USAARL-TECH-FR--2021-21

UNITED STATES ARMY AEROMEDICAL RESEARCH LABORATORY

Task Development to Validate Psychophysiological Monitoring of Cognitive Workload for Real-Time Operator State Monitoring

Jon Vogl, William Irvin, Christopher Aura, Kevin O'Brien, Hillary Plummer, Kevin Andres, Leonard Temme, Cristina Delgado-Howard, & Paul St. Onge

Notice

Qualified Requesters

Qualified requesters may obtain copies from the Defense Technical Information Center (DTIC), Fort Belvoir, Virginia 22060. Orders will be expedited if placed through the librarian or other person designated to request documents from DTIC.

Change of Address

Organizations receiving reports from the U.S. Army Aeromedical Research Laboratory on automatic mailing lists should confirm correct address when corresponding about laboratory reports.

Disposition

Destroy this document when it is no longer needed. Do not return it to the originator.

Disclaimer

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation. Citation of trade names in this report does not constitute an official Department of the Army endorsement or approval of the use of such commercial items.

Human Subjects Use Statement

The U.S. Army Aeromedical Research Laboratory Determination Official determined that this activity does not constitute research as defined under the human subjects protection regulations, as it is not "a systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge." [32 CFR 219.102(d)].

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
24-06-2021	Final	

4. TITLE AND SUBTITLE
Task Development to Validate Psychophysiological Monitoring of Cognitive Workload for Real-time Operator State Monitoring

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
6.2

6. AUTHOR(S)
Vogl, J.[1,2], Irvin, W.[1,3], Aura, C.[1,2], O'Brien, K.[1], Plummer, H.[1,3], Andres, K.[1,2], Temme, L.[1], Delgado-Howard, C.[1,3], & St. Onge, P.[1]

5d. PROJECT NUMBER
2100

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Aeromedical Research Laboratory
P.O. Box 620577
Fort Rucker, AL 36362

8. PERFORMING ORGANIZATION REPORT NUMBER
USAARL-TECH-FR--2021-21

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Development Command
504 Scot Street
Fort Detrick, MD 21702-5202

10. SPONSOR/MONITOR'S ACRONYM(S)
USAMRDC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
[1]U.S. Army Aeromedical Research Laboratory, [2]Goldbelt Frontier, LLC., [3]Oak Ridge Institute for Science and Education,

14. ABSTRACT

The Psychophysiologically Validated Assessments of Composite Workload for Real-time Operator State Monitoring project is one of a number of projects examining how to measure workload in real time at the U.S. Army Aeromedical Research Laboratory (USAARL). The current evaluation examines a subset of the iterative development process of the aforementioned project. This report aims to evaluate the effectiveness of the task battery and instrumentation in their ability to capture cognitive workload differences between gross levels of task demand across the cognitive resource space. This evaluation will also provide a data set using manufacturer-preferred data collection software for future reference. Results from this evaluation revealed areas for task battery improvement, recording instrument capabilities, and general guidelines for progressing with the overarching study.

15. SUBJECT TERMS
cognitive workload, CWL, electrocardiogram, ECG, pupillometry, eye tracking, blink rate

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UNCLAS	UNCLAS	UNCLAS	SAR	24	Loraine St. Onge, PhD
					19b. TELEPHONE NUMBER *(Include area code)* 334-255-6906

This page is intentionally blank.

Summary

The Psychophysiologically Validated Assessments of Composite Workload for Real-time Operator State Monitoring project is one of a number of projects examining how to measure workload in real time at the U.S. Army Aeromedical Research Laboratory (USAARL). The current evaluation examines a subset of the iterative development process of the aforementioned project. This report aims to evaluate the effectiveness of the task battery and instrumentation in their ability to capture cognitive workload differences between gross levels of task demand across the cognitive resource space. This evaluation will also provide a data set using manufacturer-preferred data collection software for future reference. Results from this evaluation revealed areas for task battery improvement, recording instrument capabilities, and general guidelines for progressing with the overarching study.

This page is intentionally blank.

Acknowledgements

This research was supported in part by an appointment to the Postgraduate Research Program at the U.S. Army Aeromedical Research Laboratory administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Army Medical Research and Development Command.

This page is intentionally blank.

Table of Contents

List of Figures

Table of Contents (Continued)

Page

List of Tables

Introduction

Currently, there is not a quantified dimension nor an all-encompassing definition of cognitive workload (CWL). Although many different CWL definitions have emerged over the years (Cain, 2007), a common characteristic among definitions is the interaction between the cognitive resources required to complete a task and the availability of these resources. Expanding on this, CWL can be conceptualized into three parts. First, CWL is the reaction of a specific individual to the demands of a task to achieve a criterion level of performance. The specific properties of the task (i.e., the processing limitations, multi-dimensional demand, and complexity) and the response of the individual (i.e., the perceived difficulty, motivation, strategies, state, etc.) can affect the level of CWL experienced. Second, CWL is a subjective experience of the internal physiological state of the individual, which permits a spectrum of assessment techniques (e.g., primary task, secondary task, physiological, and subjective measures) as windows into physiological correlates of CWL. Third, cognitive resources are both limited and multi-dimensional in accordance with the Wicken's Multiple Resource Theory (MRT), designed to account for a range of human experiences in regards to task performance (Wickens, 1984). Putting these elements together, we have defined precise experimental procedures to continue the modelling and assessment of CWL under specific dimensional levels of task demand.

The Psychophysiologically Validated Assessments of Composite Workload for Real-time Operator State Monitoring project is one of a number of projects examining how to measure workload in real time at the U.S. Army Aeromedical Research Laboratory (USAARL). This project will have participants complete lab-based CWL-inducing tasks (including the N-back, the Multi-Attribute Task Battery [MATB-II], and simulated aviation tasks, among others) at varying levels of task demand, and extensively train participants on the use of a subjective workload scale. Throughout this process, psychophysiological metrics will be collected and synched together to assess the changes in psychophysiological responses to small incremental changes in task demand or the level of workload of tasks. The information will be used to develop algorithms for artificial intelligence programs to analyze and calculate CWL in real-time. The combination of synchronized performance, physiological, and subjective data in response to specific levels of task demand will generate a unique database to develop these CWL assessment algorithms. The goal of this project is to develop a reliable metric to certify new helicopter flight decks as safe to fly based on the CWL demands placed upon the pilot.

This report details the initial pilot phase and experimental design of this project. In particular, this report describes the theoretical underpinning of the work, study design for our first compiled task battery, and the preliminary pilot data used to assess the feasibility of moving to the next phase of research data synchronization.

Background

Cognitive workload (CWL) is the necessary mental processing effort needed to engage in a task or set of tasks to reach a criterion level of performance. CWL requires the combination of known facts held in long-term memory, awareness of the current situation, and information processing through working memory. The amount of CWL capacity differs between individuals and is task dependent (Young et al., 2015). Due to these many facets of CWL, it is important to find structure in the models that attempt to explain CWL. This report provides a very brief look at the CWL models used to structure the current study.

Performance Models

CWL and associated task performance is conceptualized in two theoretical frameworks. The first is the inverted-U hypothesis, which states that performance is optimal between the extremes of “too low” and “too high” levels of arousal (Hebb, 1955). The initial work of Yerkes and Dodson (1908) revealed that exposure to medium strength electric shocks was more effective for learning a habit in mice, compared to both low and high strength electric shocks. While Yerkes and Dodson (1908) focused on the relationship between state (i.e., the transient characteristics of cognition and physiology) and learning, the idea has been expanded and generalized to the effects of arousal on performance (Hebb, 1955; Kahneman, 1973; Teigen, 1994). The inverted-U relationship between performance and arousal and the impact of task complexity is best depicted in the figures provided by Hebb (1955), reproduced in Figure 1a.

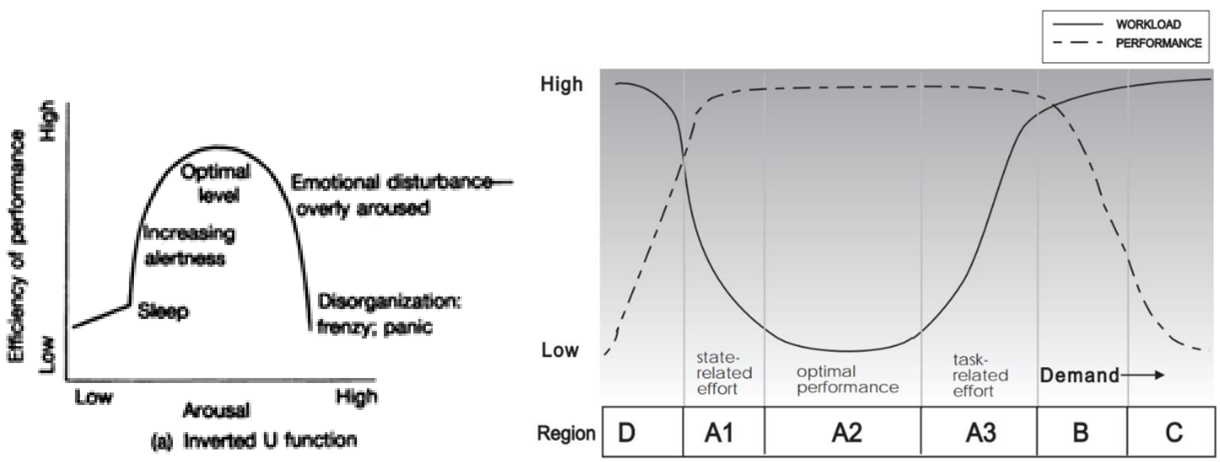


Figure 1. Theoretical frameworks detailing the relationships between aspects of CWL and overall task performance. The Yerkes and Dodson (1908) inverted-U model (a) and the Region Model of Performance (b).

