. The specific conditions and resulting actions for each mission are provided in Appendix G.

4.2.2 Subjective Mental Workload Assessment

Supervised learning requires that a training dataset be provided in which the outcome variable is “known” to train a model and learn the weights of unknown predictors. In this case, where the objective is to predict mental workload, training data is required with an associated mental workload “truth.” However, because mental workload is largely a subjective phenomenon, generating a continuous-valued “perfectly true” metric of mental workload over some duration of time is impossible complicating the application of supervised machine learning in this scenario. Despite the challenge, however, approximations can be reasoned. As presented in Section 1.1, researchers in the field apply a wide range of techniques for assessing mental workload in these scenarios including via performance on secondary tasks ([55]), questionnaires (e.g. NASA-TLX) ([38], [40], [56]) and intermittent requests of the participant for a subjective assessment ([82]). However, assessing mental workload using these approaches, is not ideal for this scenario. Measuring mental workload based on the performance of a secondary task may distract from or hinder performance on the primary task. On the other hand, assessing mental workload through questionnaires or intermittent requests provides only discrete snapshots in time where it is difficult to extrapolate between samples. Thus, in this work, an approximation of subjective mental workload is obtained through a novel approach yielding a continuous value throughout the flight scenario. The resulting metric can then be applied as the “known” outcome variable by which a variety of supervised learning prediction models can be trained.

In this work, a continuous subjective mental workload rating is provided over the course of a simulated helicopter mission by the participant following mission execution. Fig. 4.4 subfigure (a) shows the Likert scale provided to the participants to evaluate their subjective

mental workload on a scale from 1 to 10. The coloring of the scale and the adjectives used to describe a low, intermediate, and high mental workload level provide a common reference for all participants when using the scale. Care was taken in the selection of adjectives presented in conjunction with the scale. “Relaxed,” “focused,” and “overwhelmed,” describe unique positions on a continuum of arousal. Others have suggested that “boredom” may be an appropriate adjective for the far left position on the spectrum [57], [80], [85]. Boredom, however, describes a lack of interest or engagement and is not a state of arousal. Yes, a lack of arousal may lead to boredom and this state should certainly be avoided in aircraft operations, but boredom was not a focus of study in this work.

With the Likert scale in mind, participants provided their subjective assessment while viewing a video and audio playback of the executed mission immediately following mission completion. Each participant was supported by the same experimenter during this phase of the experiment who attempted to remain neutral and indifferent to the participant’s input. As part of this work, a tool was built to facilitate this subjective mental workload data collection. This tool is shown in Fig. 4.4. To aid the participant in their immersion and recall, the video playback portion of the tool provides views of the external and internal displays as recorded during flight. A snapshot of the video playback is shown in the upper section of subfigure (b). During video playback, the pilot’s gaze fixation location is also displayed (as seen by the green marker in the upper third of the display). These visual and auditory cues support the participant in their ability to recall their state of mind during the mission and provide a continuous subjective mental workload assessment over the duration of the mission. The plot in the lower portion of subfigure (b) is manipulated with the support of the experimenter to reflect the participant’s subjective mental workload throughout the recently-concluded simulated helicopter mission. The process of reviewing the previously-flown mission and obtaining the participant’s subjective mental workload assessment using this tool typically requires approximately one-quarter of the total mission duration. Thus, for a 20-minute mission, it could be expected that this process would last approximately 5 minutes.



Fig. 4.4 Tool developed to gather a participant's subjective mental workload post-flight. Subfigure (a) shows the Likert scale used by the participants to evaluate their own subjective mental workload on a scale from 1 to 10. Subfigure (b) shows the graphical user interface (GUI) used by the experimenter and participant to collect the subject's subjective mental workload over the duration of a recently-concluded simulated helicopter mission. Views internal and external to the cockpit are provided along with the pilot's gaze location (green icon) providing rich contextual information to aid in subjectively assessing mental workload. The data in the plot is edited by the experimenter as directed by the participant to reflect the participant's subjective mental workload over the course of the simulated mission.

After the participant's subjective mental workload assessment had been provided and the data was recorded, a plot of the system-generated predicted mental workload was overlaid on the plot with their subjective assessment. Each participant was then asked by the experimenter "According to what you see here, would you like to change your initial assessment in any way?" Participants were then permitted to re-consider their assessment and re-address any portion of the mission they desired. If changes were requested, they were made and the data was saved separately from the initial assessment. This procedure facilitated the collection of data required to determine what proportion of the participants would edit their subjective assessment after having seen a system-generated prediction, one of the study's secondary research questions.

4.2.3 Physiological and Behavioral Data Acquisition and Baseline Measurement

Many signals and features were extracted from functional near-infrared spectroscopy (fNIRS) recordings, a cockpit-mounted eye-tracking system, electrocardiography (ECG) recordings, electrodermal activity (EDA) recordings, and recordings from stretch-based respiration belts. The wearing of these sensors can be seen in Fig. 4.5. A list of all extracted signals and features is given in Table 4.3. The extraction of these signals and features are described in detail in Section 2 of this work. A summary of these data acquisition and feature extraction methods is provided here for readability.

fNIRS is collected from the subject's pre-frontal cortex using the fNIR203C headband from fNIR Devices. The device transmits 730 nm and 850 nm wavelength light and has 16 channels with 25 mm optode spacing and two "short channels" with 10 mm spacing. Before calculating the change in oxygenated and deoxygenated hemoglobin, signal pre-processing and filtering is conducted on the raw light intensity data to remove interference caused by the cockpit-mounted eye-tracking system (see Section 2.1.1 for the details of this noise removal process). Following noise removal, change in oxygenated and deoxygenated hemoglobin concentrations are calculated using the modified Lambert-Beer law. Change in concentration was calculated with respect to the median of each channel over the previous two minutes. The short channels were then subtracted from the signals on their respective hemisphere. Finally, features were extracted as noted in Table 4.3.

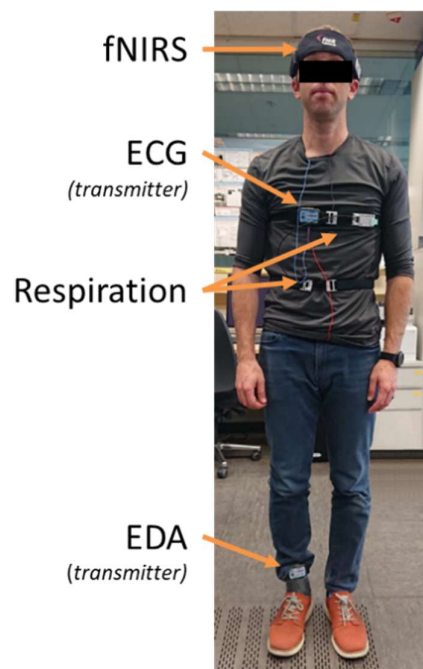


Fig. 4.5 All sensors worn by each participant during simulated missions.

Table 4.3 Summary of Extracted Features

Abbreviation	Description	Used in	
		Baseline Taken	MWL Pred.
HR	Mean inter-beat-interval (IBI) over the last 20 seconds	✓	✓
HRV	Heart rate variability calculated using RMSSD [76] over last 30 seconds	✓	✓
LF/HF	IBI low-frequency to high-frequency power ratio over last 120 seconds	✓	✓
LF/HF_slope	Change in LF/HF over one second	✓	✓
EDA	Electrodermal activity	✓	-
EDA_slope	Change in EDA over one second	✓	✓
EDA_WL_20s	Time-domain waveform length of EDA over last 20 seconds	✓	✓
respRate	Respiration rate calculated using an adaptive IRR notch filter [72], [73]	✓	✓
μ_{HHb}	Mean deoxygenated hemoglobin across all 16 fNIRS channels	✓	✓
μ_{O2Hb}	Mean oxygenated hemoglobin across all 16 fNIRS channels	✓	✓
σ_{HHb}	Instantaneous standard deviation across all 16 deoxygenated hemoglobin signals	✓	✓
σ_{O2Hb}	Instantaneous standard deviation across all 16 oxygenated hemoglobin signals	✓	✓
Spatial Asymmetry O2Hb	Difference between right and left hemisphere mean oxygenated hemoglobin	✓	✓
gazeBlinkPerMin_60s	Number of eye blinks in the last minute	-	✓
gazeChangePerSec_5s	Frequency of fixation change from above to within cockpit over last 5 seconds	-	✓
gazeChangePerSec_60s	Frequency of fixation change from above to within cockpit over last 60 seconds	-	✓
gazeSaccadePerSec_5s	Frequency of saccades in the last 5 seconds	-	✓
gazeSaccadePerSec_60s	Frequency of saccades in the last 60 seconds	-	✓

Eye movement is captured using the commercial SmartEye Pro eye-tracking system. System outputs including blink occurrences, saccadic movements, and gaze location are utilized to extract the eye-related features noted in Table 4.3.

Single-channel ECG and EDA data are collected at 200 Hz using the BIOPAC MP160 data acquisition system. ECG data are collected using single-use disposable Ag/AgCl pre-gelled electrodes on the participant's chest arranged to sample the Lead II vector. The participant's skin is cleaned using alcohol wipes and slightly abraded prior to electrode placement. EDA data are collected using single-use disposable Ag/AgCl electrodes pre-gelled with isotonic gel secured to the medial arch of the participant's right foot with approximately 8 cm spacing between electrodes.

Finally, respiration is captured using two custom-built stretch sensors worn across the chest. The signals from the individual belt sensors are aggregated and filtered before respiration rate is extracted using an adaptive IIR notch filter (ANF) [71], [72].

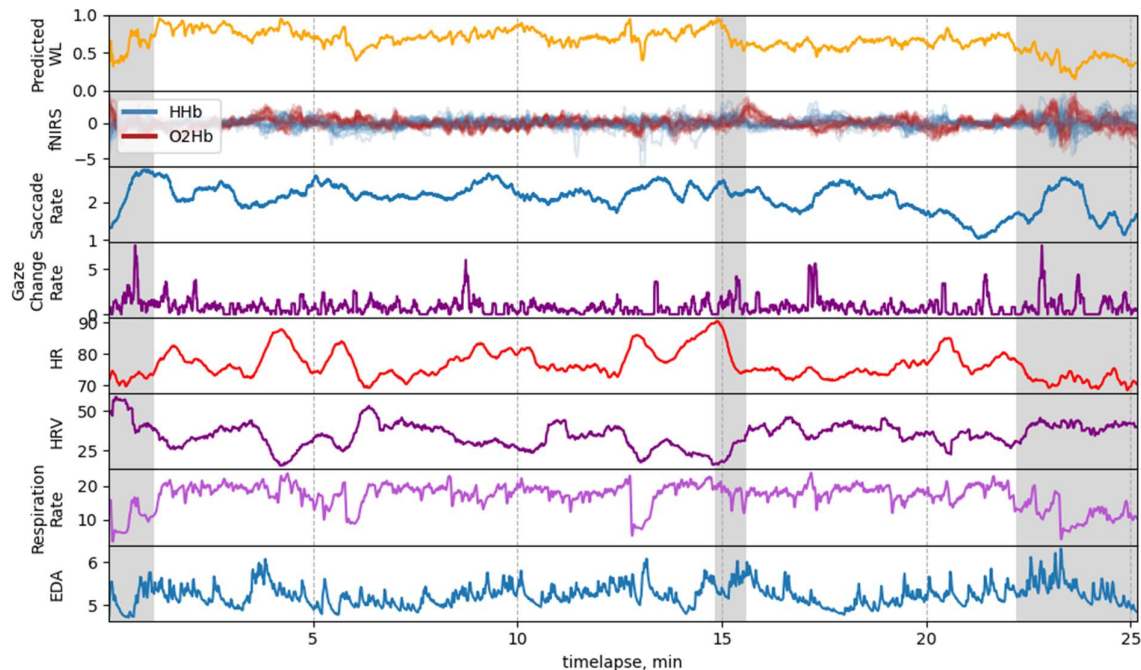


Fig. 4.6 Extracted signals and features over the course of a simulated helicopter mission from an active-duty military helicopter pilot (Pilot 2). Areas shaded in grey denote time during which the helicopter was on the ground.

A selection of all the extracted features over the course of a simulated helicopter mission collected in preparation for the experiment documented in this work are plotted in Fig. 4.6. In addition to the extracted physiological signals, a system-generated predicted mental workload plot is also provided. As can be visually appreciated by the plot, it is evident that each signal exhibits a unique temporal profile over the course of the mission. Some signals change rapidly with high-frequency components such as “EDA” and “Gaze Change Rate,” while others exhibit more gradual temporal changes, such as “HR” (heart rate).

As noted in Fig. 4.2, immediately before the execution of each simulated mission, a four-minute baseline measurement of all physiological signals was collected. It was reasoned that once seated in the cockpit, the participant may experience a heightened level of stress or arousal in anticipation of task execution. Thus, the baseline measurement was collected before the participant entered the cockpit. Signals and extracted features for which a baseline measurement was taken can be seen in Table 4.3. The data collected during the baseline measurement are a critical element of the mental workload prediction as described in the subsequent section.

4.2.4 Real-time Mental Workload Prediction through Supervised Learning

One primary objective of this work was to generate a scalar metric that represents a pilot’s mental workload in real-time from multiple input physiological signals and features

with the intent that it could then be used as a form of pilot feedback or as an input to some form of an assistant system. As discussed in Section 1.1, mental workload is a rather nebulous concept and is influenced by multiple factors including task load (and all relevant environmental factors) as well as the unique characteristics, thoughts, and perceptions of the individual performing the task. It could be reasoned then, that mental workload cannot or should not be distilled into a scalar metric. This reasoning however prevents effective operationalization and system development. Of course, it must be acknowledged that any single-valued metric purporting to “be” mental workload is incomplete. However, as suggested by Honecker et al. in [85], an absolute metric is required for an adaptive system to “make in-situ deliberative, and hence, absolute decisions.” As further argued in that work, to subsequently utilize the absolute metric, however, a considerable understanding of the context is required. Given then a rich contextual understanding, whether from a human co-pilot or an advanced autonomous system, a single-valued metric would support simple and transparent rules for triggering assistance or intervention.

Initially, attempts were made to apply various classification algorithms for producing a discrete mental workload output (e.g., “low,” “medium,” and “high”). This initial experimentation (some of which was published in [61]) yielded inconclusive results. Supervised machine learning classification methods tested included: Linear Discriminate Analysis (LDA), Support Vector Machine (SVM), Convolutional Neural Networks (CNN), and decision tree classifiers. An analysis of these initial results suggested that pursuing a continuous single-valued output may be more fitting for the desired use case. Rather than prematurely binning mental workload into a small number of discrete levels, a continuous value may provide greater transparency to an operator and would allow for greater flexibility to a system (human or machine) utilizing the value to provide support or assistance. Because of the desire for a continuous output, regression models capable of producing a continuous output were pursued including linear regression, lasso regression (linear regression with L1 regularization), ridge regression (linear regression with L2 regularization), lasso regression with polynomial features, K-Nearest Neighbors regression, and Support Vector Regression. Prior to experimentation with active-duty helicopter pilots, preliminary experimentation was conducted with multiple students and researchers at the German Armed Forces University of Munich to determine the best candidate for effective prediction during real-time implementation. Through this testing, it was found that the mental workload prediction generated through linear regression most strongly correlated with the user’s subjective mental

workload. This analysis highlighted the strength of linear models over higher-order, kernel-based, or neural-network-based models in this application. Although this may, in part, be due to an insufficient quantity of data on which the more complex models were trained, it is also unlikely that complex, higher-order relationships between the signals exist thus suggesting a linear model may not only be adequate, but appropriate.

Due to these findings as well as the algorithm's ease of interpretation and visualization (which ultimately increases transparency and trust), subsequent mental workload prediction experimentation was conducted utilizing multivariate linear regression. The outcome variable y (subjective mental workload) would be regressed onto the predictors X (multivariate physiological data) to learn the unknown model parameters. Once trained, the model could be applied to predict subjective mental workload given new physiological data.

The potential success of a linear regression approach to predicting subjective mental workload using physiological measures can be assessed by determining whether or not linear correlations exist between the outcome variable (subjective mental workload) and the individual predictors (physiological measures). A common method for assessing the linear correlation between two signals is by calculating Pearson's Correlation Coefficient for the two signals. This metric ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. Pearson's Correlation Coefficient (r) is calculated according to equation (4.1).