The second framework encompasses multiple models. The Region Model of Performance, initially proposed by Meister (1976) and extended by de Waard (de Waard, 1997), expands on the inverted-U models by modelling performance as a function of increasing task demand, as seen in Figure 1b. The performance curve (the dashed lined) of the Region Model mimics the inverted-U curve of the Yerkes-Dodson Law, with the domains of low and high task demand being correlated with low levels of performance and moderate task demand being correlated with optimal performance.

The Region Model offers more granularity by partitioning the performance curve into specific regions relative to experienced CWL and the trends of performance demonstrated by the operator. For example, CWL overload is represented in region B of the model where the expenditure of cognitive resources (i.e., CWL) has reached its capacity and task performance is deteriorating with increasing task demand due to an inability to recruit and apply additional cognitive resources. Thus, transitioning into region B of the model demonstrates the transition into catastrophic failure (Hebb, 1955; Kahneman, 1973; Teigen, 1994), as well. Once task demand pushes beyond the boundary of region A3 to region B (i.e., a threshold traditionally called 'the redline'), task performance deteriorates rapidly (at a rate dependent upon the nature of the task). This redline can be determined by measuring performance alone. However, as seen in Figure 2, it is clear that performance metrics are insensitive to some changes in CWL (for example, in regions A1 and A3), so other CWL assessment techniques are required.

Several approaches to tackling the challenge of CWL assessment are reported in the literature. One approach uses standard CWL subjective rating instruments, the second approach uses standard task performance, and the third approach uses physiologically-based biomarker measures. There are numerous comprehensive review papers in this area, for examples see (Cain, 2007; Casner & Gore, 2010; Korbach et al., 2018; Naismith et al., 2015; O'Donnell & Eggemeier, 1986; Young et al., 2015). In the current study, each of these CWL assessment techniques were employed to obtain a comprehensive review of the effect of controlled changes in task demand on CWL metrics.

Breaking the Mold – Cognitive Resources and Incremental Task Demand

Not all tasks are created equally. A great disparity exists between tasks in terms of demand, complexity, performance criteria, performance-resource relationships, and the nature of the task components' interactions with operators. In a CWL analysis, it is necessary to understand and control these features of a task in order to recognize how much the task is taxing a subject in terms of CWL. While many of these task-specific factors can be easily evaluated, it may be less intuitive to understand how specific components of a task interact with the cognitive resources of the subject. MRT can be used to better understand the resources a task will tax across all subjects (Wickens, 1984). Instead of assuming a single cognitive resource exists that is shared across all methods of processing[1], these pools are based on four dichotomous dimensions: stages (perceptual + cognitive/response), modalities (visual/auditory; can also expand to haptic), codes (spatial/verbal), and visual channels (focal/ambient). These dimensions can be seen in Figure 2, a depiction of the 'Wickens Cube' (Wickens, 2002). Using this model, all of the tasks used in the current study were characterized by the resources depicted in the Wickens Cube. Additionally, the N-back task battery used in this study was specifically designed to tax each resource detailed in MRT (with the exception of the ambient/focal dichotomy). This approach allows for analyses of CWL as a function of the resources being taxed in addition to the standard approach of examining CWL as a function of task demand or complexity.

---

[1] The 'modal' view, as was initially commonly thought to be characteristic of CWL, assumes a single pool of cognitive resources exists and all cognitive processes draw from the same pool (Kahneman, 1973). However, evidence suggests that some tasks can be more efficiently time-shared than others (e.g., talking to a passenger while driving a vehicle versus texting while driving), lending support to the idea of multiple resource pools (Wickens, 2002).

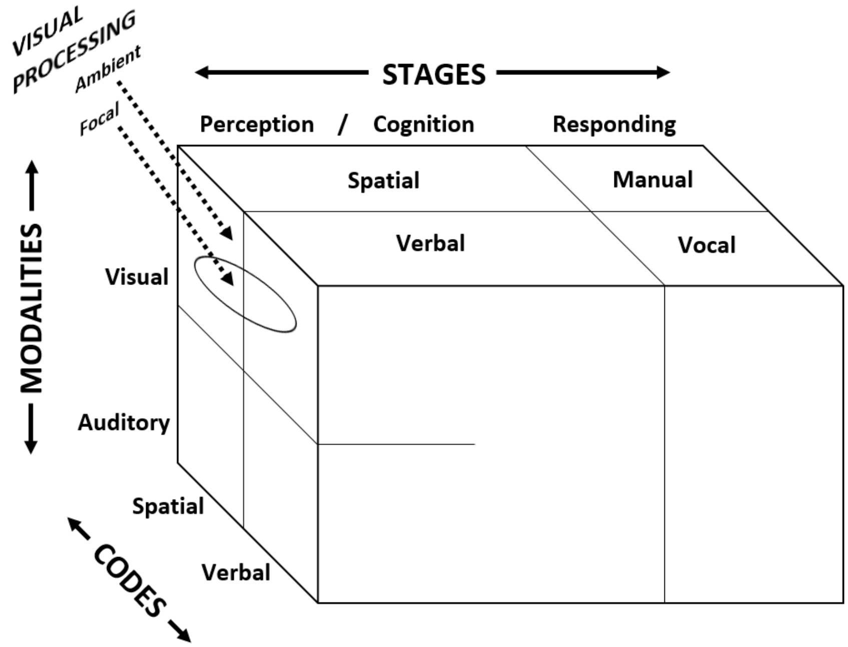


Figure 2. Visualization of the MRT model; the 'Wickens Cube' adapted from Wickens (2002).

Figure 2 shows four dichotomous dimensions: Stages (x-axis), modalities (y-axis), codes (z-axis), and visual channels (nested in the visual modality). Stages include the perception + cognition of data and the response selected and executed. Modalities are the channels used to deliver the information to be perceived and processed via cognition. The modality dimension represents sensory systems that can be efficiently time-shared and, in MRT, is often separated into visual and auditory modalities (but can be expanded to other senses, such as the haptic modality). Codes represent the time-sharing efficiency between the analogue/spatial and the categorical/symbolic process parts of perception, cognition, and responses that depend on separate resources divided across the two cerebral hemispheres (akin to Baddeley and Logie's [1999]) model of working memory which features a "visuo-spatial sketchpad" and "phonological loop" compartmentalizing the spatial/verbal dichotomy). Visual processing is divided into ambient and focal processing, with ambient representing unique aspects of peripheral vision and focal representing the information flow in central vision.

Currently there is not a quantified dimension of cognitive workload akin to classic psychophysical functions (i.e., such as with luminance of a visual target or pressure level of a sound). In the literature, workload is defined *a priori* a number of different ways. Most often, workload is specified by task-specific variables such as the number of steps in a task (i.e., complexity), the perceived difficulty of the task, or even the number of concurrently performed sub-tasks required to complete the overall task. This process almost universally leads to a specific dichotomy formed for the studies found in the literature: Low workload versus high workload task evaluation. For example, a task that requires the completion of two steps can be defined as low CWL and a task that requires completion of four steps can be considered high CWL. Research participants subjectively rate two steps as a lower level of CWL than four steps. Performance and objective physiological metrics show mixed results as detailed in "Workload and Performance: Associations, Insensitivities, and Dissociations" by Hancock and Matthews

(2019). With this approach, it is not possible to establish a psychophysical function of workload based on physiological metrics. Two and four tasks are only two points along the continuum of workload; more points are needed. To deliver these additional points, the current study aims to utilize small, incremental steps in task demand to record the just noticeable difference (JND) in perceived CWL to develop a psychophysical curve for CWL. More information on this methodology will be released in a future report.