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (4.1)$$

Where:

r is the correlation coefficient,

x_i are the values of the x-variable in a sample,

\bar{x} is the mean of the values of the x-variable,

y_i are the values of the y-variable in a sample,

and

\bar{y} is the mean of the values of the y-variable.

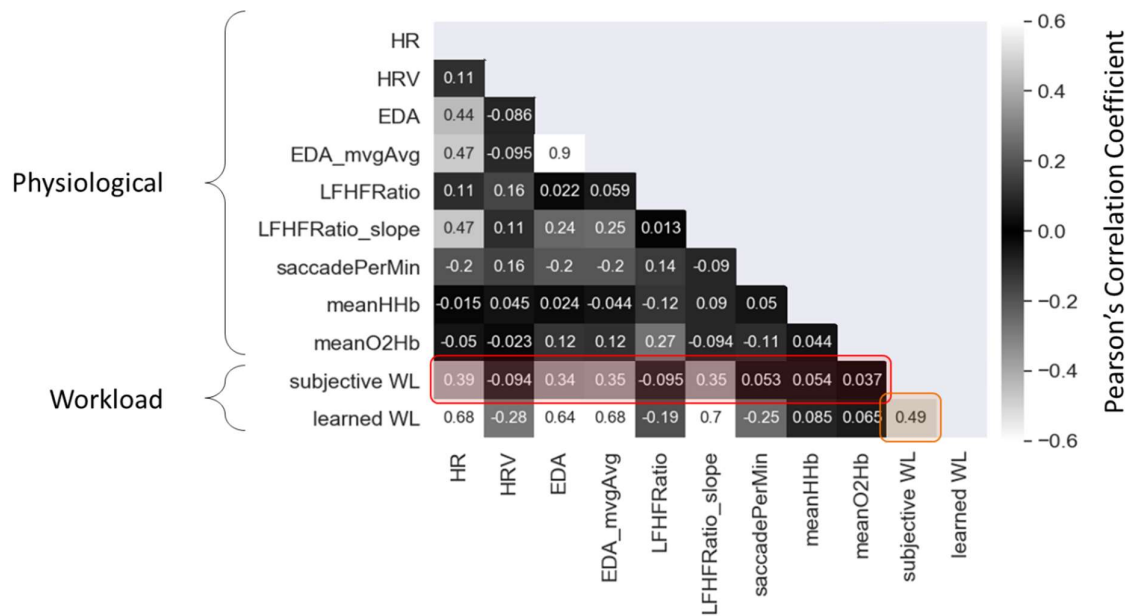


Fig. 4.7 Correlation coefficient matrix showing the linear relationship between subjective mental workload and various physiological signals.

Fig. 4.7 provides a matrix of Pearson's Correlation Coefficients for a variety of physiological measures and a participant's subjective mental workload for a dataset collected during preliminary testing. As noted in the highlighted row of data, linear correlations between subjective mental workload and physiological measures exist between -0.01 and 0.39. Significantly, it is also noted that the correlation between subjective mental workload and the predicted (or "learned") mental workload assessed through linear regression, is greater than the correlation of any one signal. This suggests that a linear combination of physiological features provides more predictive value than any individual feature alone.

Armed with the confirmation that linear correlations exist between physiological signals and subjective mental workload, a procedure was established to apply linear regression-based prediction to two consecutive simulated flights each lasting approximately 20 to 30 minutes. This procedure is illustrated in Fig. 4.8. Following the first simulated mission during which physiological data is recorded (see Table 4.3 for a complete listing of signals and extracted features), the participant's subjective mental workload is obtained through a mission playback and analysis tool as outlined in Section 4.2.1. Using the data collected during the first training mission and the participant's subjective mental workload evaluation over the course of the mission, the model's parameters are subsequently fitted using ordinary least-squares linear regression [86]. This process would ideally establish a model with parameters unique to

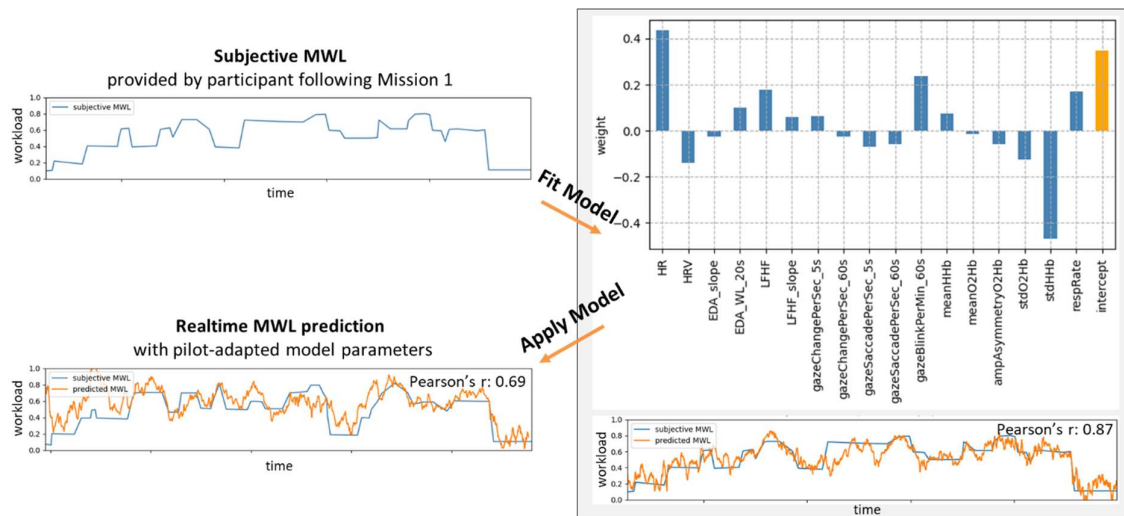


Fig. 4.8 Procedure for fitting and applying a linear regression model to two consecutive simulated helicopter missions. Following the first mission, the participant provides their subjective mental workload assessment as described previously (upper-left). Parameters for a linear regression model are fitted using the data obtained from the first mission (right). During the second simulated flight, mental workload is predicted in real-time using the pilot-adapted linear model. The data shown here is from participant *Pilot 10*.

that participant. The learned weights for one example participant are plotted in the bar plot of Fig. 4.8. Finally, during the second simulated mission, the pilot-adapted model is applied to generate a real-time prediction of mental workload.

It is noted, that for this participant, the fitted weights generally align with previously-reported correlations (see Section 1.3). For example, the weight learned for heart rate is rather large and positive (0.44) suggesting that this feature correlates strongly with the participant's subjective mental workload which aligns with the oft-published relationship between heart rate and mental workload. Similarly, the weight for heart rate variability is negative and the weight for increased electrodermal activity is positive which both agree with previously-published generalizations about these physiological signals and their relation to mental workload. Other weights are less easily interpreted, however, such as that for blink frequency. In this case, the large learned weight suggests that for this participant, an increased blink frequency is positively correlated with an increased mental workload. This relationship contradicts the general findings of others which typically show blink frequency decreases with an increased mental workload (again see Section 1.3). These findings highlight the potential value of this pilot-adapted mental workload prediction approach. It may be possible to generate a pilot-specific mental workload prediction model which outperforms a pilot-agnostic model by accounting for individual peculiarities.

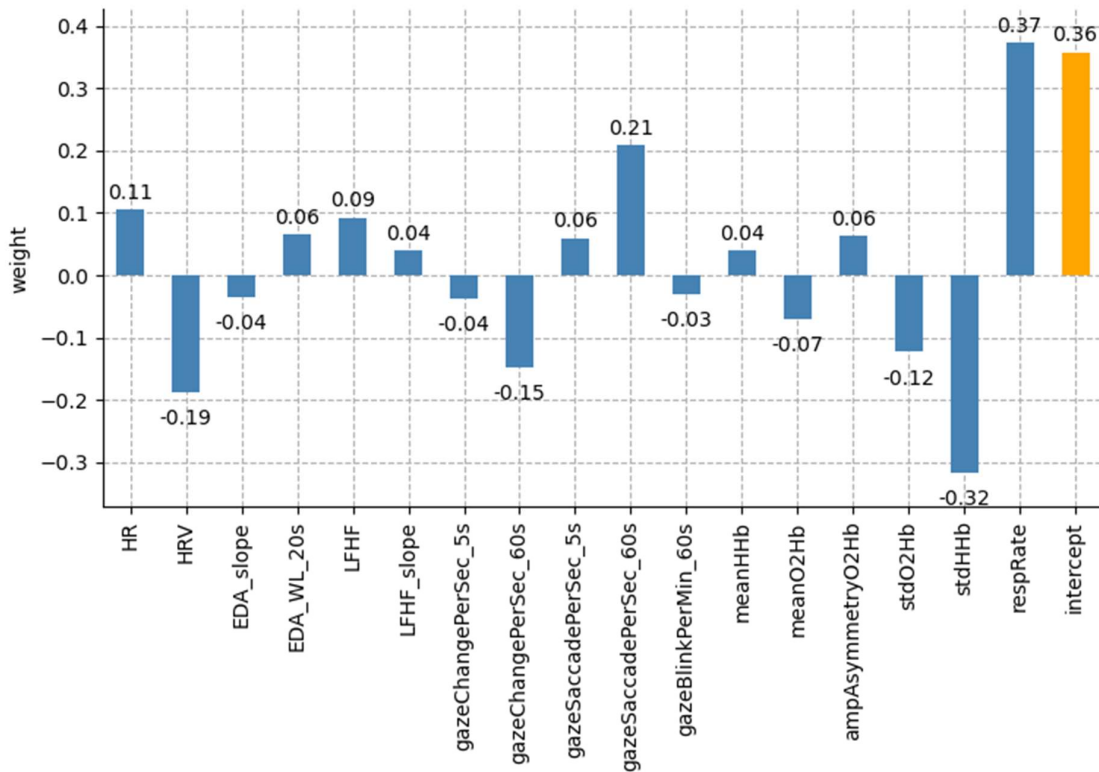


Fig. 4.9 Default model weights used to predict mental workload for each study participant during the first of two simulated helicopter missions. Mental workload prediction on the first mission supplemented the data available for assessing how frequently study participants changed their subjective mental workload assessment after having been shown a system-generated prediction.

As stated previously, a secondary objective of this experiment was to determine the proportion of the participants who would change their subjective mental workload assessment after having been presented with a system-generated mental workload prediction. To generate more data for this analysis, in addition to the presentation of a subject-specific mental workload prediction following the second mission, a subject-agnostic mental workload prediction was presented following the first mission as well. Mental workload prediction during the first mission was performed using a linear regression model with default weights having been obtained through prior data collection and model fitting and are shown in Fig. 4.9. The default weights were used from a model which performed particularly well on a dataset collected from a student researcher who was not a participant in the final experiment. Thus, each participant was asked twice whether or not they would update their subjective mental workload after having been presented with a system-generated prediction. This procedure resulted in 40 samples on which this particular analysis could be performed. It was hypothesized that the mental workload prediction on the second mission would be more strongly correlated with the

participant's subjective mental workload than that from the first for which the default model was used.

To enable the use of the default linear model during the first mission by all participants, it is necessary to scale the input features. Scaling the input features accounts for the variability in baseline measurements across participants. For example, a participant with a baseline heart rate of 60 bpm whose current heart rate is 60 bpm should not be predicted as having a lower mental workload than a participant with a baseline heart rate of 70 bpm whose current heart rate is 70 bpm (if heart rate was the only input feature). Rather, both should be predicted as having the same level of mental workload. To facilitate this normalization across participants (as well as the application of a wide variety of machine learning prediction methods, some of which require scaled/standardized data), min-max scaling was applied to the collected data which scales the range of each feature to $[0, 1]$. Minimum and maximum values observed during the baseline data collection before mission execution were used to initialize this scaling for real-time processing. During mission execution, the minimum and maximum values used for feature scaling were dynamically updated as new data was received.

4.2.5 Post-Flight Questionnaire and Data Analysis

Following the execution of the second simulated flight and the corresponding mental workload assessment, each participant completes a post-flight questionnaire. Questions are posed to assess the participant's engagement with each mission and their commitment to a successful outcome. As discussed in Section 1.1, a person's personal engagement and/or commitment to a task influences the experienced mental workload. In addition to these questions, others are posed to elicit subjective perceptions and beliefs regarding the physiological monitoring of pilots in actual flight. Multiple questions are presented only to the active-duty military pilots as they are uniquely suited to address the topic. All questions, as well as participant responses, are provided in Appendix D.

After all participants had completed the experiment, subsequent analysis is performed to evaluate the study's research questions. The strength of the real-time mental workload prediction is assessed by determining Pearson's Correlation Coefficient between the predicted and subjective mental workload for both Mission 1 and Mission 2.

To assess the relative utility of each feature in predicting the participants' subjective mental workload, the correlation between the participant's subjective mental workload and each feature for both Mission 1 and Mission 2 is determined and sorted by the magnitude of

their average correlation. Thus, the relative contribution of each feature to a linear model of subjective mental workload could be assessed.

Additionally, as introduced in Section 2.1.3, offline analysis is conducted to determine the relationship between head tilt and change in oxygenated hemoglobin. This analysis is performed to assess whether head tilt should be accounted for in future fNIRS signal extraction procedures. For each participant and each mission, the linear correlation between the two signals is calculated using Pearson's Correlation Coefficient. An average coefficient and its standard deviation are then reported across all trials.

To appraise the willingness of participants to modify their assessed subjective workload after being presented with a system-generated prediction, the proportion of participants who changed their assessment is provided. Additionally, the magnitude of these changes is evaluated by noting the change in correlation between their subjective assessment and the system-generated prediction before and after making these changes.

Next, to evaluate how study participants would objectively respond to receiving notifications of perceived high mental workload, all scenarios in which a real-time auto-notification is triggered due to a predicted mental workload greater than 0.8 out of 1 are analyzed. This analysis involves extracting the mental workload signal before and after each notification for all scenarios, time-synchronizing the data to the moment of notification, and plotting the resulting data. A grand average curve is also plotted to show the mean response over all scenarios.

Finally, responses to the post-flight questionnaires are evaluated to assess the participant's subjective perceptions of an in-cockpit mental workload prediction system.

4.3 Results

Through the post-flight questionnaire, participant engagement and commitment to the successful completion of each mission were assessed to be sufficiently high to facilitate subsequent mental workload analysis for all participants. To the question "On a scale from 1 to 5 where 1 is "fully disengaged" and 5 is "fully invested," to what extent do you feel you were mentally invested in the successful completion of the first mission?" the average response is 4.6 ± 0.58 . Regarding the second mission, the average response is 4.7 ± 0.46 . Only one participant (Student 3), reported a value less than "4" and this response (of "3") was reported for only the first mission. See Table A.0.3 in Appendix D for the complete list of responses for

all participants. Additionally, through short answer responses, pilot participants provided mixed responses regarding the difference between the mental workload induced in the simulator and during actual flight. Three participants reported experiencing more stress in the simulator, six reported experiencing more stress during actual flight, and one was indifferent. In general, pilot participants expressed that the level of concentration required is similar in both scenarios, but the overall mental workload is less in the simulator than in actual flight (see Question 9 in Appendix D for complete responses).

Initial evaluation of the collected data uncovered five of the forty simulated missions flown for which the system-generated mental workload prediction needed to be re-calculated before computing the correlation between participant subjective workload and the system-generated predicted mental workload. These five corrections are explained below.

- Student 4, Mission 2: Due to an undetermined cause, the real-time mental workload prediction was not logged. The prediction was made offline as if it were being done in real-time including the use of the baseline measurements to initialize scaling and the continual updating of the minimum and maximum scaling values as is performed in the real-time system.
- Student 8, Mission 2: Due to a pilot-induced helicopter crash during Mission 1 causing unforeseen issues, only approximately one-half the data from Mission 1 was used to train the linear regression model applied during Mission 2. To correct for this, the model was re-trained using the complete dataset from Mission 1 and applied during Mission 2 simulating real-time techniques to predict the subject's mental workload.
- Student 10, Mission 2: For an unknown reason, the saccade rate was not calculated for much of Mission 1. The resulting model, fit using data from only the short amount of time during which all features including saccade rate were available, was ineffective. To correct for this error made during real-time processing, saccade rate was extracted using the raw eye tracking data, the prediction model was re-trained, and the new model was applied to Mission 2 as if executed in real-time.
- Pilot 4, Mission 1 and Mission 2: Due to a uniquely noisy and previously-unseen ECG waveform, the ECG-extracted features (HR, HRV, LF/HF, LF/HF_slope) were not extracted properly. Thus, the prediction model trained on Mission 1 and applied to Mission 2 in real-time lacked an accurate representation of these parameters as it was trained on the small portion of data for which these and all other features existed. To correct this error, the ECG-related features were extracted from the logged raw ECG

signal using a manual process to ensure accurate peak extraction, the model was re-trained, and the new model was applied to Mission 2 as if executed in real-time.

These offline corrections were possible due to the meticulous logging of all individual data streams.

After making the necessary corrections and concluding that no participant data would be excluded from use, analysis was conducted to assess the study's multiple research questions beginning with the correlation between subjective and predicted mental workload. Representative plots from two participants showing the participants' subjective and predicted mental workload over the course of a simulated helicopter mission are given in Fig. 4.10. Plotted are the datasets with the largest and smallest Pearson correlation coefficients of all collected (0.69 and -0.31 respectively). The other datasets are well represented by these two examples.

For each study participant and both missions, Pearson Correlation Coefficients were calculated between the system-generated predicted mental workload and each participant's assessed subjective mental workload (both before, and after they had been shown the system-generated prediction). These correlation coefficients are given in Table 4.4. For the five missions for which the system-generated predicted mental workload was corrected post-mission execution, the coefficients for both the real-time and the corrected datasets are provided. Summary statistics for data provided in Table 4.4 are given in Table 4.5.

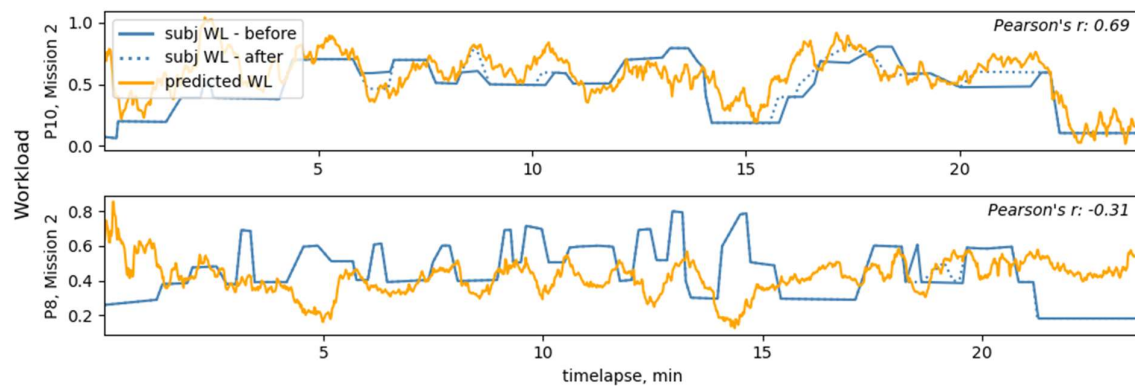


Fig. 4.10 Subjective and predicted mental workload over the course of simulated helicopter missions. The participant's subjective mental workload both before and after seeing the system-generated prediction are plotted. The top plot shows the data yielding the highest Pearson's correlation coefficient (0.69) and the bottom plot shows the data yielding the lowest Pearson's correlation coefficient (-0.31).

Table 4.4 Complete mental workload correlation statistics for each participant

Participant ID	Mission ID	Real-Time/Uncorrected		Corrected	
		Before View Prediction	After View Prediction	Before View Prediction	After View Prediction
S1	M1	0.3601	0.3601		
	M2	0.5742	0.6486		
S2	M1	0.3745	0.4595		
	M2	0.1045	0.1284		
S3	M1	0.6576	0.6576		
	M2	0.6095	0.6095		
S4	M1	0.4007	0.3947		
	M2	-	-	0.5184	0.5184
S5	M1	0.1916	0.2933		
	M2	0.2127	0.4063		
S6	M1	0.5083	0.5013		
	M2	0.1828	0.1828		
S7	M1	0.2722	0.3045		
	M2	0.4422	0.4601		
S8	M1	0.3993	0.4248		
	M2	0.4101	0.4166	0.4755	0.4527
S9	M1	0.3850	0.4808		
	M2	0.3815	0.4787		
S10	M1	0.2359	0.2457		
	M2	-0.1350	-0.1284	0.3855	0.3199
P1	M1	0.3362	0.3362		
	M2	0.1962	0.4024		
P2	M1	0.4935	0.4935		
	M2	0.4544	0.4544		
P3	M1	0.1910	0.2590		
	M2	0.4154	0.4154		
P4	M1	0.5126	0.5126	0.5077	0.5077
	M2	0.2444	0.2470	0.3352	0.2894
P5	M1	0.0699	0.0699		
	M2	0.0492	0.0492		
P6	M1	0.4108	0.4108		
	M2	0.1418	0.1418		
P7	M1	0.4518	0.4677		
	M2	0.5880	0.6054		
P8	M1	-0.1866	-0.1441		
	M2	-0.3291	-0.3125		
P9	M1	0.2667	0.2765		
	M2	0.1271	0.1720		
P10	M1	0.4002	0.4677		
	M2	0.6121	0.6904		

All values represent the Pearson Correlation Coefficient between participant-provided subjective mental workload and system-generated predicted mental workload.

Table 4.5 Complete summary statistics of all mental workload correlation coefficients

		Real-time/Uncorrected			Corrected		
		Mission 1	Mission 2	Both	Mission 1	Mission 2	Both
Before View Prediction	All Participants	0.34±0.18	0.28±0.25	0.31±0.21	0.34±0.18	0.32±0.23	0.33±0.21
	Students	0.38±0.13	0.31±0.23	0.35±0.18	0.38±0.13	0.39±0.16	0.38±0.15
	Pilots	0.29±0.21	0.25±0.27	0.27±0.24	0.29±0.21	0.26±0.27	0.28±0.24
After View Prediction	All Participants	0.36±0.17	0.32±0.26	0.34±0.22	0.36±0.17	0.36±0.24	0.36±0.21
	Students	0.41±0.11	0.36±0.24	0.39±0.18	0.41±0.11	0.42±0.16	0.42±0.14
	Pilots	0.31±0.20	0.29±0.28	0.30±0.24	0.31±0.20	0.29±0.28	0.30±0.24

All values show the population mean and one standard deviation from the mean.

Table 4.6. Focused summary of the correlation analysis between subjective and predicted mental workload. The data in this table is taken from the lower-right quadrant of Table 4.5 representing the correlation between the participant's subjective workload after viewing the real-time prediction and the system-generated prediction (corrected in 5 of the 40 cases).

	Mission 1	Mission 2	Both
All Participants	0.36±0.17	0.36±0.24	0.36±0.21
Students Only	0.41±0.11	0.42±0.16	0.42±0.14
Pilots Only	0.31±0.20	0.29±0.29	0.30±0.24

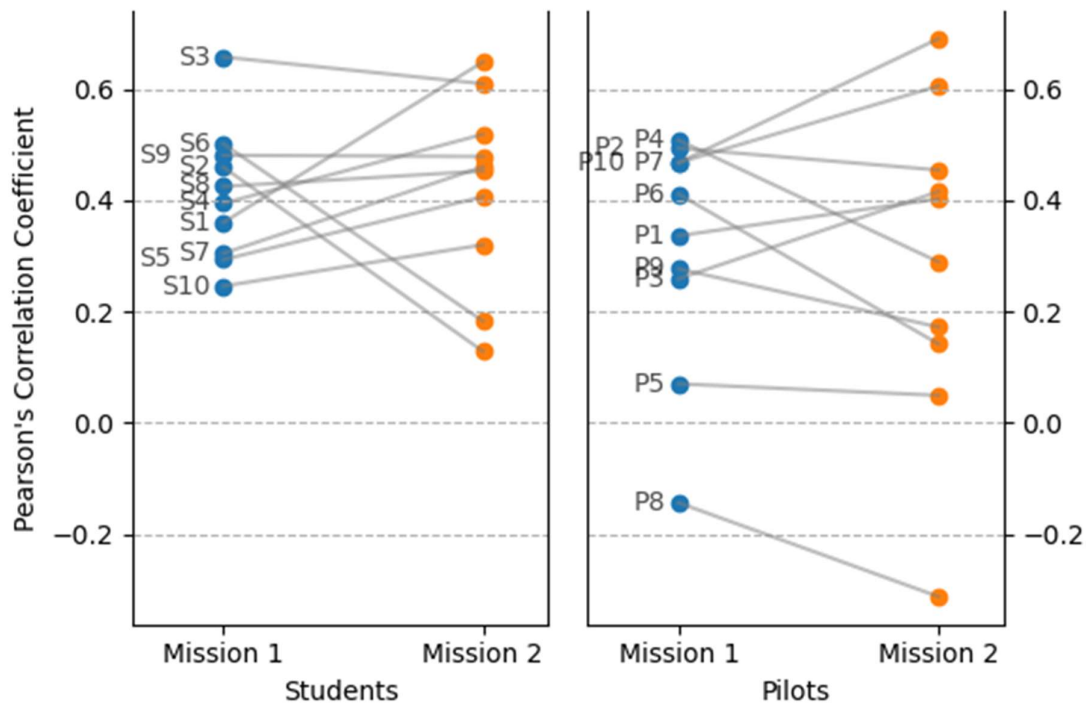


Fig. 4.11 Grouped scatter plots of the correlation between subjective and predicted mental workload. The plots show the increase or decrease in correlation between the first and second simulated missions for student (left) and pilot (right) participants.

A focused summary of the correlation analysis between participant subjective mental workload and predicted mental workload is provided in Table 4.6. Across all participants and missions, the average Pearson correlation coefficient is 0.36 with a standard deviation of 0.21. Correlation coefficients for each participant across both missions (given numerically in Table 4.4) are visualized in Fig. 4.11. This figure enables the visual assessment of how well subjective mental workload was predicted for each participant individually. To determine whether or not there was a statistical difference between Mission 1 and Mission 2 across all participants, paired t-tests were performed. The assumption that the differences between pairs should be approximately normally distributed was verified via a Shapiro-Wilk test for normality which concluded there was insufficient evidence to reject this hypothesis. The paired t-tests yielded p-values greater than 0.05 providing insufficient evidence to reject the null hypothesis of identical averages. Thus, it cannot be concluded that the mean correlation coefficient differed from Mission 1 to Mission 2.

The results of the feature utility analysis are summarized in Table 4.7. It is noted that the feature yielding the largest average correlation coefficient (in magnitude) was σ_{HHb} . Thus, on average, this feature correlated most strongly with the participant's subjective mental workload. It is also noted that this feature's average correlation of -0.41 is greater in magnitude

Table 4.7 Correlation summary statistics between subjective mental workload and individual features. All values are mean Pearson correlation coefficients and one standard deviation from the mean across all participants. The individual features are sorted by the absolute value of the aggregate mean correlation coefficient.

Feature	Aggregate	Mission 1	Mission 2	Paired Δ
<i>Pred. Mental Workload*</i>	0.36±0.21	0.36±0.17	0.35±0.24	-0.01±0.18
σ_{HHb}	-0.41±0.20	-0.43±0.22	-0.39±0.18	0.04±0.19
σ_{O2Hb}	-0.38±0.20	-0.38±0.22	-0.37±0.18	0.00±0.14
HR	0.27±0.24	0.25±0.25	0.29±0.24	0.04±0.24
HRV	-0.23±0.24	-0.24±0.23	-0.22±0.25	0.01±0.20
respRate	0.19±0.17	0.22±0.16	0.15±0.18	-0.06±0.16
gazeBlinkPerMin_60s	-0.18±0.21	-0.26±0.18	-0.11±0.21	0.14±0.24
gazeSaccadePerSec_60s	0.16±0.28	0.13±0.31	0.19±0.24	0.06±0.33
gazeChangePerSec_60s	-0.14±0.27	-0.20±0.26	-0.07±0.27	0.13±0.39
gazeSaccadePerSec_5s	0.08±0.16	0.06±0.17	0.10±0.15	0.04±0.19
EDA	-0.07±0.27	-0.07±0.30	-0.06±0.24	0.01±0.27
EDA_WL_20s	-0.06±0.22	-0.06±0.25	-0.07±0.19	-0.01±0.24
mean_O2Hb	0.06±0.15	0.04±0.16	0.09±0.15	0.04±0.19
gazeChangePerSec_5s	-0.06±0.15	-0.09±0.15	-0.03±0.14	0.07±0.21
LF/HF	-0.04±0.25	-0.01±0.26	-0.07±0.25	-0.05±0.22
<i>Spatial Asymmetry O2Hb</i>	0.03±0.17	0.01±0.15	0.05±0.19	0.03±0.27
LF/HF_slope	0.02±0.10	0.01±0.11	0.02±0.10	0.00±0.14
EDA_slope	-0.01±0.04	-0.02±0.05	-0.01±0.03	0.00±0.04
μ_{HHb}	0.00±0.13	0.03±0.12	-0.02±0.14	-0.04±0.16

* Predicted Mental Workload is not an individual feature but the prediction resulting from the linear regression model fusing all listed features.

than the average correlation coefficient between subjective and predicted mental workload of 0.36. The table also provides the summary statistics for the paired difference between Mission 1 and Mission 2 across participants in the right-most column “Paired Δ .” From this dataset, it is noted that “gazeBlinkPerMin_60s” yielded correlation coefficients of the greatest difference between missions.

A portion of the data presented in Table 4.7 is visualized in Fig. 4.12 with paired box-and-whisker plots of the linear correlation between individual features and subjective mental workload. There, again sorted by the magnitude of the mean correlation coefficient across participants, one can visualize the linear correlation between individual features and subjective mental workload.

Pertaining to the relationship between head tilt and fNIRS introduced in Section 2.1.3, the correlation between head tilt and change in oxygenated hemoglobin across all participants and missions resulted in an average Pearson’s correlation coefficient of -0.08 with a standard deviation of 0.16. Of the 40 missions flown by the 20 participants, the correlation was found to be weakly positive for 12 of the 40 missions. The correlation was found to be weakly positive

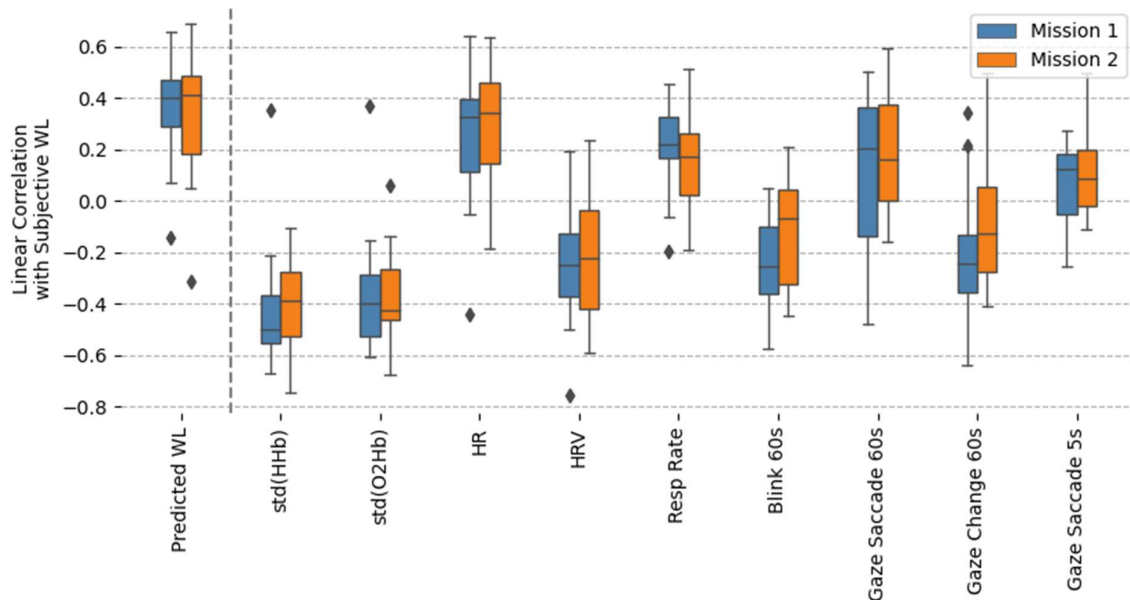


Fig. 4.12 Paired box-and-whisker plots of the linear correlation between individual features and subjective mental workload. Features are sorted by the absolute value of their mean Pearson correlation coefficient. Plotted are the nine features with correlation coefficients of the largest magnitude. Additionally, the correlation between predicted and subjective mental workload is also plotted (far left). Box edges show the quartiles of the dataset while the whiskers extend to show the rest of the distribution (except for points determined to be outliers which are denoted as small diamonds).

for 8 of the 20 participants during at least one of the two missions. It is noted that the average standard deviation of head tilt across all participants is 5.77 degrees.