The Current Evaluation – Pilot Data Analysis

The current evaluation examines a subset of the larger experiment described above. This report aims to evaluate the effectiveness of the instrumentation used for its ability to detect CWL differences between gross levels of task demand across the cognitive resource space. This evaluation will also provide a dataset using manufacturer-preferred data collection software (to be used as a comparison in future evaluations utilizing custom software to synchronize the data streams). Additionally, this evaluation served as a stepping-stone in the iterative development of the tasks and measures used in the current study. As such, we will report the data collected from these evaluation efforts and reflect on our path moving forward to increase overall efficiency and productivity of the database of performance, physiological, and subjective CWL data being collected for the larger scope of this project. This study was conducted under a U.S. Army Medical Research and Development Command Institutional Review Board-approved protocol #10881 and U.S. Army Aeromedical Research Laboratory (USAARL) protocol #2020-013.

Methods

Subjects

The pilot data presented in this report was collected from two U.S. Army Aeromedical Research Laboratory (USAARL) personnel. Each subject read and signed an informed consent document prior to participating.

Materials and Instruments

Performance Measurement Instruments.

N-Back Task Battery.

The task set implemented in this evaluation was based on the classic N-back task paradigm. The N-back task is a sequential memory storage and manipulation task where subjects are required to hold a sequence of previously presented information in their short-term memory while evaluating newly presented information against the stored sequence. The “N” in N-back is a variable (typically spanning from 0 to 6) that determines the length of the sequence of stimuli the subject is required to store in their working memory. As more stimuli are presented, the subject will need to update the stored sequence with each stimulus presentation. The primary goal is to compare the stimulus currently presented to another stimulus presented “n” presentations prior; in the event of a match, the subject would press a button marked “Yes,” while subjects would press a button marked “No” for non-matches. Figure 3 depicts an example of an N-back task using number stimuli with N = 2.

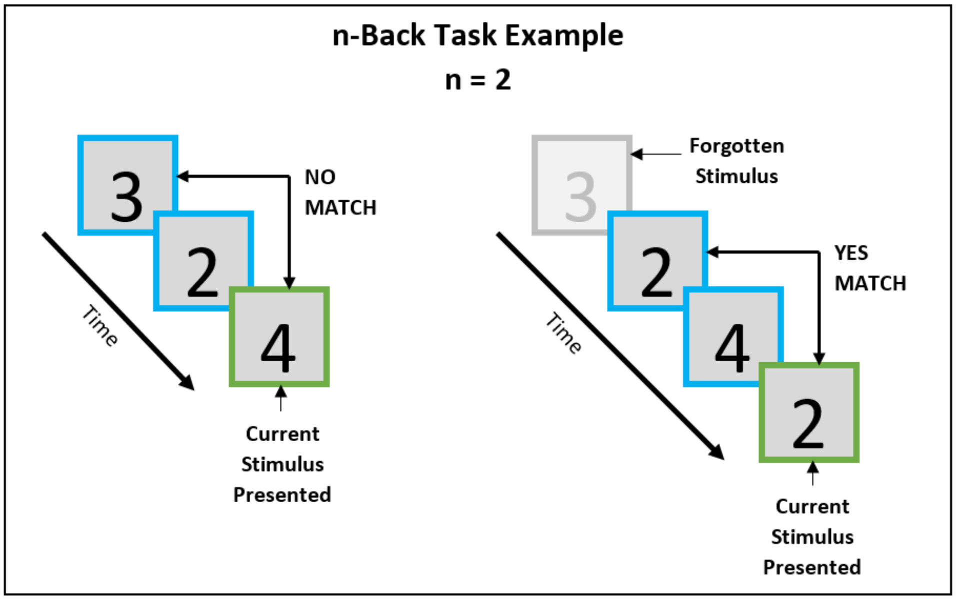


Figure 3. Example of the standard N-back task paradigm.

The task battery utilized in the current evaluation consisted of four N-back tasks that drew upon the spectrum of cognitive resources defined by MRT. These variable N-back tasks are most easily defined by the nature of the stimuli presented to the subject. The tasks in the N-back battery are as follows: Visual-Verbal, Visual-Spatial, Auditory-Verbal, and Auditory-Spatial N-back. The Visual-Verbal N-back task presents to the subject a sequence of images containing a single alphanumeric character, as seen in Figure 4a. As such, the Visual-Verbal task gets its name due to the type of stimuli used for the task, where the stimulus is presented to the visual modality (i.e., displayed on a screen) of the subject and requires storage of the verbally coded information in the subject's articulatory loop (Baddeley, 1986). The other N-back tasks follow suit with their naming conventions, as seen in the stimuli examples provided in Figure 4b for Visual-Spatial, Figure 4c for Auditory-Verbal, and Figure 4d for Auditory-Spatial. Additionally, a modified cognitive footprint, a depiction of each cognitive resource utilized for a task as described by MRT, is provided for each task in Figure 4 (Seeber, 2007). For each task, a stimulus was presented for a duration of one second (s) and was followed by either one second of a visual white noise mask during the visual-based tasks or one second of silence for the auditory-based tasks. All stimuli were designed to be similar in luminance and sound intensity. The subjects input their responses manually by pressing a key labeled either "Y" for "Yes Match" or "N" for "No Match."

In the current evaluation, accuracy and reaction time data were collected using the N-back task battery for a duration of seven minutes for each task at N levels of 0, 2, and 5. The stimulus set size (i.e., the number of stimuli available as options in the task presentation) was 10 for each task and the random number generator seed was always set to 1 for each task and subject (ensuring each subject received the same stimuli presentation). Using these task battery settings, the task demand was manipulated to weigh on specific cognitive resources at varying levels of perceived difficulty by the subject, with N = 0 equating to easy/boring difficulty, N = 2 equating to moderate difficulty, and N = 5 equating to very difficult/impossible difficulty. In addition, because CWL can be measured using not only performance but also physiological and subjective metrics, a suite of physiological measures were collected as well as a subjective responses to the Crew Status Survey (CSS).

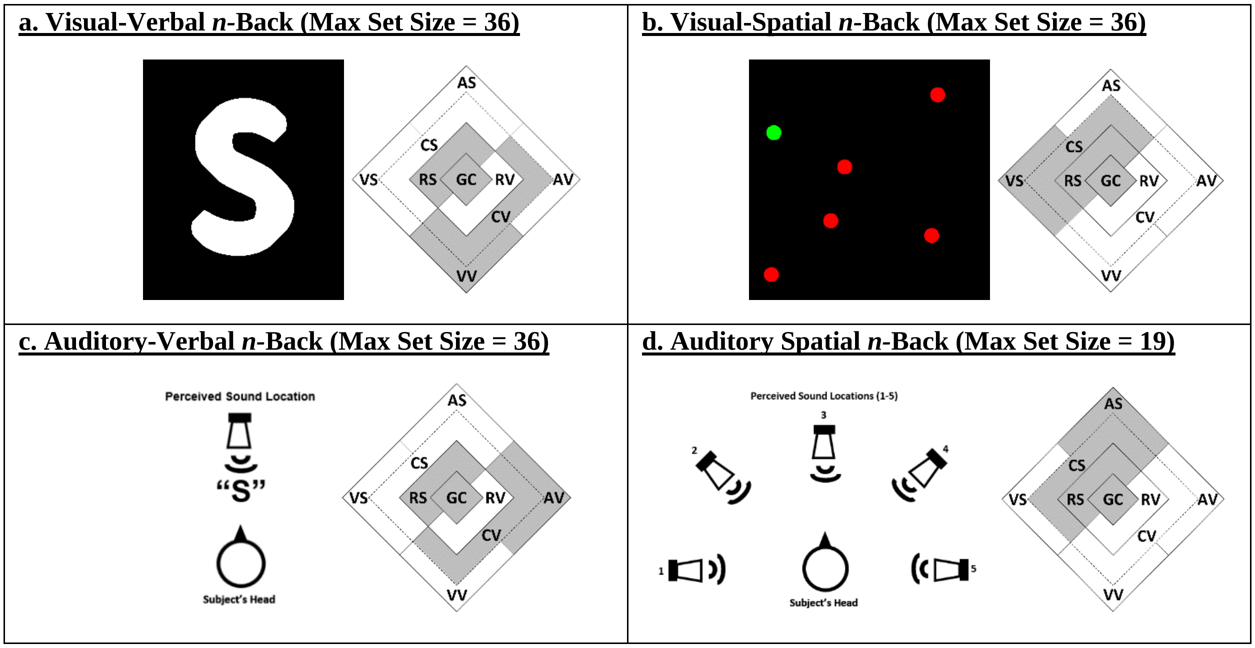


Figure 4. Depiction of stimuli and cognitive footprints of each task in the N-back battery.

Physiological Measurement Instruments.

Biopac MP160.

The Biopac® MP160 system (Goleta, CA) was used to collect electrocardiography data and participants’ respiration rate with a sample rate of 1 kilohertz (kHz), while they were engaged in the tasks described above using the ECG100D and RESP100D modules, respectively. Traces were collected using the Acq*Knowledge* software suite (Biopac®; Goleta, CA), and data were analyzed using custom software written in Python, MATLAB (Mathworks®; Natick, MA), and Excel (Microsoft; Seattle, WA). Separate recording files were created for each task block to facilitate analysis, but the sensors remained on the participants throughout the entire collection session.