Regarding the changing of one's subjective mental workload when presented with a system-generated prediction, 16 of the 20 participants (80%) modified their subjective mental workload assessment to more closely align with the system-generated prediction after being presented with the system's predicted mental workload for at least one of the two simulated missions flown (9 of the 10 student participants and 7 of the 10 trained pilots). Across both missions, 26 of the 40 subjective assessments (65%) were modified (15 of the 20 missions flown by the student participants and 11 of the 20 missions flown by the pilots). Shown previously, Fig. 4.10 presents representative plots of these changes. In the upper plot, it can be seen that Participant P10 modified their initial assessment at multiple points to more closely match the predicted value. In the lower plot, it is shown that Participant P8 modified their initial assessment only once (near the 19-minute mark) and only slightly. The data in Table 4.5 quantifies the effect of these changes. Before viewing the system-generated prediction, across all participants and both missions, the average correlation coefficient between predicted and subjective mental workload is 0.33 ± 0.21 . After viewing the system-generated prediction and having an opportunity to update their subjective assessment, the average correlation coefficient increases slightly to 0.36 ± 0.21 .

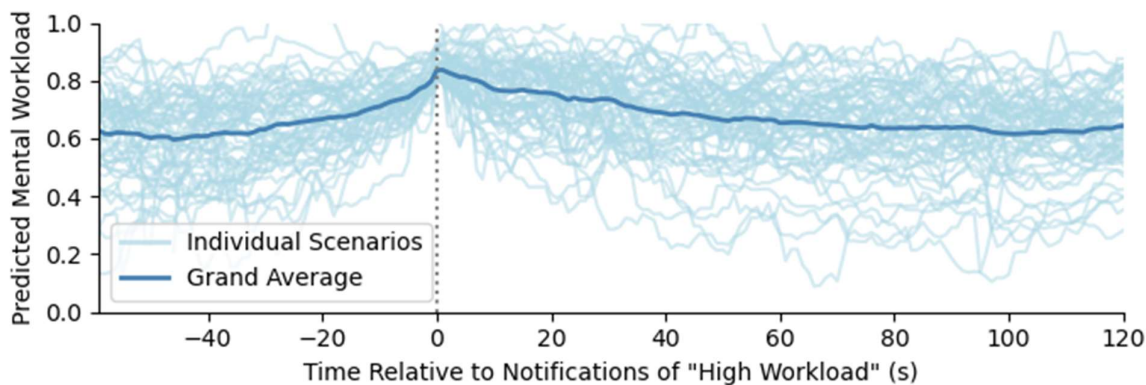


Fig. 4.13 Time-synchronized plots of predicted mental workload centered on notifications of high workload. Time $t = 0$ corresponds to the moment notifications of perceived high workload were provided aurally to study participants. From the grand average plot, it is evident that despite the large variability across scenarios, on average, the notifications have a rapid and significant impact on the temporal dynamics of the predicted mental workload signal. It is hypothesized that study participants applied self-regulatory strategies to regulate the recorded physiological signals used to predict mental workload.

Next, to assess the objective response of participants to aural notifications of perceived high mental workload, all scenarios in which a real-time auto-notification were triggered due to a predicted mental workload greater than 0.8 out of 1 are time-synchronized and plotted in Fig. 4.13. The plot includes 58 scenarios across 8 study participants during which the aforementioned criteria were met and the notifications were delivered. No form of assistance was provided to the participants following notification of high mental workload. The figure illustrates that, on average, the notifications bring about a rapid and significant reduction in the participant's predicted mental workload. It is hypothesized that upon receiving the notifications, the study participants applied self-regulatory strategies which influenced the recorded physiological signals used to predict mental workload.

Finally, from the post-flight questionnaire, it is concluded that the subjective workload analysis tool facilitated the acquisition of responses from the participants. To the question "On a scale from 1 to 5 where 1 is "very easy" and 5 is "very hard," how difficult was it for you to subjectively assess your mental workload after mission completion using the provided video playback tool?" the average response is 2.1 ± 0.62 suggesting it was moderately simple to provide the assessment. The highest rating of "4," from Participant P2, suggests that even with the video playback tool, the process remains difficult for some.

The post-flight questionnaire provided the pilot and student participants an opportunity to share their subjective perceptions and feelings about various aspects of an in-cockpit mental workload prediction system (not necessarily the specific system tested in this experiment). Although not a large population from which to establish engineering or design requirements, the responses provide a considerable range of thoughtful responses and are given in Appendix

D. Responses range from the negative (“such systems will be more of a distraction than a help”) to the positive (“such systems could significantly increase flight safety”), and include multiple suggestions to future system researchers and developers. In general, the surveyed active-duty military pilots are receptive to and welcome the technology.

4.4 Discussion

This work presents the first known study in which fNIRS, ECG, EDA, respiration rate, and eye movement data were used to predict the mental workload level of pilots during simulated flight. Previous studies have used one or more of these sensors for related purposes yet this work demonstrates a furthering of the work by applying a practical multi-modal approach to predicting the pilot’s mental workload in real-time using this set of psychophysiological signals.

Across all participants and both missions, the average Pearson correlation coefficient between the participant’s subjective mental workload and the system’s real-time predicted mental workload is 0.36 with a standard deviation of 0.21. This value suggests that the multi-modal physiological-based mental workload prediction correlates moderately with the participant’s subjectively-assessed mental workload throughout simulated flight.

It was determined that the mean correlation coefficient did not differ statistically significantly from Mission 1 to Mission 2. This finding speaks against the hypothesis that the prediction model tuned to the individual pilot would outperform a model with default weights. Possible explanations for this observation include: 1) differing workload-contributing factors between Mission 1 and Mission 2) the inability of the participants to provide a reliable subjective mental workload assessment.

This first explanation centers around the idea that if mental workload is induced which requires focused visual attention outside the cockpit for example (such as low-level flight through a canyon), and a prediction model is trained on this dataset, it will be “tuned” for this visually-demanding environment. If then a high level of mental workload is induced in another setting not requiring focused visual attention, the trained prediction model would be inaccurate. This, in essence, describes a foundational requirement for supervised machine learning, that the test environment very nearly resembles the training environment. The data presented in Table 4.7 supports this argument. It is noted that the eye-movement-related features (specifically “gazeChangePerSec_60s”) differed more between Mission 1 and Mission 2 than the other features. Early work on pilot mental workload prediction showed that eye-related

features were sensitive to visual workload [35]. The finding that these eye-movement-related features differed between missions suggests that the visual demands were different between the two missions.

The second possible explanation for the observed result points to another foundational requirement for supervised machine learning – that the “truth” is properly labeled. In this case, the model is trained on a subjective evaluation of mental workload which, although great care was taken to assess as accurately as possible, is imperfect. It is likely the reported assessment is influenced by the participant’s perceived performance and not an unbiased reflection of their actual mental workload experienced during flight. This influence has been suggested and reported by others in previously-published works [36]. Additionally, after observing significant deviations between physiological responses during flight and participant assessment of mental workload, support is given to the conclusion presented in [56] that it is not possible to evaluate a pilot’s mental workload by subjective measures alone. Thus, we cannot assume that the subjective assessment is the “truth” we are seeking to predict. The difficulties of training a model on the entire flight-experience envelope and of obtaining properly-labeled “truth” highlight the challenge associated with this approach.

Another interesting aspect of these results is the fact that the predicted mental workload signal yields an average correlation coefficient of less magnitude than the individual features σ_{HHb} and σ_{O_2Hb} (see Table 4.7). This is an unexpected result. It was hypothesized that a signal incorporating a variety of physiological inputs would correlate more strongly with subjective mental workload than any one signal alone. It may thus be suggested that a system aimed at predicting subjective mental workload should utilize these features alone and set aside the others. Furthermore, it could be suggested that σ_{O_2Hb} could be used over σ_{HHb} as a single indicator of subjective mental workload due to the minimal difference in mean correlation between missions (mean paired difference of 0.00 ± 0.14). This conclusion may be misleading, however. First of all, as discussed previously, the subjective mental workload assessment provided by the participants is very likely not the absolute “truth” being sought. Rather, it is likely that variations in the other signals contain information predictive of mental workload of which the participants are less cognizant. Additionally, it may be found that the fused signal is more robust to variations in mental workload conditions as is suggested by the data reported in Table 4.7 where it is noted that the difference in mean correlation between missions is very small.

The finding that the two fNIRS features (σ_{HHb} and σ_{O2Hb}) correlated more strongly with participant subjective mental workload than any of the other features is significant. The significance of these features has not been previously reported. This is an exciting finding. Existing literature suggests that μ_{HHb} and μ_{O2Hb} are sensitive to mental workload, yet the correlation results summarized in Table 4.7 suggest these particular features correlate quite poorly with subjective mental workload. Indeed, μ_{HHb} is determined to be the feature correlating least strongly of all eighteen features analyzed. On the other hand, the instantaneous variance of oxygenated and deoxygenated hemoglobin concentration changes across channels are most strongly correlated with subjective mental workload of all features analyzed. Although the acquisition and extraction of these fNIRS features is more challenging than others such as heart rate, there is reason to believe it may be possible to incorporate an fNIRS acquisition system into a pilot's helmet enabling the use of these signals in actual flight. Between EEG and fNIRS cortical-activity-monitoring technologies, fNIRS is more likely to be realized operationally due to its ease of setup and its robustness against motion artifacts. Although preliminary testing showed a relationship between head tilt and fNIRS signals (see Section 2.1.3), data from this study showed only a weak correlation between head tilt and change in oxygenated hemoglobin in the pre-frontal cortex suggesting the influence of head tilt on recorded fNIRS signals in this environment is not a significant confounding factor. This is likely because approximately 95% of the time, participants' heads were tilted within the small range of only approximately 20 degrees (four times the standard deviation of 5.77 degrees). Certainly, challenges remain related to the acquisition of fNIRS signals in real-flight conditions which have not been addressed in this work such as variations in ultraviolet light (UV) exposure and movement-induced g-forces.

Additionally, the finding that 80% of participants changed their subjective assessment of workload after having been presented with a system-generated prediction provides evidence that such subjective assessments can be aided by the use of objective physiological data. Reviewing one's physiological data supports recall when evaluating one's mental workload. It was poignant that one of the participants would state in their post-flight questionnaire regarding the post-flight workload analysis “[the system] seemed to capture my mental state better than I could...”

Finally, visualizing the rapid effect notifications of high workload have on a prediction of mental workload is exciting. As seen in Fig. 4.13, across many instances in which such notifications were given, the average response shows a marked response within seconds of

notification delivery. It is noteworthy that the near immediate interruption of the increasing predicted mental workload and its subsequent decline following notification is not the result of task elimination or simplification assistance. Rather, the effect is attributed to the participant's self-regulation of their physiological state. This self-regulation may be just what is needed to ensure a pilot remains in a productive and effective mental state while flying. The generation of this compelling illustration was enabled by the novel real-time continuous mental workload prediction system and online and transparent triggering system presented in this work.

5 Conclusion

As the aircraft cockpit tends towards single-pilot operations, the monitoring of pilot mental state must be augmented by tools designed for this purpose. This work has presented an approach to estimating pilot mental workload in real-time through an analysis and aggregation of various physiological and behavioral signals through a linear model. The estimated, or predicted, mental workload signal correlated moderately with the subjective evaluation provided by the student and active-duty pilot participants. Correlation between predicted and subjective mental workload averaged 0.36 ± 0.21 across all participants with the strongest correlation being 0.69.

Due to the lack of precision with the measurement, it is suggested that the metric be used as a broad indicator, rather than a sole driver of a multi-level adaptive automation system. It may likely be effectively used as a trigger to notify a crew member, supporting ground-station personnel, or even the pilot themselves of potentially undesirable states. Although not rigorously tested in this experiment, notifications of high mental workload were provided on multiple occasions and it was noted that these notifications resulted in the near-immediate application of self-adaptive strategies which yielded a decrease in predicted mental workload. These self-adaptive strategies included weight shifting, deep breathing, and verbalization of the situation. Future experimentation may determine these observations were not circumstantial and that such notifications aid the pilot in maintaining a safe and productive mental workload level.

Additionally, it is offered that the moderate correlation between predicted and subjective mental workload is not as insignificant as it may appear. As mentioned previously, the subjective mental workload provided by the participants may not be the full “truth” we aim to predict. Rather, the fusion of various physiological signals shown to respond to stress and mental workload may generate a metric even more sensitive to experienced mental workload than the participant can effectively assess. This argument is strengthened noting that of the 20 participants, 16 modified their subjective mental workload assessment after having been

presented with the predicted value without any significant encouragement to do so. When comparing their subjective assessment with the newly-displayed predicted value, many would express ideas such as “Yes, I suppose that was more difficult during that time than I previously noted” or “Yes, I did start feeling anxious about the situation earlier than I reported.” These edits to the participants’ subjective mental workload support the idea that the physiological-based mental workload metric may, at times, be more representative of a person’s mental workload than they can express themselves.

5.1 Summary of Contributions

This work has furthered both the theoretical and scientific basis as well as the practical application of physiological monitoring in the cockpit to assess pilot mental workload in real-time.

5.1.1 Theoretical and Scientific Contributions

First, a theoretical basis for monitoring physiological measures of pilots was established and supported by a systematic review and summary of previously-published works. This foundational work established a relationship between task difficulty, performance, and mental workload. It supported the conclusion that mental workload is experienced uniquely by an individual and that it cannot be deduced through an analysis of the task load alone. Ultimately, this theoretical work concluded that physiological monitoring may provide an important input source to a human-machine system aimed at optimizing performance. The summary of previously-published works illustrated the potential utility of various physiological signals in the pursuit of this goal.

Next, a unique combination of physiological signals was selected and utilized for real-time mental workload prediction. The selected signals supported a “full-body assessment” by monitoring many of the human body’s physiological sub-systems. The central nervous system was probed through functional near-infrared spectroscopy (fNIRS). The activity of the sympathetic nervous system was observed through the collection of electrodermal activity (EDA). The respiratory system was monitored through chest and stomach stretch sensors. Data pertinent to the cardiovascular system was collected through 3-lead electrocardiography (ECG). Finally, multiple features related to eye movement were also collected. Through the utilization of these many signals and extracted features, a prediction of mental workload was generated correlating with the participants’ subjective mental workload with an average Pearson’s correlation coefficient of 0.36 across all 20 participants.

An assessment was conducted of the 18 features utilized in the real-time mental workload prediction system. It was shown that features representing the instantaneous variance of the signals measuring oxygenated and deoxygenated hemoglobin in the outer surface of the prefrontal cortex had the strongest linear correlation with participant subjective mental workload. Previously-published works have not presented this fNIRS-extracted feature as being sensitive to mental workload. Although this finding is significant, it is also argued that the other signals which did not have as strong a linear correlation with subjective mental workload may nonetheless be sensitive to or predictive of mental states not captured by one's subjective assessment of mental workload.

This work presented a novel approach for evaluating subjective mental workload and obtaining a continuous metric of its value over the course of a defined period. Rather than through mid-task questioning, post-task questionnaires, or other methods used commonly in the field, this work utilized a post-task immersion to enable the continuous assessment of mental workload over the duration of the task. This immersion was supported by a video and audio playback of the mission (including all displays in and out of the cockpit) as well as a presentation of the gaze location of the participant. This continuous-valued metric could then be used in conjunction with the other continuous-valued physiological signals in various machine learning and statistical applications. The continuous-valued metric, in conjunction with the transparent triggering system (built into PhysHub), also enabled the novel analysis of mental workload following notifications of high workload. The resulting figure (Fig. 4.13) and the clarity by which it illustrates the potential benefit of real-time workload notifications in the cockpit is original in the field.

Finally, the results presented in this work highlight the challenge associated with generating a predictive model from one flight scenario and applying it in a second flight scenario when the flight conditions are not identical between flights. The differences in high-workload-producing tasks between the first and second experimental missions likely contributed to the result that the correlation between subjective and predicted mental workload on the second mission was not as strong as expected. It is thus noted that future work ensures all sources of mental workload are considered in designing a training mission on which a pilot-adapted model is trained.

5.1.2 Practical Application Contributions

This work furthered the practical implementation of a real-time pilot physiological monitoring system in multiple ways. The work demonstrated a working system integrated into a research helicopter simulator capable of real-time physiological signal processing and visualization. The developed system incorporates the simultaneous processing and exploitation of more physiological signals than any previously documented system. The system's code was written in multiple programming languages including C++, Python, and other shell scripting languages. The system demonstrated an edge-computing paradigm where low latency, high-frequency data processing was conducted at the nodes and compressed results were transmitted to a centralized monitoring and processing center ("PhysHub") at low frequencies.

One specific element of the implemented system of particular novelty is the prototype in-cockpit display and pilot interface. Through this display, pilots could gain insights into their physiological state and that of their co-pilot. These systems (the humans) are arguably the most important systems within the human-machine team and until now, the assessment of these systems in the cockpit of a helicopter was unsupported by technological means. It is strongly suggested that space be allocated to the presentation of this important information within cockpits. The prototyped display also provides the pilots with an interface to the mental workload notification system. Through the display, the pilots can set the mental workload threshold and frequency at which they wish to receive notifications both for themselves and for their co-pilot. This feature allows for the personal customization of the human-machine team by the pilots in real-time.

Another practical contribution made by this work is the design and implementation of the centralized monitoring and processing center, "PhysHub." This tool enables a rapid assessment of the system including its many sub-systems. It also offers a novel approach for implementing a transparent and customizable triggering system capable of reacting to various measured states. Triggers can be easily imported, exported, and modified with unique thresholds for each pilot or mission. For example, a collision warning trigger can be easily modified by the pilot based on their preferred risk tolerance. Additionally, the tool supports the live activation or deactivation of individual triggers.

Together, the centralized processing center "PhysHub," and the in-cockpit system interface provide transparency for both the pilots and the experimenter. For an effective human-machine team, transparency into system states, modes, and settings is critically important. I

believe any pilot adaptive-assistant system designed without the transparency of these tools will not be accepted or found useful to the pilots for whom they are intended to assist.

5.2 Future Work

The tools and findings presented in this work could be built upon in multiple avenues to further advance the field.

5.2.1 Longitudinal Studies of Pilot Physiological Data and Subjective Mental Workload

Pilot physiology as recorded through the various means presented in this work as well as their subjective assessment of mental workload should be assessed over multiple simulated flights spanning multiple days or weeks. Rather than observing each pilot during only two missions, it would be informative to assess these metrics over the course of many (e.g., 10) missions spanning multiple weeks. Longitudinal studies of this sort would provide critical insights into the utility of a system over these longer timescales. It is hypothesized that taking into account the pilot's baseline physiology (as was done in this work) is very important. It may, however, also be found that taking into account the pilot's baseline state alone is insufficient for informing the prediction model and that some method of online re-training of the prediction model is required. Of particular value would be the assessment of pilot subjective mental workload over these missions. It would be informative to see how participants assessed their mental workload on identical (or nearly identical) missions flown days apart. Additionally, analyzing the stability of individual physiological features over these many missions would help identify features of most utility in this setting.

5.2.2 Further Exploration of fNIRS Features

This work found two previously unreported fNIRS features more strongly correlated with participant subjective mental workload than any of the other more traditional features such as heart rate or respiration rate. Specifically, it was found that the instantaneous variability of all oxygenated and deoxygenated hemoglobin channels declined as subjective mental workload increased. Because this finding is unique among the published literature, further exploration of these features should be conducted to validate or contradict this finding. Additionally, experimentation could be conducted to determine the task conditions which most significantly affect this feature. It could be determined that the feature is more sensitive to tasks of one type than another.

5.2.3 Suitability Studies of fNIRS in Real Flight Conditions

Existing published literature suggests it is possible to collect and extract features from fNIRS in real flight conditions [25], [26]. Due to the significant interference observed between the eye-tracking and fNIRS systems however (documented in section 2.1.1), substantial challenges are anticipated in processing the fNIRS signal given the impact of g-forces, head tilt, and changing light conditions (specifically infrared light) experienced during actual flight. If it is found that the extraction of robust workload-relevant features from fNIRS in real flight conditions is not feasible, effort could be directed toward the utilization of the signals in stationary systems such as training simulators or remotely-piloted aircraft control systems.

5.2.4 Integration into an Assistant System and Acceptability Studies

Finally, this work could be advanced by applying the developed mental workload prediction algorithm to a pilot assistant system based on transparent and editable triggering. The triggering system described in sections 3.2 and 3.4 could be further developed to enable the pilot full control of all triggers while in the cockpit. Additionally, the eye-related features could be removed as inputs to the overall mental workload prediction. Instead, these features could be used as inputs to a “visual capacity” metric with other applications. As discussed in section 4.4, these features are highly situation dependent and may be more appropriately applied to assessing the visual modality of workload (when looked at through the lens of Multiple Resource Theory [6]).

In addition to other potential triggers built into the system, notifications of high mental workload should be explicitly evaluated for their utility and acceptability. Situations should be simulated which lead to the triggering of these notifications and quantitative and qualitative assessments should be conducted to evaluate their impact on the participant and their performance. The physiological data of the participants should be assessed for signs of self-adaptive strategies to lower their physiological strain. Task performance should be assessed throughout the increase in predicted mental workload and moments following notification of high mental workload. Additionally, questionnaires should be utilized to evaluate the subjective acceptability of the notifications.

Another aspect of the triggering system which could be studied is the appropriate recipient of high mental workload notifications and how they should be delivered. It may be determined that a co-pilot or ground-station operator is a more appropriate recipient of this

information than the pilot experiencing the high mental workload. These individuals could then adapt to the situation and communicate with the pilot if necessary and appropriate.

Ultimately, the system will need to be intuitive, reliable, and valuable to the human pilots for whom it is intended to support. As with all human-machine systems, assessing and achieving this state will require ergonomic and usability studies.

Bibliography

- [1] M. Cooley, "Human Centered Systems: An Urgent Problem for Systems Designers," *AI Soc.*, vol. 1, pp. 37–46, 1987.
- [2] C. E. Billings, "Human-Centered Aircraft Automation: A Concept and Guidelines," *NASA Tech. Memo.*, no. 103885, 1991.
- [3] C. E. Billings, "Human-Centered Aviation Automation: Principles and Guidelines," *NASA Tech. Memo.*, no. 110381, 1996.
- [4] P. M. Fitts, "Human Engineering for an Effective Air-Navigation and Traffic-Control System," Ohio State University Research Foundation Report, Columbus, OH, 1951.
- [5] A. T. Welford, *Skilled performance: Perceptual and motor skills*. Glenview, Ill: Scott & Foresman, 1976.
- [6] C. D. Wickens, "Multiple Resources and Mental Workload," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 50, no. 3, pp. 449–455, Jun. 2008, doi: 10.1518/001872008X288394.
- [7] R. K. Dismukes, T. E. Goldsmith, and J. A. Kochan, "Effects of Acute Stress on Aircrew Performance: Literature Review and Analysis of Operational Aspects," *NASA Tech. Memo.*, 2015, doi: 10.13140/RG.2.1.2898.3449.
- [8] A. Schulte, D. Donath, and D. S. Lange, "Design Patterns for Human-Cognitive Agent Teaming," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9736, pp. 231–243, 2011, doi: 10.1007/978-3-319-40030-3_24.
- [9] R. Onken and A. Schulte, *System-ergonomic Design of Cognitive Automation: Dual-Mode Cognitive Design of Vehicle Guidance and Control Work Systems*. Heidelberg: Springer Berlin, 2010.
- [10] B. P. Bailey, J. A. Konstan, and J. A. Konstan, "On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state," *Comput. Human Behav.*, vol. 22, pp. 685–708, 2006, doi: 10.1016/j.chb.2005.12.009.
- [11] N. Peters, G. Romigh, G. Bradley, and B. Raj, "When to Interrupt: A Comparative Analysis of Interruption Timings Within Collaborative Communication Tasks," in *Advances in Intelligent Systems and Computing*, 2017, vol. 497, pp. 177–187, doi: 10.1007/978-3-319-41956-5_17.
- [12] H. Kharoufah, J. Murray, G. Baxter, and G. Wild, "A review of human factors causations in commercial air transport accidents and incidents: From to 2000–2016," *Prog. Aerosp. Sci.*, vol. 99, pp. 1–13, May 2018, doi: 10.1016/J.PAEROSCI.2018.03.002.
- [13] A. Schell and M. E. Dawson, "Psychophysiology," *Int. Encycl. Soc. Behav. Sci.*, pp. 12448–12452, Jan. 2001, doi: 10.1016/B0-08-043076-7/03424-0.
- [14] S. H. Fairclough, "Fundamentals of physiological computing," *Interact. Comput.*, vol. 21, no. 1–2, pp. 133–145, Jan. 2009, doi: 10.1016/J.INTCOM.2008.10.011.
- [15] R. M. Yerkes and J. D. Dodson, "The Relation of Strength of Stimulus to Rapidity of Habit-Formation," *J. Comp. Neurol. Psychol.*, vol. 18, no. 5, pp. 459–482, Nov. 1908, doi: 10.1002/CNE.920180503.

-
- [16] D. M. Diamond, A. M. Campbell, C. R. Park, J. Halonen, and P. R. Zoladz, "The Temporal Dynamics Model of Emotional Memory Processing: A Synthesis on the Neurobiological Basis of Stress-Induced Amnesia, Flashbulb and Traumatic Memories, and the Yerkes-Dodson Law," *Neural Plast.*, vol. 2007, p. 33, 2007, doi: 10.1155/2007/60803.
- [17] W. Rohmert and J. Rutenfranz, *Arbeitswissenschaftliche Beurteilung der Belastung und Beanspruchung an unterschiedlichen industriellen Arbeitsplätzen*. Bundesminister für Arbeit und Sozialordnung, 1975.
- [18] G. Durantin, J. F. Gagnon, S. Tremblay, and F. Dehais, "Using near infrared spectroscopy and heart rate variability to detect mental overload," *Behav. Brain Res.*, vol. 259, pp. 16–23, 2014, doi: 10.1016/j.bbr.2013.10.042.
- [19] F. Honecker, Y. Brand, and A. Schulte, "A Task-centered Approach for Workload-adaptive Pilot Associate Systems," *Proc. 32nd Conf. Eur. Assoc. Aviat. Psychol. Cascais, Port.*, pp. 485–507, 2016.
- [20] F. Honecker and A. Schulte, "Automated Online Determination of Pilot Activity Under Uncertainty by Using Evidential Reasoning," *Eng. Psychol. Cogn. Ergon. Cogn. Des.*, vol. 10276, pp. 231–250, 2017, doi: 10.1007/978-3-319-58475-1.
- [21] F. Honecker and A. Schulte, "Full-Mission Human-in-the-Loop Experiments to Evaluate an Automatic Activity Determination System for Adaptive Automation," *Adv. Intell. Syst. Comput.*, vol. 903, pp. 731–737, 2019, doi: 10.1007/978-3-030-11051-2_111.
- [22] D. Mund, E. Pavlidis, M. Masters, and A. Schulte, "A Conceptual Augmentation of a Pilot Assistant System with Physiological Measures," *Proc. 3rd Int. Conf. Intell. Hum. Syst. Integr.*, pp. 959–965, 2020, doi: 10.1007/978-3-030-39512-4_146.
- [23] J. A. Veltman and C. Jansen, "The Role of Operator State Assessment in Adaptive Automation," *TNO Defence, Secur. Saf.*, vol. TNO-DV3, no. A245, 2006.
- [24] A. H. Roscoe and B. S. Grieve, "Assessment of Pilot Workload During Boeing 767 Normal and Abnormal Operating Conditions," *SAE Tech. Pap. 881382*, vol. 97, pp. 968–972, 1988, doi: 10.4271/881382.
- [25] F. Dehais *et al.*, "Monitoring Pilot's Cognitive Fatigue with Engagement Features in Simulated and Actual Flight Conditions Using an Hybrid fNIRS-EEG Passive BCI," *IEEE Int. Conf. Syst. Man, Cybern.*, pp. 544–549, Oct. 2018, doi: 10.1109/SMC.2018.00102.
- [26] T. Gateau, H. Ayaz, and F. Dehais, "In silico vs. Over the Clouds: On-the-Fly Mental State Estimation of Aircraft Pilots, Using a Functional Near Infrared Spectroscopy Based Passive-BCI," *Front. Hum. Neurosci.*, vol. 12, p. 187, May 2018, doi: 10.3389/fnhum.2018.00187.
- [27] K. J. Verdière, R. N. Roy, and F. Dehais, "Detecting Pilot's Engagement Using fNIRS Connectivity Features in an Automated vs. Manual Landing Scenario," *Front. Hum. Neurosci.*, vol. 12, p. 6, Jan. 2018, doi: 10.3389/fnhum.2018.00006.
- [28] F. Dehais *et al.*, "Monitoring Pilot's Mental Workload Using ERPs and Spectral Power with a Six-Dry-Electrode EEG System in Real Flight Conditions," *Sensors*, vol. 19, no. 6, p. 1324, Mar. 2019, doi: 10.3390/s19061324.
- [29] J. A. Veltman, "A Comparative Study of Psychophysiological Reactions During Simulator and Real Flight," *Int. J. Aviat. Psychol.*, vol. 12, no. 1 SPEC, pp. 33–48, 2002, doi: 10.1207/s15327108ijap1201_4.
- [30] M. Causse, Z. Chua, V. Peysakhovich, N. Del Campo, and N. Matton, "Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS," *Sci. Rep.*, vol. 7, no. 1, p. 5222, Dec. 2017, doi: 10.1038/s41598-017-05378-x.
- [31] S. Scannella, V. Peysakhovich, F. Ehrig, E. Lepron, and F. Dehais, "Assessment of