Pupil Labs Core Monocular Eye Tracking System.

Pupil area, line-of-sight, and eye-blinks were captured using the Pupil Core monocular eye tracking system (Pupil Labs; Berlin, Germany) and recorded using the Pupil Record Software (Pupil Labs; Berlin, Germany). This system employs a small infrared (IR) camera and illuminator suspended from a head mounted frame (akin to eyeglasses) to sit just below the right eye. This camera captures images of the eye at 120 hertz (Hz), with a selectable resolution of to 1024 x 1024 pixels. A second, visible-light, camera mounted to the front of the frame captured a forward facing view at a rate of 200 Hz which was used to register line of sight (LOS) to the forward scene. The software performs and logs real-time estimates of LOS, pupil area, and the incidence of eye-blinks, but also stores the video feed from each camera to allow for post-hoc reparsing and re-analysis. Separate recording files were created for each block to facilitate analyses. These data were collected on the same computer as the physiological data and were synchronized using the system clock. Prior to each recording block participants' LOS were calibrated using the Pupil Record software's internal calibration procedure on the same screen used to administer the CWL tasks. The software itself used a 3-dimensional (3D) modeling algorithm to detect the location of the eyeball in relation to the camera. This estimate was then used to convert the camera pixel area occupied by the pupil to estimate the diameter in millimeters (mm). In our preliminary analyses, this approach was effective at reliably estimating pupil diameter in all but one recording session. Further use and assessment is required, however, to determine the incidence of this miscalculation, and how to avoid this issue in future data collections.

Subjective Measurement Instruments.

Crew Status Survey (CSS).

The Crew Status Survey (CSS) is a seven-point unidimensional CWL assessment scale that was validated and verified through the testing of trained pilots and aircrew members (Ames & George, 1993). Operators were asked to assess their experienced fatigue levels and subjectively perceived level of CWL using two responses on the seven-point scale (1 = low workload; 7 = high workload) as shown in Table 1. For the CWL assessment, the first response requires the operator to assess the maximum level of CWL that was experienced during the task by selecting one of the seven levels. The second response asks operators to identify the statement that best describes the average level of CWL experienced during the task. Addressing both the maximum and average values allows for more strategic analysis of the data compared to assuming operators were not biased towards maximum workload inducing events throughout the task. This quick assessment offered experimenters the ability to assess subjective CWL during the task and/or following the task.

The CSS is a revision of the School of Aerospace Medicine (SAM) Form 202 CWL estimate that offered many advantages such as ease of use, minimal training effort, and scale steps that were anchored in absolute terms. However, the SAM Form 202 lacked technical verification of the continuous underlying CWL dimension being tested. The development of the CSS involved revising the SAM Form 202 using a pair comparison and rank order estimation test. The results of Ames and George's (1993) efforts yielded a revised scale (adopted as the CSS) with verified ordinal steps and nearly equal psychological intervals between steps. Based on this set of assumptions the authors argue that the CSS can be considered an interval scale; thus, allowing more types of statistical analyses to be performed with the resulting data than an ordinal scale. The potential usefulness of the CSS in flight test applications is supported by the many advantages of the SAM Form 202 including the absolute CWL score output (compared to relative workload score output), and the fact that the revisions were performed with trained pilots and aircrew members.

Table 1. Crew Status Survey Mental Workload Scale Levels and Descriptions

Level	Description
1	Nothing to do; No system demands.
2	Light Activity; minimal demands.
3	Moderate activity; easily managed considerable spare time.
4	Busy; Challenging but manageable; Adequate time available.
5	Very busy; Demanding to manage; Barely enough time.
6	Extremely Busy; Very difficult; Non-essential tasks postponed.
7	Overloaded; System unmanageable; Essential tasks undone; Unsafe

Dependent Variables

Performance Measures.

The three performance metrics collected from the N-back task battery were accuracy, response latency, and response category. The output of the N-back task battery provided summary statistics for each performance metric. Accuracy, here, was described in terms of percent of correct responses to the N-back task. Response latency was measured as a response time in milliseconds (ms) between stimulus presentation and the subject's response. For the response category metric, the subject's responses were categorized using a confusion matrix to tally the number of hit, miss, false alarm, and correct rejection responses for each task.

Subjective Measures.

Upon completion of each N-back task configuration, the subjects completed the CSS to assess their perceived fatigue and CWL levels. The CSS was presented digitally along with digital instructions (e.g., changing the response method from marking a sheet of paper to pressing numeric keys). Subjects input their responses for the fatigue, average CWL, and maximum CWL scales into the software where it was stored alongside the performance data for the task.

Physiological Measures.

A number of common physiological measures were collected for this report including electrocardiogram (ECG), electro-dermal activity (EDA), respiration rate, and changes in pupil diameter and LOS. These physiological dependent variables (DVs) will be used in this larger research effort to develop biomarker-based response profiles associated with calibrated surrogate task CWL stimulus level. These physiological DVs, with their response profiles to known and calibrated CWL conditions, serve as input data to develop multivariate CWL predictors. Standard statistical techniques such as factor analysis, neural networks, multivariate regressions, support vector machines, etc. will be used to create algorithms and statistical predictors that classify and discriminate CWL from the physiological DVs. With the success of predictive discriminations based on physiological status DVs recorded during task performance, an objective, valid, and generalizable standard scale will be developed to monitor, assess, and predict CWL. This standard scale is expected to provide a transparent, real-time operator state monitoring index. For the current evaluation, preliminary analyses of these signals were completed to demonstrate effective data collection and data quality.

Electrocardiogram (ECG).

Data were collected through the Biopac® MP160 linked ECG100D electrocardiogram amplifier module (Biopac®, Goleta, CA). The ECG100 is a single channel, high gain, differential input, biopotential amplifier designed specifically for monitoring the heart's electrical activity. Single-lead electrodes were placed on each of the subject's clavicles and one below the left pectoral area. ECG data was collected at a rate of 2,000 Hz. The ECG data were filtered using a high pass infinite impulse response filter at 0.5 Hz. After the data were filtered, each subject's ECG data were visually inspected for artifacts.

ECG recordings were saved in the Acq*Knowlege* (4.4.1) ACQ format. A custom python script was developed using the bioread (2.1.2) library (Vack & Ollinger, 2020) to automatically separate ECG from other data channels. The ECG data along with recording timestamp information (referred to as "relative time," as the start of the recording has a time of zero) pulled from the acquisition (ACQ) file were inserted into a Pandas (1.1.3) data frame (The pandas development team, 2020). Local time for the start of recording was pulled manually from separate records and incorporated into the data frame (referred to as "absolute time") before it was stored as a comma separated value (CSV) file.

The CSV files containing the ECG traces were fed into a peak-detection algorithm which was built in-house for a prior experiment (O'Brien et al., 2020). The outputs of which were peak-labeled ECG traces suitable for spot inspections and CSV files containing inter-beat intervals (IBIs) in milliseconds along with corresponding absolute and relative time information (O'Brien et al., 2020). Examination of a small, random sample of trace segments did not show any missed or incorrectly labeled R-wave peaks, as would be expected for controlled laboratory data collected with a precision instrument. The CSV files containing IBIs were temporally binned into 30- and 60-second intervals relative to the start and end of both absolute and relative time, and time domain heart rate variability (HRV) metrics were calculated for each temporal bin using the HRV-analysis library (Champseix, 2018). Based on guidelines in Shaffer and Ginsberg (2017), these temporal bins are sufficient for calculating minimum, maximum, mean, and standard deviation of heart rate (HR) as well as the root mean square of successive differences (RMSSD) for IBIs as variability metrics. These metrics were also visualized as step functions for ease of review. Future analyses may include longer temporal bins to permit additional time-domain HRV metrics or frequency-domain metrics.

Respiration Data.

Respiration data was collected using the Biopac® respiration pneumogram amplifier module (RSP100D), which is a single channel, differential amplifier designed specifically for recording respiration effort. Respiration was collected through a transducer strap placed on the chest or abdomen of the subject and was sampled at a rate of 50 Hz. After data collection, the respiration data were filtered using a high-band pass infinite impulse response (IIR) filter at 0.05 Hz, and then visually inspected for artifacts, which were corrected using linear interpolation. The Acq*Knowledge* 5.1 (Biopac, 2021) software was used to extract respiration rate in breaths per minute using the peak-to-peak values between respiration waves. The average beats per minute (BPM) for each of the tasks was used in data analyses presented herein.

Oculometrics.

Eye tracking was conducted using a small infrared (860 nanometer), video-based, camera system (Pupil-Labs Core Monocular; Pupil-Labs, Berlin, Germany) suspended below the right eye from a head-mounted frame. Images of the eye were collected at a sample rate of 120 Hz using Pupil-Labs' Record software suite. The pupil was identified and measured using a custom algorithm developed by Pupil Labs. The system and algorithms are detailed by Kassner, Patera, and Bulling (2014). Participants completed a 4-point calibration procedure, which allowed the system to develop a 3D model of the eye, from which Pupil diameter was derived (in mm), and LOS coordinates were registered to a forward facing visible light world camera. This system also recorded disruptions of the pupil visibility to the camera as blinks, and stored a time-stamped log of these events, which was used to perform the blink activity analysis included in this report.

Table 2. Summary of Data Collection Methods

Method	**Construct Measured**	**CWL Metrics**
Biopac® MP160 – ECG100D; EDA100D; RESP100D	Electrocardiogram Respiration	Heart rate (BPM) Respiration rate
Pupil Labs Core Monocular Eye Tracking System	Pupilometery Blink Dynamics	Pupil diameter (mm) Blink count Blink duration (s)
Crew Status Survey	Subjective fatigue Subjective workload	Fatigue rating Average workload rating Maximum workload rating
N-Back Task Battery	Accuracy Response latency	Accuracy (%) Response category Reaction time (ms)

Procedure

Subjects first read and signed an informed consent document. Prior to starting the tasks, each subject was instrumented with physiological sensors. First, two ECG100D Biopac® module electrodes were placed on the subject's clavicles and one below the left pectoral area to collect ECG data during the evaluation. Second, the pneumogram amplifier module (RSP100D) transducer strap was placed around the subject's chest, taking care not to interfere with the ECG sensors. Once instrumented with the Biopac® sensors, the subject was asked to don the Pupil Labs Core head-mounted monocular eye tracking system. Once the headset was placed on the subject's head, the camera was properly adjusted to capture the subject's eye in the video stream. The subject was then situated roughly 2 feet in front of the experiment presentation monitor.

The N-back task battery was completed next, to serve as the cognitively demanding task set for this evaluation. Each subject experienced the same tasks, in the same order, with the same random seed for stimulus list generation. The N-back task configuration order was as follows: Visual-Verbal, Auditory-Spatial, Auditory-Verbal, and Visual-Spatial. For each task configuration, the subject first completed the N = 0 level first, followed by N = 2, followed by N = 5. Each task presentation session was prefaced by an explanation of the task and instructions, verification of the subject's understanding of the task configuration and N-level, and answering any questions the subject may have. Each task presentation session consisted of 210 trials and lasted 7 minutes. Subjects were offered a break between each task presentation session.

The first subject completed their sessions across two days, while the second subject completed their session in one day. For the first subject, the first task configuration (visual-verbal, N = 0, 2, and 5) was completed on day one, and the remaining task configurations were completed the following day.

Data Analysis

For the current evaluation, the analysis of the collected CWL metrics remained simple to verify the working status of the apparatus and to determine the sensitivity of the sensors to changes in CWL as a function of task demand and resource domain. Summary statistics for each CWL metric were analyzed and the quality of the data output reported.

It is important to note, to achieve a seven-minute task duration with two seconds per trial (i.e., one second of stimulus presentation followed by one second of mask/silence), 210 total trials were recorded for each task and N-level combination. Due to the unsynchronized nature of the evaluation data set, the data used across each CWL metric analysis may vary. For example, performance metric analysis will use the entire 210-trial dataset for each task, as performance data were recorded using the same software that presented the task. However, for the physiological metrics, only the middle 5 minutes of the collected dataset (removing the first and last minute of the dataset) for each task was analyzed, to avoid any effects of the desynchronized collection. This ensured that the data analyzed for the physiological metrics was data collected during task engagement, and not during the starting/stopping of the recording equipment.

Results

Summary statistics of the CWL response metrics are provided for each dependent variable below. Results from each task are reported, and each task is represented by an abbreviation of the task's name (i.e., Auditory-Spatial = AS, Auditory-Verbal = AV, Visual-Spatial = VS, Visual-Verbal = VV) followed by the N level of the task (i.e., 0, 2, 5).

Performance Metrics

Performance was measured in terms of accuracy and reaction time. Visualizations of the output metrics for performance on the N-back task battery can be seen in Figure 5. The results across both subjects were similar; as such, the average accuracy and reaction time between subjects are presented here. A general trend of decreasing accuracy with increased N-level is present, although more exaggerated for some tasks. Additionally, the reaction time results indicate a trend of longer reaction times during an N-level of 2 compared to N-levels of 0 or 5.

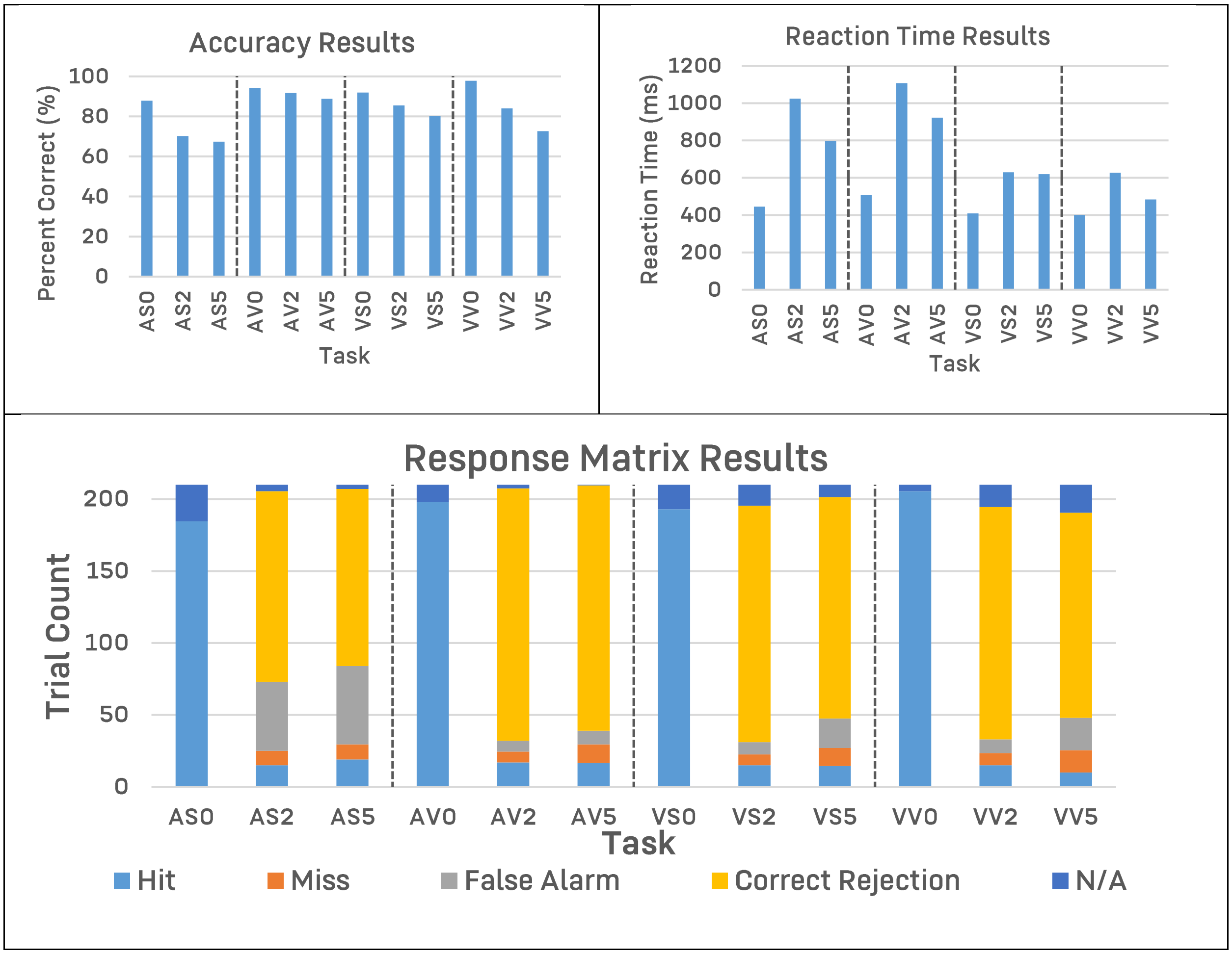


Figure 5. Performance data. Each task is named with an abbreviation of the task's name (i.e., Auditory-Spatial = AS, Auditory-Verbal = AV, Visual-Spatial = VS, Visual-Verbal = VV) followed by the N-level of the task (i.e., 0, 2, 5). Dotted lines separate each task set.

A response matrix was also tallied using the response data for each trial. Each response was categorized as a "Hit" (i.e., correct answer = Yes and subject response = Yes), "Miss" (i.e., correct answer = Yes and subject response = No), "False Alarm" (i.e., correct answer = No and subject response = Yes), "Correct Rejection" (i.e., correct answer = No and subject response = No), or "N/A" (i.e., no recorded response, equivalent to a "Miss" but separated for this evaluation stage). A trend of increasing misses and false alarms as the N-level increases is present for each task.