- Ocular and Physiological Metrics to Discriminate Flight Phases in Real Light Aircraft.,” *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 60, no. 7, pp. 922–935, Nov. 2018, doi: 10.1177/0018720818787135.
- [32] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research,” *Adv. Psychol.*, vol. 52, no. C, pp. 139–183, Jan. 1988, doi: 10.1016/S0166-4115(08)62386-9.
- [33] Air Force Research Lab, “Integrated Cockpit Sensing (ICS) Program Overview,” 2022.
- [34] “National Commission on Military Aviation Safety: Report to the President and the Congress of the United States,” 2020.
- [35] T. C. Hankins and G. F. Wilson, “A Comparison of Heart Rate, Eye Activity, EEG and Subjective Measures of Pilot Mental Workload During Flight,” *Aviat. Sp. Environ. Med.*, vol. 69, no. 4, pp. 360–367, 1998, doi: 10.1207/s15327752jpa8502.
- [36] S. Miyake, “Multivariate workload evaluation combining physiological and subjective measures,” *Int. J. Psychophysiol.*, vol. 40, no. 3, pp. 233–238, Apr. 2001, doi: 10.1016/S0167-8760(00)00191-4.
- [37] G. F. Wilson, “An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures,” *International Journal of Aviation Psychology*, vol. 12, no. 1 SPEC. Lawrence Erlbaum Associates, Inc., pp. 3–18, 2002, doi: 10.1207/s15327108ijap1201_2.
- [38] Y. H. Lee and B. S. Liu, “Inflight Workload Assessment: Comparison of Subjective and Physiological Measurements,” *Aviat. Sp. Environ. Med.*, vol. 74, no. 10, pp. 1078–1084, Oct. 2003.
- [39] P. Nickel and F. Nachreiner, “Sensitivity and Diagnosticity of the 0.1-Hz Component of Heart Rate Variability as an Indicator of Mental Workload,” in *Human Factors*, Dec. 2003, vol. 45, no. 4, pp. 575–590, doi: 10.1518/hfes.45.4.575.27094.
- [40] F. Di Nocera, M. Camilli, and M. Terenzi, “A Random Glance at the Flight Deck: Pilots’ Scanning Strategies and the Real-Time Assessment of Mental Workload,” *J. Cogn. Eng. Decis. Mak.*, vol. 1, no. 3, pp. 271–285, Sep. 2007, doi: 10.1518/155534307X255627.
- [41] F. Dehais, M. Causse, and J. Pastor, “Embedded eye tracker in a real aircraft: new perspectives on pilot/aircraft interaction monitoring,” in *3rd International Conference on Research in Air Transportation*, 2008.
- [42] A. Kikukawa, A. Kobayashi, and Y. Miyamoto, “Monitoring of pre-frontal oxygen status in helicopter pilots using near-infrared spectrophotometers,” *Dyn. Med.*, vol. 7, no. 1, 2008, doi: 10.1186/1476-5918-7-10.
- [43] A. Girouard *et al.*, “Distinguishing Difficulty Levels with Non-invasive Brain Activity Measurements,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5726 LNCS, no. PART 1, pp. 440–452, doi: 10.1007/978-3-642-03655-2_50.
- [44] L. Luigi, D. Stasi, J. R. Helmert, and J. J. Cañas, “Saccadic Peak Velocity Sensitivity to Variations in Mental Workload Designing for Life: A Human Perspective on Technology Development View project,” *Aviat. Space. Environ. Med.*, vol. 81, no. 4, Apr. 2010, doi: 10.3357/ASEM.2579.2010.
- [45] S. D. Power, T. H. Falk, and T. Chau, “Classification of prefrontal activity due to mental arithmetic and music imagery using hidden Markov models and frequency domain near-infrared spectroscopy,” *J. Neural Eng.*, vol. 7, no. 2, 2010, doi: 10.1088/1741-2560/7/2/026002.
- [46] N. Dahlstrom, S. Nahlinder, G. F. Wilson, and E. Svensson, “Recording of Psychophysiological Data During Aerobic Training,” *Int. J. Aviat. Psychol.*, vol. 21, no. 2, pp. 105–122, Mar. 2011, doi: 10.1080/10508414.2011.556443.
- [47] S. Tokuda, G. Obinata, E. Palmer, and A. Chaparro, “Estimation of Mental Workload

- Using Saccadic Eye Movements in a Free-Viewing Task,” in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2011, pp. 4523–4529, doi: 10.1109/IEMBS.2011.6091121.
- [48] C. Herff, D. Heger, O. Fortmann, J. Hennrich, F. Putze, and T. Schultz, “Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS,” *Front. Hum. Neurosci.*, vol. 7, p. 935, Jan. 2014, doi: 10.3389/fnhum.2013.00935.
- [49] G. Derosi re, S. Dalhoumi, S. Perrey, G. Dray, and T. Ward, “Towards a Near Infrared Spectroscopy-Based Estimation of Operator Attentional State,” *PLoS One*, vol. 9, no. 3, p. e92045, Mar. 2014, doi: 10.1371/journal.pone.0092045.
- [50] T. Gateau, G. Durantin, F. Lancelot, S. Scannella, and F. Dehais, “Real-Time State Estimation in a Flight Simulator Using fNIRS,” *PLoS One*, vol. 10, no. 3, p. e0121279, Mar. 2015, doi: 10.1371/journal.pone.0121279.
- [51] F. Dehais, V. Peysakhovich, S. Scannella, J. Fongue, and T. Gateau, “Automation Surprise in Aviation: Real-Time Solutions,” in *Conference on Human Factors in Computing Systems - Proceedings*, Apr. 2015, pp. 2525–2534, doi: 10.1145/2702123.2702521.
- [52] M. Causse, V. Peysakhovich, and E. F. Fabre, “High Working Memory Load Impairs Language Processing During a Simulated Piloting Task: An ERP and Pupillometry Study,” *Front. Hum. Neurosci.*, vol. 10, May 2016, doi: 10.3389/fnhum.2016.00240.
- [53] H. Mansikka, K. Virtanen, D. Harris, and P. Simola, “Fighter pilots’ heart rate, heart rate variation and performance during an instrument flight rules proficiency test,” *Appl. Ergon.*, vol. 56, no. April, pp. 213–219, 2016, doi: 10.1016/j.apergo.2016.04.006.
- [54] H. Aghajani, M. Garbey, and A. Omurtag, “Measuring Mental Workload with EEG+fNIRS,” *Front. Hum. Neurosci.*, vol. 11, p. 359, Jul. 2017, doi: 10.3389/fnhum.2017.00359.
- [55] A. R. Hidalgo-Mu noz, D. Mouratille, N. Matton, M. Causse, Y. Rouillard, and R. El-Yagoubi, “Cardiovascular Correlates of Emotional Ddate, Cognitive Workload and Time-on-Task Effect During a Realistic Flight Simulation,” *Int. J. Psychophysiol.*, vol. 128, pp. 62–69, Jun. 2018, doi: 10.1016/j.ijpsycho.2018.04.002.
- [56] A. Alaimo, A. Esposito, C. Orlando, and A. Simoncini, “Aircraft Pilots Workload Analysis: Heart Rate Variability Objective Measures and NASA-Task Load Index Subjective Evaluation,” *Aerospace*, vol. 7, no. 9, Sep. 2020, doi: 10.3390/AEROSPACE7090137.
- [57] P. A. Hebbar, K. Bhattacharya, G. Prabhakar, A. A. Pashilkar, and P. Biswas, “Correlation Between Physiological and Performance-Based Metrics to Estimate Pilots’ Cognitive Workload,” *Front. Psychol.*, vol. 12, p. 954, Apr. 2021, doi: 10.3389/FPSYG.2021.555446/BIBTEX.
- [58] R. L. Charles and J. Nixon, “Measuring mental workload using physiological measures: A systematic review,” *Appl. Ergon.*, vol. 74, pp. 221–232, Jan. 2019, doi: 10.1016/J.APERGO.2018.08.028.
- [59] A. R. Harrivel, A. G. Hylton, and T. A. Hearn, “Best Practices for the Application of Functional Near Infrared Spectroscopy to Operator State Sensing,” *NASA Tech. Memo.*, no. 217615, 2012.
- [60] G. Borghini, L. Astolfi, G. Vecchiato, D. Mattia, and F. Babiloni, “Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness,” *Neurosci. Biobehav. Rev.*, vol. 44, pp. 58–75, Jul. 2014, doi: 10.1016/j.neubiorev.2012.10.003.
- [61] M. Masters and A. Schulte, “Investigating the Utility of fNIRS to Assess Mental Workload in a Simulated Helicopter Environment,” *Proc. 2020 IEEE Int. Conf. Human-Machine Syst.*, Sep. 2020, doi: 10.1109/ICHMS49158.2020.9209549.

-
- [62] M. Strait and M. Scheutz, "What we can and cannot (yet) do with functional near infrared spectroscopy.," *Front. Neurosci.*, vol. 8, p. 117, 2014, doi: 10.3389/fnins.2014.00117.
- [63] S. Lloyd-Fox, A. Blasi, and C. E. Elwell, "Illuminating the developing brain: The past, present and future of functional near infrared spectroscopy," *Neuroscience and Biobehavioral Reviews*, vol. 34, no. 3. pp. 269–284, Feb. 2010, doi: 10.1016/j.neubiorev.2009.07.008.
- [64] Y. Hoshi, "Towards the next generation of near-infrared spectroscopy," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 369, no. 1955. Royal Society, pp. 4425–4439, Nov. 28, 2011, doi: 10.1098/rsta.2011.0262.
- [65] F. Scholkmann *et al.*, "A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology," *NeuroImage*, vol. 85. pp. 6–27, Jan. 15, 2014, doi: 10.1016/j.neuroimage.2013.05.004.
- [66] L. Kocsis, P. Herman, and A. Eke, "The modified Beer–Lambert law revisited," *Phys. Med. Biol.*, vol. 51, no. 5, p. N91, Feb. 2006, doi: 10.1088/0031-9155/51/5/N02.
- [67] F. Scholkmann and M. Wolf, "General equation for the differential pathlength factor of the frontal human head depending on wavelength and age," *J. Biomed. Opt.*, vol. 18, no. 10, p. 105004, Oct. 2013, doi: 10.1117/1.JBO.18.10.105004.
- [68] G. Aranyi, M. Cavazza, and F. Charles, "Using fNIRS for Prefrontal-Asymmetry Neurofeedback: Methods and Challenges," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9359, pp. 7–20, doi: 10.1007/978-3-319-24917-9_2.
- [69] M. A. Just and P. A. Carpenter, "A Theory of Reading: From Eye Fixations to Comprehension," *Psychol. Rev.*, vol. 87, no. 4, Jul. 1980.
- [70] Y. Rai and P. Le Callet, "Visual attention, visual salience, and perceived interest in multimedia applications," *Acad. Press Libr. Signal Process. Image Video Process. Anal. Comput. Vis.*, vol. 6, pp. 113–161, Jan. 2018, doi: 10.1016/B978-0-12-811889-4.00003-8.
- [71] L. Tan and J. Jiang, "Novel Adaptive IIR Filter for Frequency Estimation and Tracking," *IEEE Signal Process. Mag.*, vol. 26, no. 6, pp. 186–189, 2009, doi: 10.1109/MSP.2009.934189.
- [72] T. Ballal, R. B. Shouldice, C. Heneghan, and A. Zhu, "Breathing Rate Estimation from a Non-Contact Biosensor Using an Adaptive IIR Notch Filter," in *2012 IEEE Topical Conference on Biomedical Wireless Technologies, Networks, and Sensing Systems*, 2012, pp. 5–8, doi: 10.1109/BIOWIRELESS.2012.6172727.
- [73] H. Kim, J. Y. Kim, and C. H. Im, "Fast and Robust Real-Time Estimation of Respiratory Rate from Photoplethysmography," *Sensors*, vol. 16, no. 9, Sep. 2016, doi: 10.3390/S16091494.
- [74] C. Yifan *et al.*, "Non-Invasive Respiration Rate Estimation Using Ultra-Wideband Distributed Cognitive Radar System," in *Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2006, pp. 920–923, doi: 10.1109/IEMBS.2006.260759.
- [75] F. Shaffer, R. McCraty, and C. L. Zerr, "A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability," *Front. Psychol.*, vol. 5, p. 1040, 2014, doi: 10.3389/fpsyg.2014.01040.
- [76] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," *Front. Public Heal.*, vol. 5, p. 258, Sep. 2017, doi: 10.3389/FPUBH.2017.00258.
- [77] P. D. Welch, "The Use of Fast Fourier Transform for the Estimation of Power Spectra:

- A Method Based on Time Averaging Over Short, Modified Periodograms,” *IEEE Trans. Audio Electroacoust.*, vol. 15, no. 2, pp. 70–73, 1967, doi: 10.1109/TAU.1967.1161901.
- [78] I. Sommerville, *Software Engineering*, 9th ed. Boston: Pearson Education, Inc, 2011.
- [79] M. Masters and A. Schulte, “Physiological Sensor Fusion for Real-Time Pilot Workload Prediction in a Helicopter Simulator,” *Proc. AIAA SciTech 2022 Forum*, Jan. 2022, doi: 10.2514/6.2022-2344.
- [80] A. H. Roscoe, “Assessing pilot workload. Why measure heart rate, HRV and respiration?,” *Biol. Psychol.*, vol. 34, no. 2–3, pp. 259–287, Nov. 1992, doi: 10.1016/0301-0511(92)90018-P.
- [81] P. G. A. M. Jorna, “Heart rate and workload variations in actual and simulated flight,” *Ergonomics*, vol. 36, no. 9, pp. 1043–1054, 1993, doi: 10.1080/00140139308967976.
- [82] M. Masters, D. Donath, and A. Schulte, “An Exploratory Analysis of Physiological Data Aiming to Support an Assistant System for Helicopter Crews,” *Proc. 2nd Int. Conf. Intell. Hum. Syst. Integr.*, pp. 744–750, Feb. 2019, doi: 10.1007/978-3-030-11051-2_113.
- [83] J. Schwarz and S. Fuchs, “Validating a ‘Real-Time Assessment of Multidimensional User State’ (RASMUS) for Adaptive Human-Computer Interaction,” in *2018 IEEE International Conference on Systems, Man, and Cybernetics*, 2019, pp. 704–709, doi: 10.1109/SMC.2018.00128.
- [84] K. M. Feigh, M. C. Dorneich, and C. C. Hayes, “Toward a Characterization of Adaptive Systems: A Framework for Researchers and System Designers,” *Hum. Factors*, vol. 54, no. 6, pp. 1008–1024, Dec. 2012, doi: 10.1177/0018720812443983.
- [85] A. Schulte, D. Donath, and F. Honecker, “Human-System Interaction Analysis for Military Pilot Activity and Mental Workload Determination,” in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, vol. 00, no. c, pp. 1375–1380, doi: 10.1109/SMC.2015.244.
- [86] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.
- [87] Fang Chen *et al.*, *Robust Multimodal Cognitive Load Measurement*, 1st ed. Springer Cham, 2016.
- [88] F. Togo and M. Takahashi, “Heart Rate Variability in Occupational Health —A Systematic Review,” *Ind. Health*, vol. 47, no. 6, pp. 589–602, Nov. 2009, doi: 10.2486/indhealth.47.589.
- [89] D. Tao, H. Tan, H. Wang, X. Zhang, X. Qu, and T. Zhang, “A Systematic Review of Physiological Measures of Mental Workload,” *Int. J. Environ. Res. Public Health*, vol. 16, no. 15, Aug. 2019, doi: 10.3390/IJERPH16152716.
- [90] A. Kobayashi, A. Tong, and A. Kikukawa, “Pilot cerebral oxygen status during air-to-air combat maneuvering,” *Aviat. Sp. Environ. Med.*, Sep. 2002.

Appendices

Appendix A Output Variables Broadcast by the Proprietary SmartEye Pro Software

Table A.0.1 provides a list of the variables broadcast by the proprietary SmartEye Pro software over Transmission Control Protocol (TCP). The non-real-time stream was found to follow the real-time stream by approximately 0.74 seconds. For a description of each variable, readers are directed to the company's User's Guide.

Table A.0.1 Output variables from the proprietary SmartEye Pro software.

Real-time output	Non-real-time output
timestamp	timestamp
object	saccade
object x (pixel)	fixation
object y (pixel)	blink
object stdev x (pixel)	frame number
object stdev y (pixel)	estimated delay
pearson rho	real-time clock
gaze direction x	frame rate
gaze direction y	
gaze direction z	
gaze heading	
gaze pitch	
head position x	
head position y	
head position z	
head heading	
head pitch	
head roll	
pupil diameter	
pupil diameter quality	
filtered pupil diameter	
filtered pupil diameter quality	
frame number	
estimated delay	
real-time clock	
frame rate	

Appendix B Custom-Built ECG and EDA Processing GUI

Below is a screenshot of the graphical user interface (GUI) enabling the experimenter to manipulate the various ECG and EDA processing tools developed for this work. The tool can process and display real-time ECG and EDA from two participants simultaneously (“Pilot Left” and “Pilot Right”).