The collection of performance data did not reveal any major flaws in the data acquisition system. Timing of stimulus presentation was consistent, with the mean time between stimuli measured to be 2001.53 ms (*SD* = 0.68 ms) for the auditory tasks (AS and AV), 2038.32 ms (*SD* = 1.57 ms) for the visual-spatial (VS) task, and 2010.18 ms (*SD* = 1.09 ms) for the VV task. Note that the reported times take into account that each data collection system was connected to the one computer. The consistency in mean presentation times suggests minimal interference in task presentation with multiple recording systems running on one computer. Future efforts will be made to coordinate further the mean presentation time of each stimuli to 2000 ms.

Subjective Metrics

Visualizations of the output of the CSS are presented in Figure 6. Similar to the physiological results shown above, the subjects responded similarly on the subjective metrics employed. Average response scores are presented here. A general trend of increasing subjective appraisals of average and maximum experienced CWL is present for each task. Interestingly, fatigue levels stagnated or decreased with increasing N-levels.

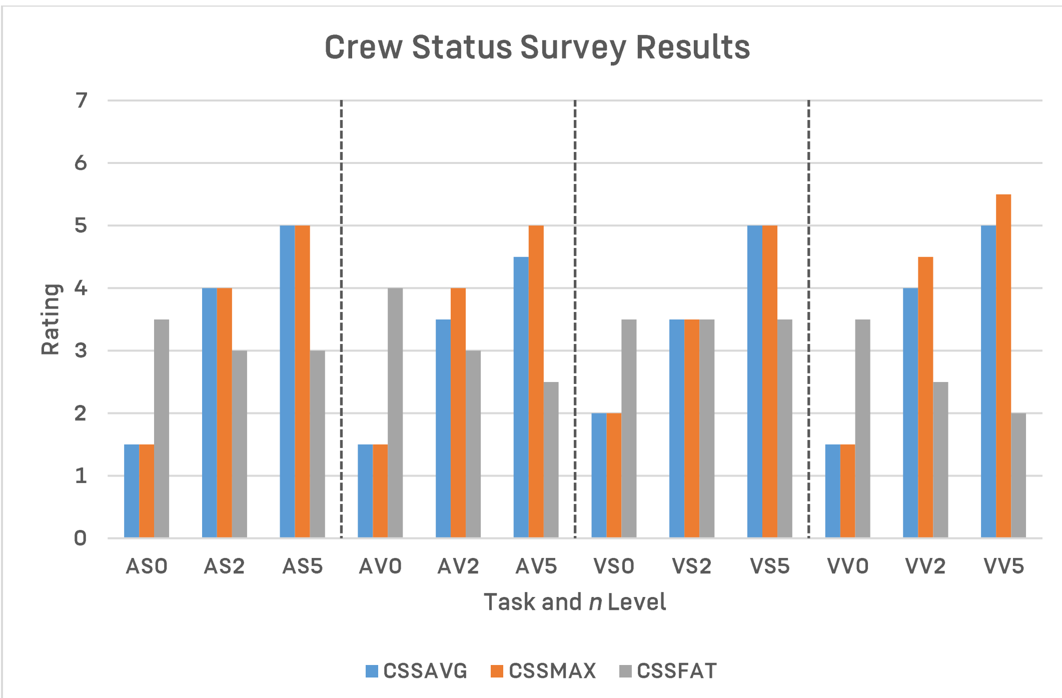


Figure 6. Subjective data from the Crew Status Survey.

Physiological Metrics

Descriptive statistics and simple visualizations for physiological measures are provided below to demonstrate data capture of CWL metrics as a function of task demand.

ECG.

The ECG signal for each data collection appeared to be free of major artifacts, as seen in a representative snapshot of the ECG trace (after being processed by the peak detection process detailed above) in Figure 7.

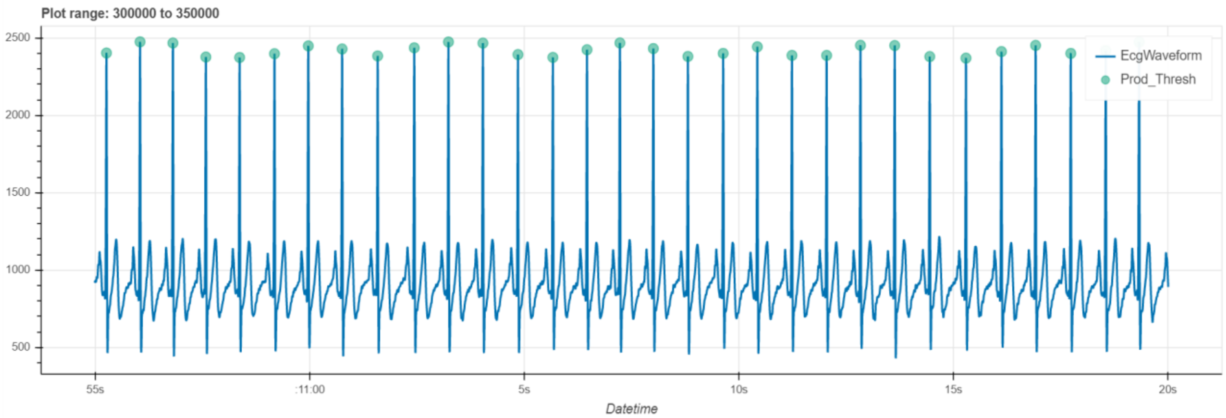


Figure 7. Example ECG trace output with marked peak values.

Using the ECG traces for each task configuration, mean heart rate, and heart rate variability across both subjects (as the results were very similar) were calculated using the output of the peak detection algorithm. No significantly noticeable changes in heart rate as a function of cognitive demand appear to be present in the descriptive statistics of the collected data, as seen in Figure 8. The derived RMSSD (heart rate variability) metric values can be seen in Figure 9. Again, no general trend appears to be apparent from this smaller sample size.

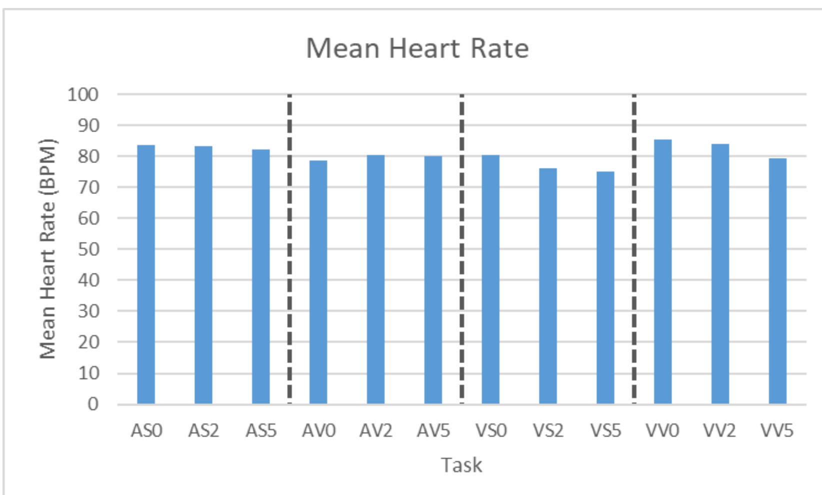


Figure 8. Mean heart rate data.

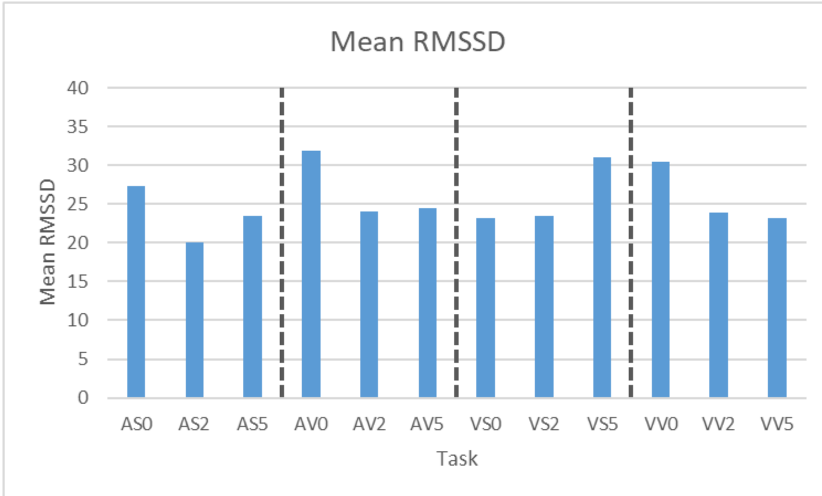


Figure 9. Mean RMSSD (heart rate variability) data.

Respiration.