IFS
Real Time ECG & EDA Processing

General Collection Settings

Run in "dummy" mode Channels

Sampling Rate (hz)

Pilot Left

Settings

Log Folder

Log Prefix

Run Number

ECG Channel

EDA Channel

Sampling Rate for ECG Peak Findinn (hz)

Sampling Rate for EDA Event Detection (hz)

Baseline

Duration (min)

Receive Raw Process ECG Process EDA

Plotting

Plot Raw Plot HR/HRV/EDA

Raw @ 200 Hz Log Send Anycom Send LSL

HB & EDA Events

HR, HRV @ 1 Hz

EDA, EDAslope @ 10 Hz

Listen External Log

Pilot Right

Settings

Log Folder

Log Prefix

Run Number

ECG Channel

EDA Channel

Sampling Rate for ECG Peak Findinn (hz)

Sampling Rate for EDA Event Detection (hz)

Baseline

Duration (min)

Receive Raw Process ECG Process EDA

Plotting

Plot Raw Plot HR/HRV/EDA

Raw @ 200 Hz Log Send Anycom Send LSL

HB & EDA Events

HR, HRV @ 1 Hz

EDA, EDAslope @ 10 Hz

Listen External Log

Appendix C Study Participant Consent Form



3.5.2021

Einverständniserklärung zur Erhebung und Nutzung von audiovisuellen und peripher Physiologische Daten für wissenschaftliche Versuchszwecke

Titel der Studie: Eine Analyse der peripher physiologische Reaktionen auf Arbeitssituationen, die während eines simulierten Fluges in einem Hubschrauber ausgelöst werden

Studiendauer: Mai – Juli 2021

Versuchsleiter: Matthew Masters

Studienteilnehmer/-in:

- 1) Ich wurde über den Inhalt und die Vorgehensweise der Studie in verständlicher Form verbal aufgeklärt. Meine Fragen wurden ausreichend und verständlich beantwortet.
- 2) Ich bin damit einverstanden, dass im Rahmen dieser Studie Audio-, Video- und Physiologischer Aufzeichnungen (inclusive ECG, EDA, fNIRS, und Brustausdehnung) von mir für eine spätere Analyse erhoben werden.
- 3) Die studienbezogenen Daten werden pseudonymisiert und gespeichert und analysiert. Alle Daten werden ausschließlich zu wissenschaftlichen Zwecken im Rahmen der o.g. Studie oder Folgestudie durch den Versuchsleiter verwendet, vertraulich behandelt und ohne mein Einverständnis nicht an Dritte weitergegeben.
- 4) Die Teilnahme an dieser Studie und die Einwilligung in die oben beschriebene Nutzung meiner Daten sind freiwillig. Ich kann meine Einwilligung jederzeit ohne Angabe von Gründen mit Wirkung für die Zukunft widerrufen. Durch eine Verweigerung oder einen Widerruf der Einwilligung entstehen mir keine Nachteile. Im Falle meines Widerrufs werden meine Daten gelöscht und aus der Auswertung entfernt. Diese Löschung kann jedoch nicht mehr erfolgen, nachdem die Ergebnisse des Experiments bei einem Verlag zur Begutachtung (als Artikel- oder Buchpublikation) eingereicht wurden.
- 5) Als Studienteilnehmer/-in erhalte ich keine finanzielle Aufwandsentschädigung.

Ort, Datum

Unterschrift des Studienteilnehmers/der Studienteilnehmerin

Appendix D Pre- and Post-Experiment Questionnaire Results

All study participants completed pre- and post-experiment questionnaires in conjunction with the experiment presented in Section 4 of this work. The pre-experiment questionnaire was conducted to assess flight experience and determine participant readiness to engage in the experiment. The post-experiment questionnaire was conducted to assess subjective assessment of the experience and gather feedback regarding an in-cockpit physiological monitoring and notification system. Questions and answers have been translated from German into English for this report.

Pre-Experiment Questionnaire

Questions of and participant responses to the pre-experiment questionnaire are shown in Table A.0.2.

Table A.0.2 Participant responses to the pre-experiment questionnaire

<i>Participant</i>	<i>Age</i>	<i>Gender</i>	<i>Flight Experience</i>
Pilot 1	43	M	1900 hrs
Pilot 2	29	M	170 hrs
Pilot 3	51	M	3000 hrs
Pilot 4	43	M	3000 hrs
Pilot 5	43	M	2200 hrs
Pilot 6	51	M	3600 hrs
Pilot 7	44	M	2500 hrs
Pilot 8	48	M	4000 hrs
Pilot 9	25	M	Test Flight Engineer. Many hours helicopter simulator
Pilot 10	44	M	2000 hrs
Student 1	21	F	2 hrs lab simulator
Student 2	22	F	Flight screening tests, 15 min lab simulator
Student 3	23	M	120 hrs glider and motor aircraft
Student 4	25	M	100 hrs home computer simulator, 1 hr lab simulator
Student 5	22	M	No flight experience (real or sim)
Student 6	22	M	200 hrs glider and motor aircraft, 1 hr lab simulator, flight screening test
Student 7	23	M	600 hrs glider, home computer simulator, 30 min lab sim
Student 8	24	M	Flight screening tests, 4 hrs helicopter simulator
Student 9	20	M	Flight screening tests
Student 10	25	M	Flight screening tests, home computer simulator

<i>Participant</i>	<i>Current local time</i>	<i>How do you currently feel?</i>	<i>Hours of sleep last night</i>	<i>Coffee/caffeine today?</i>
Pilot 1	07:01	Good/As Usual	6	No
Pilot 2	07:14	Good/As Usual	7	No
Pilot 3	07:02	Good/As Usual	6	No
Pilot 4	06:49	Good/As Usual	7-8	No
Pilot 5	06:55	Good/As Usual	7.5	No
Pilot 6	07:06	Good/As Usual	8	No
Pilot 7	07:02	Good/As Usual	6	No
Pilot 8	07:01	Good/As Usual	6	No
Pilot 9	06:55	Good/As Usual	6	No
Pilot 10	07:01	Good/As Usual	7	No
Student 1	09:14	Good/As Usual	7	No
Student 2	08:17	Somewhat more tired than usual	6.5	No
Student 3	11:06	Good/As Usual	7.5	1 cup, 4 hr prior
Student 4	13:59	Good/As Usual	8.5	No
Student 5	10:42	Good/As Usual	8	No
Student 6	14:06	Good/As Usual	8	1 cup, 5 hr prior

Student 7	07:58	Good/As Usual	8	No
Student 8	11:00	Good/As Usual	7.5	No
Student 9	14:02	Good/As Usual	8	No
Student 10	07:55	Good/As Usual	7.5	No

* Answer options for “How do you currently feel” included: “more tired than usual,” “good/as usual,” and “more alert than usual.”

Post-Experiment Questionnaire

Participant responses to the first four questions of the post-experiment questionnaire are shown in Table A.0.3. Questions 5 through 8 were asked of all participants while questions 9 through 14 were asked only of the operational military helicopter pilots.

- Q1 On a scale from 1 to 5 where 1 is “fully disengaged” and 5 is “fully invested,” to what extent do you feel you were mentally invested in the successful completion of the **first** mission?
- Q2 On a scale from 1 to 5 where 1 is “fully disengaged” and 5 is “fully invested,” to what extent do you feel you were mentally invested in the successful completion of the **second** mission?
- Q3 On a scale from 1 to 5 where 1 is “very easy” and 5 is “very difficult,” how difficult was it for you to subjectively assess your mental workload after mission completion (using the provided video playback tool)?
- Q4 Has your physiological data ever been monitored during simulated or actual flight?

Table A.0.3 Participant responses to questions 1-4 of the post-experiment questionnaire

<i>Participant</i>	<i>Q1</i>	<i>Q2</i>	<i>Q3</i>	<i>Q4</i>
Pilot 1	4	4	2	no
Pilot 2	5	5	4	no
Pilot 3	4	4	3	no
Pilot 4	4	4	2	no
Pilot 5	5	5	2	YES
Pilot 6	5	5	2	no
Pilot 7	4	4	1	no
Pilot 8	5	5	1	YES
Pilot 9	4	5	3	no
Pilot 10	5	5	2	YES
Student 1	5	5	2	-
Student 2	3	4	2	-
Student 3	5	5	2	-
Student 4	5	5	2	-
Student 5	5	4	2	-
Student 6	5	5	2	-
Student 7	5	5	2	-
Student 8	5	5	2	-
Student 9	4	5	2	-
Student 10	5	5	2	-

* Q4 was not asked of student participants

- Q5 What do you think of a highly automated system that could respond to your physiological state while flying?

P1 *In principle, this could be very helpful. However, like everything in aviation, the interaction between pilot and machine must be thoroughly established.*

P2 *I hold the promise of the technology in high regard. It could significantly increase flight safety*

- P3 *It would be ok if it reacted with tips or notifications, but not ok if it would physically intervene.*
- P4 *Very interesting.*
- P5 *Sounds interesting if it was made to work effectively and if there was a benefit.*
- P6 *Determining how it would react is important. Providing suggestions makes sense. Suggestions to turn on the autopilot during a holding pattern or to simplify UAV command tasks could be nice.*
- P7 *I like it. People often don't accurately assess their own mental state. Receiving a reminder of my high stress could be helpful. At times, people may not know they are at their limit.*
- P8 *I'm not excited about the technology. What do I do with the data it provides? I suppose my acceptance would depend on how the system reacts. I must have trust in the system. I would appreciate receiving options to resolve problems.*
- P9 *I support autonomous assistance systems with warnings. Active intervention could only be performed with appropriate training of the aircrew (including override capabilities)*
- P10 *This would be a very good improvement to a cockpit. It is perfect for increasing flight safety.*
- S1 *I see both sides. Could be overbearing. Important to avoid falsely reporting a high workload. Could help to avoid burnout or assist with high workload.*
- S2 *Hard to say. Seems interesting.*
- S3 *Reviewing the data after flight could benefit those in training.*
- S4 *It seems extremely difficult to make it reliable. However, it seemed to capture my mental state better than I could after the fact.*
- S5 *The feedback is perhaps more useful after flight.*
- S6 *Such a system would be good as long as it only provided suggestions. If it could override pilot controls, it would be stressful.*
- S7 *Difficult to judge. Reporting or acting upon incorrect conclusions would be problematic.*
- S8 *Well, it was fun today. I think it would be cool.*
- S9 *As a feedback system, it makes lots of sense. It need not necessarily take action, however.*
- S10 *I wouldn't have a problem with it. Could alert someone when they are at their workload limit.*

Q6 What benefit could you see in a system that could accurately determine your mental workload during flight?

- P1 *This would be very beneficial, especially for the post-flight debrief.*
- P2 *It could prevent excessive demands on the pilot.*
- P3 *It could give advisory information and display abnormalities.*
- P4 *It could lead to better situational awareness concerning the remaining capabilities within the crew. Improvement of Crew Resource Management (CRM). From my work as a test pilot, I have seen this before.*
- P5 *I can't imagine such a system helping. Any warnings would likely be distracting when they come right at the moment when the workload is high.*
- P6 *Could benefit training situations by providing feedback. It is possible to get tunnel mental vision while flying but student pilots don't notice this.*
- P7 *Answered in my previous question.*
- P8 *It could also warn of low mental workload. It could prevent boredom.*
- P9 *It could benefit pilots by helping them train to better handle difficult situations.*
- P10 *It could result in a higher level of air traffic control/mission fulfillment, as it could relieve the pilot in critical situations (of high workload)*
- S1 *Could help to avoid burnout.*
- S2 *Unsure.*
- S3 *By reviewing the data after flight, stressful situations could be identified, discussed, and practiced.*
- S4 *I'm not sure how it could help.*
- S5 *It could help identify the difficult aspects/moments of flight. These moments could then be trained and improved.*
- S6 *It could be useful for training. It could hide/suppress superfluous information in stressful situations and later show it only when I have time to see and process it.*
- S7 *It could reduce the burden upon me when I have too much. It could allow me to perform a pre-landing checklist verbally for example when my hands are needed for manual flight.*
- S8 *(No answer)*

- S9 *I think the system could be helpful/meaningful in routine situations. In combat scenarios, however, I know there are situations when you have a high workload and any intervening system would be disturbing.*
- S10 *It could detect pilot performance limits and issue warnings. You could then think about whether to fly or not. It could be used to support targeted training.*

Q7 What problems could you see with a system that responds to your perceived mental workload?

- P1 *Warning... psychologizing here... I think you need qualified peer co-pilots with considerable flying experience to provide solutions to workload peaks. For example, sequencing.*
- P2 *It could be too heavily focused on an average pilot and not adapted enough to the particular person. Early or false notifications or alarms would result in a loss of trust in the system.*
- P3 *It could be more of a distraction than a help. Task saturation.*
- P4 *It could cause the pilot to rely too heavily on the system which is dangerous if a bias exists.*
- P5 *It could be distracting. When assistance or warnings are given should be established by each pilot individually.*
- P6 *Notification or feedback could be incorrect leading to a loss of trust and system acceptance.*
- P7 *Trust problems. Trust is lost quickly in computerized systems.*
- P8 *I see a potential conflict between the pilot's subjective workload and the system's prediction. False positive warning would be really bad leading to a quick loss of confidence in the system.*
- P9 *It could distract from the actual situation.*
- P10 *Workload is very subjective and difficult to determine by measurements. Thus, it is likely alarms would annoy or disrupt the pilot.*
- S1 *Incorrect assessments would be problematic. The pilot must be able to take control. It should be calibrated for each pilot individually.*
- S2 *(No answer)*
- S3 *It could be wrong and not know it.*
- S4 *Reliability seems nearly impossible.*
- S5 *It could be disturbing/distracting during the flight.*
- S6 *When someone is consciously trying to keep their workload load, it could lead to greater stress.*
- S7 *False positives are certainly a significant issue. Poorly designed/implemented interventions would do more harm than help.*
- S8 *(No answer)*
- S9 *People become insecure when they are told they have a high workload.*
- S10 *Incorrect assessments would be problematic leading to even more mental strain and/or confusion.*

Q8 Explain how you felt while subjectively assessing your workload post-flight using the video playback tool with the associated editable plot. Was it easy or difficult?

- P1 *It was easy because the tool helped me remember almost everything and also the emotions I had at the time.*
- P2 *It was hard for me to remember the details of each situation. More significant events such as landing or the tasking of the UAVs were easy to remember and assess.*
- P3 *It was a bit challenging due to the completely subjective nature of the task. It is easier to do comparisons between my experience and physiological data (such as heart rate).*
- P4 *It wasn't too difficult. Due to my work as a test pilot, I have done similar things to this in the past.*
- P5 *Undecided. Not bad, but the fundamental problem is still there. See previous responses.*
- P6 *I had to get back into the same head-space as when I was flying.*
- P7 *I could easily make the assessment without issue.*
- P8 *It was easier to do after flight than during flight as I have done before. Doing it afterward didn't add an additional strain as is the case when asked during flight.*
- P9 *Quite challenging. Trying to remember after the fact was more difficult than had I been asked in the moment.*
- P10 *It was fairly easy because I could go through the whole mission in a relaxed state. Additionally, I could compare the different situations over the entire period.*

- S1 *I found watching the situation with the video helped and it afforded me more time to think.*
- S2 *The playback tool gave a different angle of insight. I could better compare situations with each other.*
- S3 *With the video, I could put myself back into the situation. I got the feeling back from when I was flying.*
- S4 *The playback let me see more of the situation. I could compare each situation better with the others. I could compare each situation with the entire mission.*
- S5 *Putting myself back into the situation was exhausting. I think assessing the truth in real-time would have been easier.*
- S6 *The tool allowed me to fill the time between situations of high and low workload/stress with a workload assessment.*
- S7 *I liked being able to make comparisons over the entire flight. This helped me differentiate better between different situations.*
- S8 *Looking back, I could look at the situation more fully/completely and make a better assessment.*
- S9 *I could think back on the entire flight. This allowed me to better compare the individual moments.*
- S10 *It was a bit challenging having to "put myself back" into the situation, but it also wasn't very difficult being able to do it while relaxed.*

Pilot-only Questions (student participants were not asked)

Q9 How would you compare the level of mental strain/concentration required or experienced during these simulated flights to real flight in an actual helicopter?

Please explain.