The respiration data collected by the Biopac® system also appeared to be free of any major artifacts, as seen in a representative snapshot of the respiration trace (as displayed in Acq*Knowledge*) in Figure 10. The respiration rate across the middle 5 minutes of task engagement was calculated for each task configuration and is visualized in Figure 11. There appears to be a trend of increasing breaths per minute with increasing N-level across both subjects.

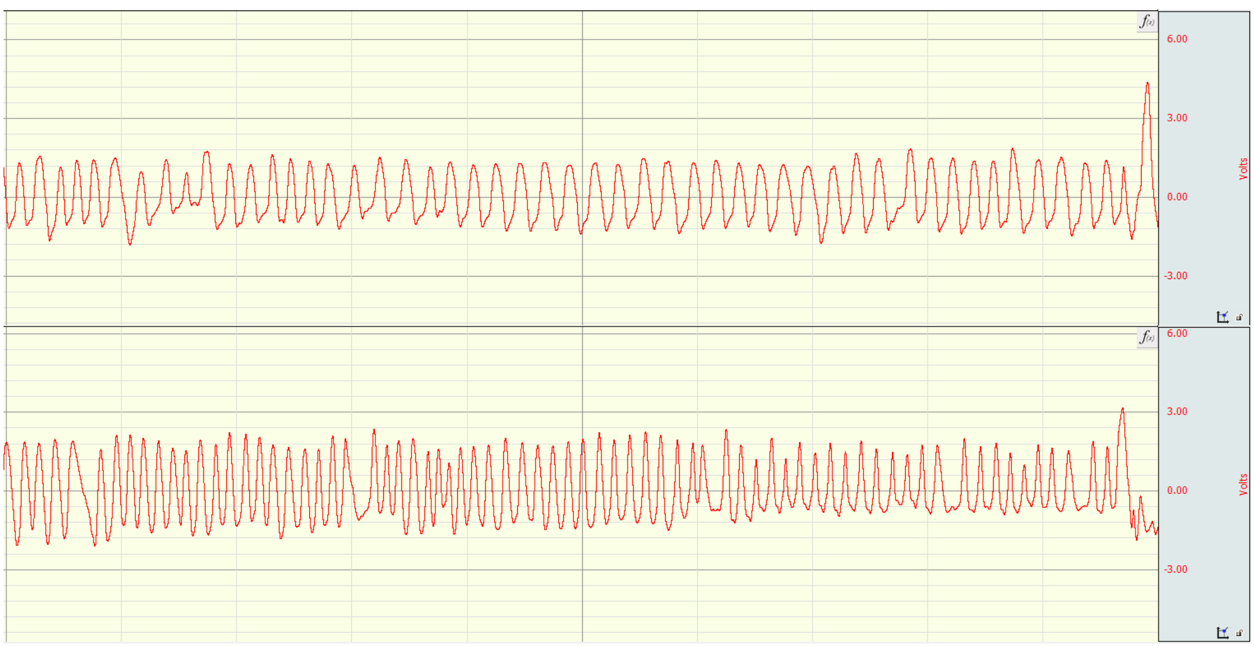


Figure 10. Representative snapshot of respiration trace. The top trace is from an AS0 session, and the bottom trace is from an AS5 session.

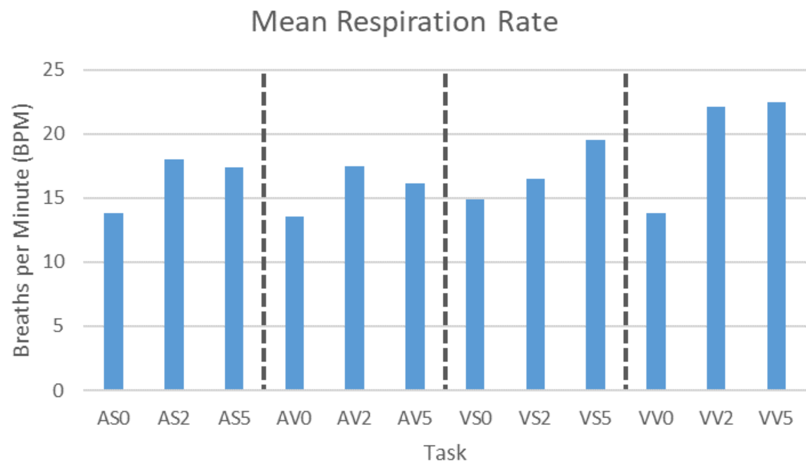
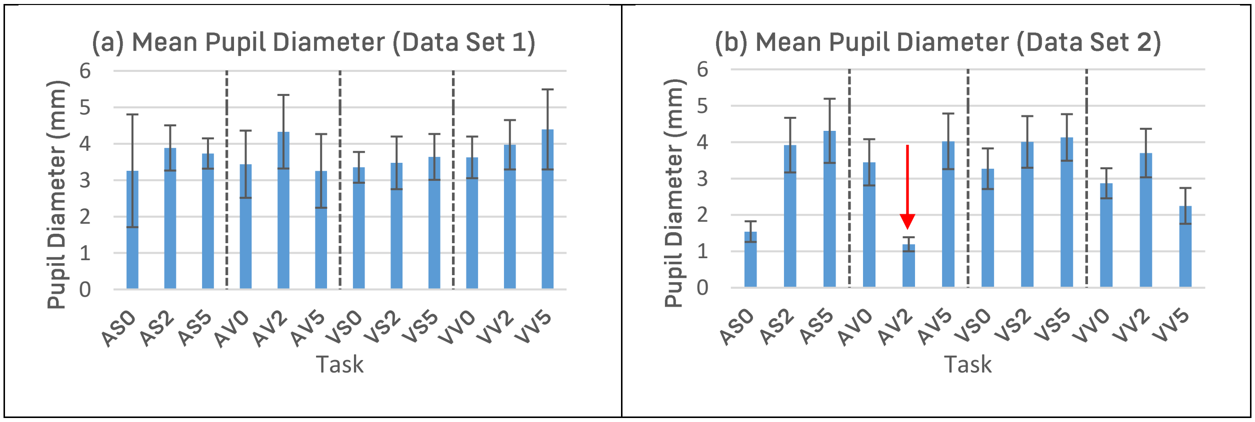


Figure 11. Mean respiration rate data.

Oculometrics.

Results of the mean pupil diameter for each subject across each task configuration (along with the mean results across both subjects) are visualized in Figure 12. Upon inspection of the recordings, it is clear that the AV2 session recorded in the second data set was inaccurate due to a calibration error in the eye model used by the eye tracking system. This error is marked with a red arrow in Figure 12. Mixed results are present in the pupil data, but, in some task configurations, the expected trend of increasing pupil diameter was observed with increasing levels of N. The general pupil size increase was also noted by the experimenters watching the live data stream as it was being recorded. A representative snapshot of the noted effect is presented in Figure 13.



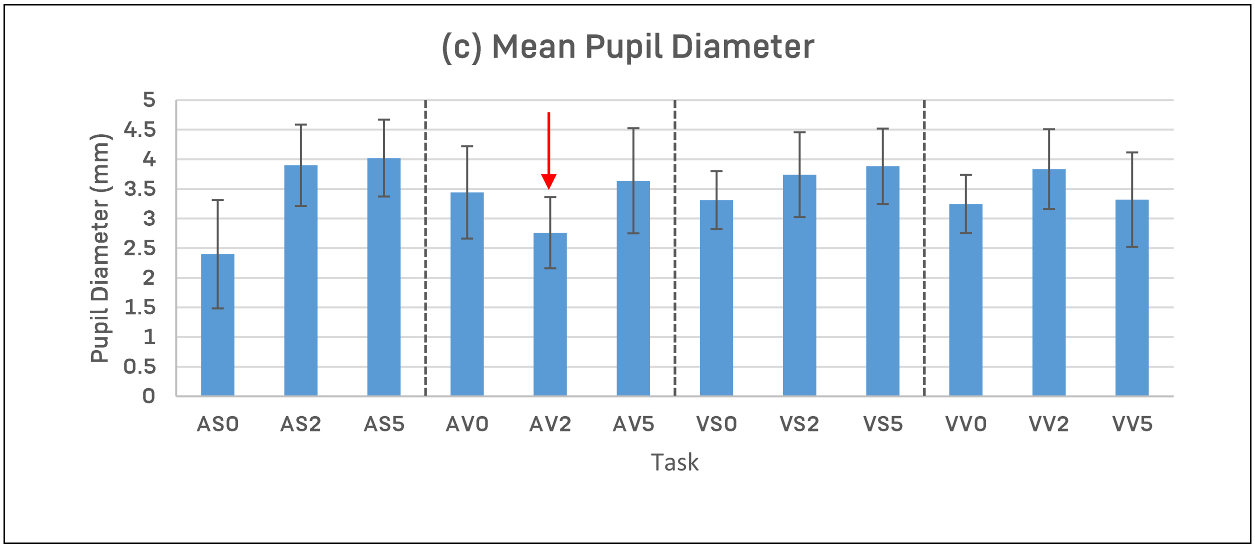


Figure 12. Pupil diameter results. Individual data sets are presented in plots a and b, while the average pupil diameter across subjects is presented in plot c. Note, the red arrow indicates a faulty recording due to an inaccurately calibrated eye model used by the recording system. This error suppressed the pupil diameter output for one data set, and overall the reported means.