- P1 *I think I am calmer in real flight because I feel I have a better command of the system. The visual references (when looking outside the cockpit) are (obviously) more realistic and the very helpful physical sensations of flight (the "butt meter") reduces stress considerably.*
- P2 *Stress during take-off and landing was comparable to a real flight. In the simulator, the flight parameters were more difficult to maintain and I was more easily/frequently distracted by small things because I was yet familiar with the aircraft.*
- P3 *I experienced less mental strain in the simulator than in real flight. This is because the consequences of mistakes are different. I found it important not to take the "flight behavior" of the simulator too seriously so I wouldn't get too frustrated. The simulator is a training device to evaluate new concepts.*
- P4 *A simulator never reflects reality and therefore the personal attitude towards flight is also different. I knew I could fly a bit more aggressively without anything serious happening.*
- P5 *My concentration level is similar in both situations. The mental load ("die Belastung") is higher in actual flight however as there are consequences with real flight.*
- P6 *Real flight requires considerably more mental strain/concentration. In the simulator, I could think "Oh it doesn't matter too much" while during real flight it must be done right. The focus of flight in the simulator is different. In the simulator, the focus is on fulfilling the simulated mission. In real flight, the main priority is flying.*
- P7 *Less mental strain/concentration in the simulator than in real flight. In the simulator, there is no danger to life. Also, more factors to consider during actual flight such as the co-pilot, weather, radio, visual distractions, glare, and other disturbances.*
- P8 *This is difficult to assess. The simulator is unrealistic, but regardless I find myself getting into the scenario. There is a greater basic load ("die Grundbelastung") in the simulator than in actual flight.*
- P9 *The two situations are not quite the same because the simulator is artificial, but they are close. The mental demand ("der Anspruch") is similar.*
- P10 *The simulator is realistic, but the workload in real flight is greater because of the "real" consequences. I found myself thinking "Oh it's just a simulation." However, the difference in workload between the simulator and real flight is not great.*

Q10 If you answered "yes" to Question 4 about whether or not your physiological data has previously been monitored during flight, please describe the situation.

- P1 *n/a*
- P2 *n/a*
- P3 *n/a*
- P4 *n/a*

- P5 *I participated in a “simulator sickness experiment about 13 years ago. I did not see or discuss the data as we did today.*
- P6 *n/a*
- P7 *n/a*
- P8 *I was monitored in both a simulator and actual aircraft for personal health reasons (details omitted for participant confidentiality)*
- P9 *n/a*
- P10 *I participated in a workload study in an A320 simulator where EEG was measured/collected. I had ECG collected during a centrifuge flight. ECG and pulse oximeter measurements were taken during a pressure chamber “flight” without oxygen.*

Q11 If data collection would not bother or interfere with you, how would you feel about your physiological data being monitored **by an autonomous system** during flight?

- P1 *It wouldn't bother me at all.*
- P2 *Well, as I said before, I think it would be okay because with such monitoring, the training of certain situations could be improved.*
- P3 *It would be okay.*
- P4 *I would be very interested in it.*
- P5 *I would ask why. In principle, I have nothing against it. If it does me some good, then don't mind. but I wouldn't like it if it was just for science or something. Drowsiness detectors in cars are not robust. I suppose if it really worked, it could be helpful.*
- P6 *I would not have a problem with it.*
- P7 *I think it would be a good thing. It would improve flight safety. Pilots bleed to death. A new flight suit with integrated sensors, even with pilot-healing properties/capabilities, would be great.*
- P8 *I would have no problem with this.*
- P9 *This would be perfectly fine as long as the data is processed securely (data protection) and reliably.*
- P10 *If the data could be used as part of an assistance system, then this would be okay for me.*

Q12 If data collection would not bother or interfere with you, how would you feel about your physiological data being monitored **by a human** (such as someone in a ground station) during flight?

- P1 *This would also be okay with me. The monitoring would likely completely fade into the background given the demands of flight.*
- P2 *This is also conceivable. I imagine the human would have a difficult time making an objective assessment of the data though.*
- P3 *This would also be okay.*
- P4 *I wouldn't have any problems with this.*
- P5 *Same as my previous answer. If it helps me, then it's okay.*
- P6 *I would not have a problem with this.*
- P7 *Same as above. This would be good.*
- P8 *This would be strange. I would have the feeling I was being controlled. I would feel uncomfortable.*
- P9 *This would be fine as long as it was evaluated the same as by the autonomous system.*
- P10 *Same answer as previous.*

Q13 If it were to exist in your cockpit, what could a highly-automated assistant system do for you? What should it respond to?

- P1 *It should ensure that the limit values of the system are complied with.*
- P2 *It could provide fatigue warning, degree of exhaustion in operational flight, vital states such as temperature and the need for fluids (water/drink/injury).*
- P3 *Notifications (provide information). Show abnormalities.*

-
- P4 *It could display an indication of the body's physical load limits. This would require several measurements to be carried out to determine minimum/maximum loads.*
- P5 *If it detects that I am only looking down (within the cockpit), it could give a warning. This wouldn't make sense for Tiger pilots though because they always fly as a pair and one often looks down while the other looks up and out.*
- P6 *I don't know if anything could really help.*
- P7 *It could provide "high workload" warnings. It could be a "limit indicator" for humans. It could share physiological states between the pilot and co-pilot.*
- P8 *When experiencing high mental workload, it could ask how it could help. It would be best to have a system that does exactly what I tell it to do. Siri in the cockpit.*
- P9 *It could provide warnings to interrupt tunnel vision.*
- P10 *The system should recognize if I am no longer able to act due to too high a workload and then support me by completing checklists, indicating abnormal system states, or autonomously executing tasks (e.g., putting down the landing gear).*

Q14 How much time would be acceptable for putting on sensors to monitor the physiological state during a normal flight?

- P1 *15 minutes*
- P2 *5 minutes*
- P3 *15 minutes*
- P4 *Maximum 3 hours. Basically, I don't care, because it's the job that counts. Due to my work as a test pilot, I have already done this for the most part.*
- P5 *5 minutes if casual training flight. It depends though what kind of a flight it is. For some flights, it would need to be integrated into clothes (helmet, gloves, shirt).*
- P6 *15 minutes*
- P7 *0 additional minutes. Sensors must be integrated into clothing (e.g., the flight jacket). The sensors/system would not be accepted if it was uncomfortable or took additional time.*
- P8 *3 minutes*
- P9 *2 minutes*
- P10 *5 - 10 minutes*

Appendix E Selectable Pages of the Multi-Function Display

Both the left and right Multi-Function Display (MFD) allow the pilot to navigate between “pages” by selecting the corresponding page in the “page selector” shown below. The pilot has the option of displaying two pages side-by-side or a single page spanning the entirety of the MFD. Below are snapshots of each page and a short description of the contents of each.

Page Selector

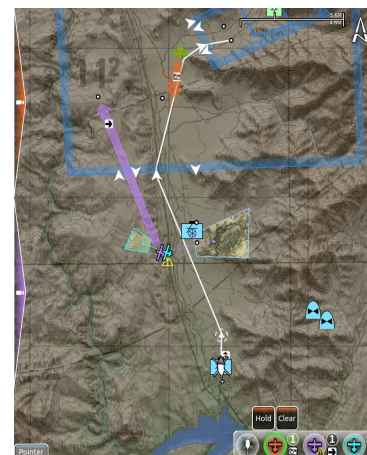


PFD



Standard IFR/VFR displays

MAP



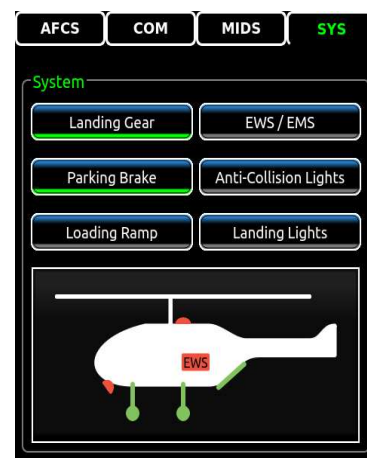
Interactive map with marked routes & UAV tasking tools

COM



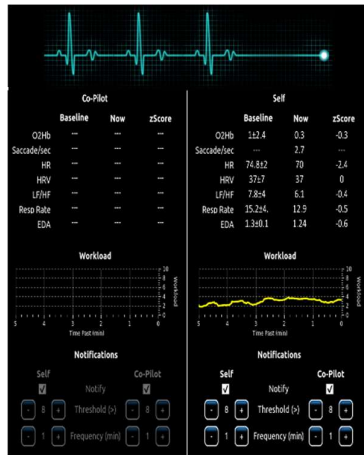
View/set communication frequencies

SYS



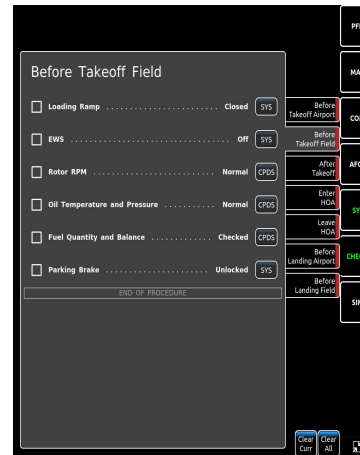
View/set system states (Landing gear, parking brake, etc.)

PHYS



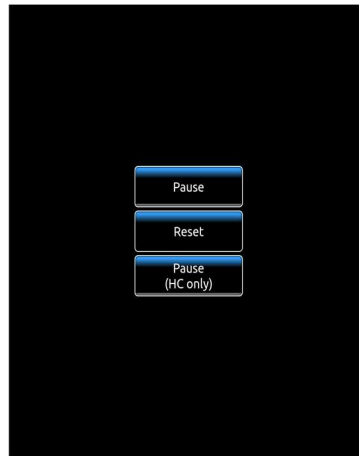
View pilot physiological state. Set triggers.

CHECK




Selectable checklists (pre-takeoff, after takeoff, etc.)

SIM




Simulator-relevant control (pause, restart, etc.)

Appendix F Pre-Brief Slides Presented to Participants Before Mission Execution



Universität der Bundeswehr München
Institut für Flugsysteme

Training Mission 1



- Missionsdauer: ca. 30 min
- Start: MOB SHOCKCENTER
- Own Callsign: BOOMER
- "Comm" settings sind nicht relevant
- Zu Starten:
 - Pre-flight Checklist bearbeiten
 - QNH: 1015
 - Troops einsteigen lassen (open loading ramp, close loading ramp)
- Auf "R2" hören



Universität der Bundeswehr München
Institut für Flugsysteme

Training Mission 2



- Missionsdauer: ca. 30 min
- Start: MOB SHOCKCENTER
- Own Callsign: BOOMER
- "Comm" settings sind nicht relevant
- Zu Starten:
 - Pre-flight Checklist bearbeiten
 - QNH: 1015
 - Troops (callsign TURTLE) einsteigen lassen (loading ramp öffnen und schließen)
- Zugang und Ausgang zur HOA nur über festgelegte Korridore, die auf der MFD-Karte eingezeichnet sind
- Auf "R2" hören

Appendix G Conditions and Triggered Actions for Simulated Missions

Mission 1

Condition Group	Condition(s)	Triggered Action(s)
1	Helicopter is on the ground Helicopter is within 500 m of START Loading ramp is open	Move soldiers to helicopter
2	Condition Group 1 is complete Loading ramp is closed	Move soldiers into helicopter
3	Pre-flight checklist is complete	
4	Condition Groups 2 and 3 are complete	Radio to Boomer: "Hello Boomer, R2 here. The assault team has boarded and the checklist is complete. Depart out of radio control point November at an altitude of 400 feet AGL."
5	Helicopter is within 1000 m of radio control point November	Radio to Boomer: "R2 here. Fly northeast into Area 51. Once there, locate the red smoke and drop off the assault team by hovering at that location."
6	Helicopter is inside Area 51	Start continuous red smoke Radio to Boomer: "R2 here. Again, locate the red smoke and drop off the assault team by hovering at that location at an altitude of 200 feet AGL. I repeat hover at an altitude of 200 feet AGL."
7	Helicopter is within 500 m of red smoke Helicopter altitude is less than 250 feet AGL Helicopter speed is less than 10 knots	Release soldiers Radio to Boomer: "The assault team is exiting the aircraft. Maintain this hover for 30 seconds then fly south to Point Sierra at the entrance of the canyon."
8	Helicopter is within 500 m of Point Sierra (entrance of the canyon)	Radio to Boomer: "R2 Here. Nice flying. Now fly east through the canyon to point Echo at an altitude of 200 to 300 feet A.G.L."
9	Helicopter is within 500 m of hidden weather marker	Weather quickly degrades (rain, wind, turbulence).
10	Helicopter is within 500 m of 2nd hidden weather marker	Weather degrades further.
11	Helicopter is within 500 m of 3rd hidden weather marker	Weather improves slightly.
12	Helicopter is within 500 m of 4th hidden weather marker (at end of canyon)	Weather returns to nominal conditions (no rain, no wind, no turbulence). Radio to Boomer: "R2 here. You made it through the canyon. Nice job. Now follow the marked route as closely as possible to the airport where you will land."
13	Helicopter is within 500 m of hidden marker "confusing_command".	Radio to Boomer: "R2 here. Change of plans. Fly north to Point Alpha. I repeat, fly south to point Bravo."
14	Helicopter is greater than 1000 m from hidden marker "confusing_command".	Radio to Boomer: "R2 here. Sorry for the confusion. Correction. Continue your approach to MOB Chineh and land at the airport. I repeat, continue your approach to MOB Chineh and land"
15	Helicopter is within 1000 m of MOB Chineh	Light turbulence induced
16	Helicopter lands at MOB Chineh	Turbulence stopped Radio to Boomer: "Well done! Please remain seated and calm for approximately two minutes."

Mission 2

Condition Group	Condition(s)	Triggered Action(s)
1	Helicopter is on ground Helicopter is within 500 m of MOB SHOCKCENTER Loading ramp is open	Move soldiers (TURTLE) to helicopter
2	Condition Group 1 is complete Loading ramp is closed	Move soldiers (TURTLE) into helicopter
3	Pre-flight checklist is complete	
4	Condition Groups 2 and 3 are complete	Radio to Boomer: "Hello Boomer, R2 here. The assault team has boarded and the checklist is complete. Depart out of radio control point November at an altitude of 400 feet AGL."
5	Helicopter is within 1000 m of radio control point November	Radio to Boomer: "BOOMER this is your assistant R2. TURTLE has loaded and the checklist is complete. Depart via radio control point SIERRA at 400 feet AGL. I repeat, depart via radio control point SIERRA at 400 feet AGL."
6	Helicopter is within 500 m of point SIERRA	Radio to Boomer: "R2 here. Fly east to point FOXTROT at 600 feet AGL and wait for further instruction. I repeat, fly east to point FOXTROT at 600 feet AGL and wait for further instruction"
7	Helicopter is within 500 m of point FOXTROT	Radio to Boomer: "Ok. Now use your UAVs to reconnoiter the marked points in the H.O.A. - Alpha, Bravo, Charlie, and Delta. Determine which location is displaying green smoke. After locating the green smoke, fly to that location and unload TURTLE there. While doing this maintain 600 feet AGL. I repeat, use your UAVs to reconnoiter the marked points in the H.O.A. - Alpha, Bravo, Charlie, and Delta. Determine which location is displaying red smoke. After locating the red smoke, fly to that location and unload TURTLE there. While doing this maintain 600 feet AGL."
8	Helicopter is within 500 m of location of red smoke	Radio to Boomer: "BOOMER this is R2. Change of plans. This is no longer a training mission, but is now an active search and rescue mission. We have been asked to locate and evacuate two pilots who crashed south of our position. Their call sign is AAROW. The crash site is marked on your map. Transport TURTLE to the crash site, evacuate the pilots, then return to MOB SHOCKCENTER. I repeat, transport TURTLE to the crash site, evacuate the pilots, then return to MOB SHOCKCENTER."
9	Helicopter is within 1500 m of crash site	Start continuous blue smoke near crash site Radio to Boomer: "BOOMER this is R2. ARROW has thrown blue smoke. Unload TURTLE at that location and evacuate AAROW."
10	Helicopter is within pre-defined crash site area Helicopter is on the ground Loading ramp is open	Downed pilots (AAROW) approach the helicopter
11	Condition Group 10 complete Loading ramp is closed	Downed pilots (AAROW) moved into helicopter. Radio to Boomer: "BOOMER, ARROW has boarded. Return immediately to MOB SHOCKCENTER. Intelligence has reported hostile activity in the area. For maximum security, fly the marked route as closely as possible at an altitude of 400 feet AGL. I repeat, return immediately to MOB SHOCKCENTER. Intelligence has reported hostile activity in the area. For maximum security, fly the marked route as closely as possible at an altitude of 400 feet AGL."
12	Helicopter lands at MOB SHOCKCENTER	Radio to Boomer: "Welcome back. Let out AAROW then remain seated and calm for approximately two minutes. Thank you."