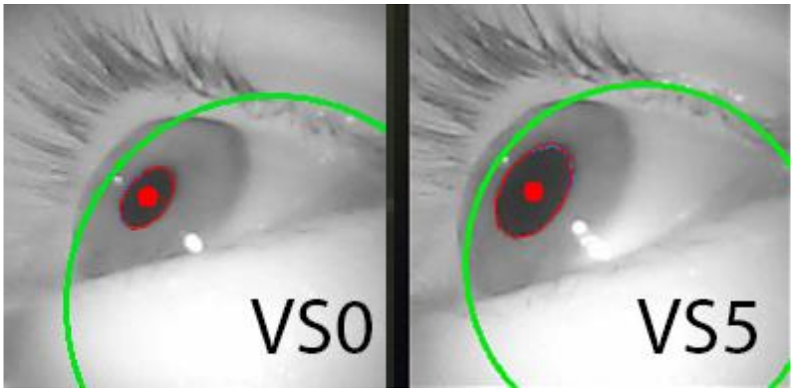


Figure 13. Representative snapshot of pupil size differences between the VS0 and VS5 task levels.

The average blink duration and blink count for each task configuration was also calculated for each subject (and across both subjects), and are visualized in the three axes plots in Figure 14. A noticeable difference in blink duration was present between subjects, with one subject, as seen in the duration differences between data sets 1 and 2 in Figure 14. The mean blink duration across subjects during the N = 0 tasks appears to be longer relative to each task configuration's accompanying N = 2 and N = 5 levels. The mean blink count data did not reveal any notable trends in terms of simple descriptive analysis.

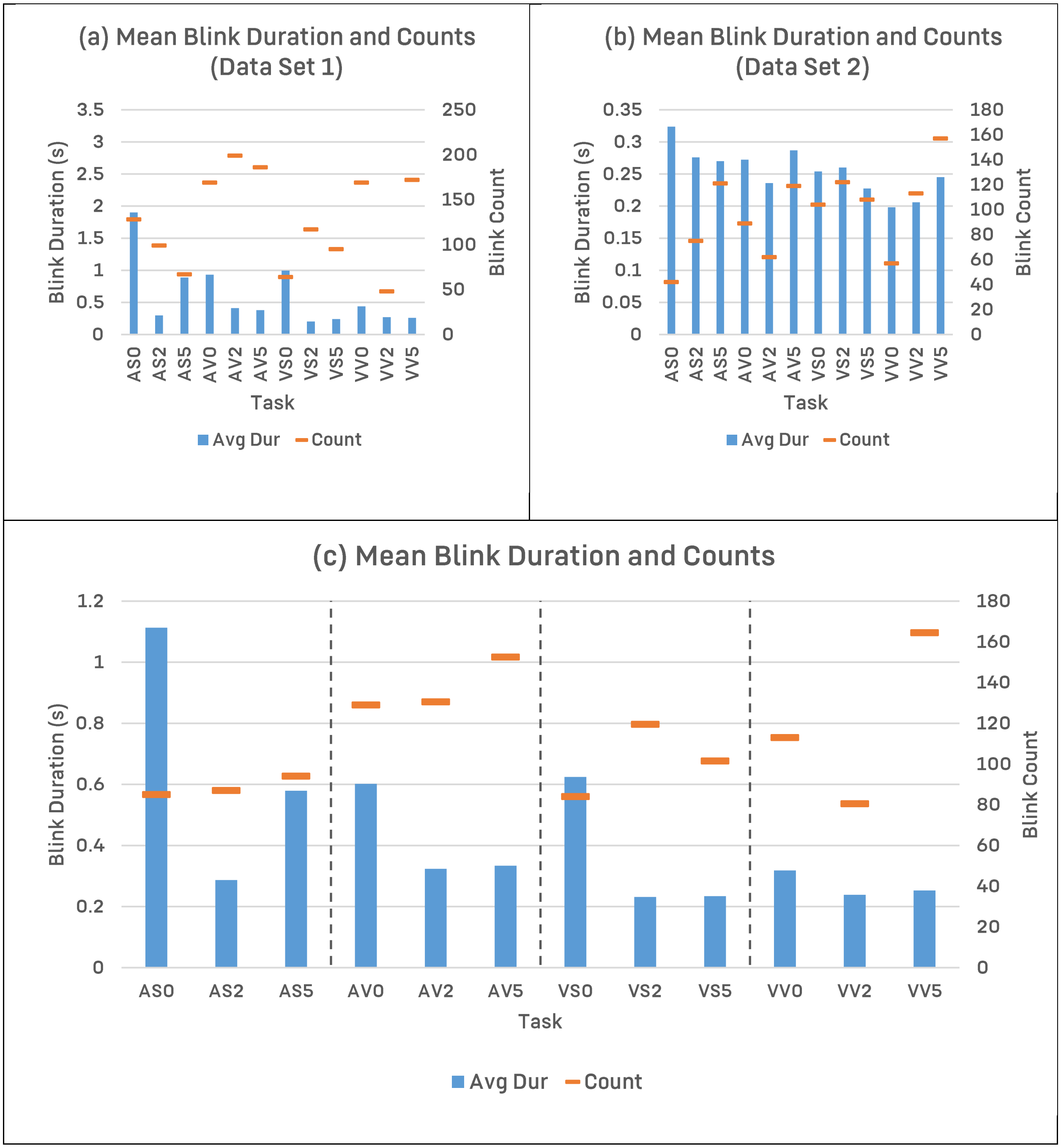


Figure 14. Average blink duration (blue bars) and count (orange points) results for each individual subject (a and b) and across both subjects (c).

Discussion

Due to the iterative nature of this study, small evaluations of the multiple components needed for task development, task presentation, physiological recording, and subjective appraisals are required to develop the large CWL database for this project. It is important to note that this is a simple evaluation of the task battery and sensors for the overall project, so the small sample size and unsynchronized data severely limit any statistical analysis of the CWL metrics presented in this report. However, in this evaluation, we have demonstrated the feasibility of the project's ambitious data collection methods and gained insight into potential areas of improvement.

The N-back task battery used in this evaluation was designed in a way to precisely tax specific cognitive resources, as defined by MRT. Results indicate that while all of the tasks follow the same N-back paradigm, there are potential differences in the CWL metrics based on the resource domains being taxed. Using performance metrics as an example, accuracy on the AS task was lower than nearly all other tasks and reaction times were far higher for auditory-based tasks than visual-based tasks. It is expected that more concrete differences between task configurations will emerge as more data is collected using the N-back task battery. Therefore, the multiple N-back battery configurations will remain as a task set in this project's future.

In a more classic approach to CWL, which divides tasks into a few different task demand levels, the expected drop in accuracy with increasing N-level (i.e., increasing task demand) was observed for each task configuration. Interestingly, the reaction time results indicate that an N-level of two required more time execute a response than an N-level of five. This is likely due to the N-level of two being perceived as more possible to complete successfully than an N-level of five (which was chosen due to its extremely high *perceived* difficulty). Thus, it is likely that the subjects felt more in control of the N = 2 task over the N = 5 task, and took the extra time to provide a more accurate response as opposed to defaulting to a single answer.

This brings up an important point in the design of the N-back task used in this evaluation. Due to the methods used to generate the stimulus lists presented to the subjects, it is far more likely that the correct response to any single trial is "No match." As seen in the response matrix results, the number of trials that were marked as "Correct Rejections" was far higher than the maximum number of possible "Hits" in the stimulus list. It is not surprising the number of "Misses" and "False Alarms" increases as the N-level increases. Once this strategy is understood by the subject, it becomes safer to respond to a trial they are unsure about with a "No match" response, as it will more often than not be correct. To alleviate this issue, the stimulus generation software will be modified to produce a specific percentage of "Match" responses as answers across a set number of trials. This will increase the perceived risk of defaulting to a single "No match" response by the subjects.

Looking at the subjective responses collected by the CSS, the reported fatigue levels dropped with increasing N-levels. This may indicate that the boredom experienced in the N = 0 configurations lead to higher levels of experienced fatigue compared to the more engaging feeling experienced with the N = 2 and 5 tasks. Indeed, the collected blink duration data aligns with the reported fatigue values, with longer blinks experienced during the $N = 0$ trials. It is also not surprising to see higher levels of reported average and maximum experienced workload with increasing levels of N. Due to the fact that the subjects knew which N-level they were completing, it would be easy to have that knowledge bias their subjective workload responses. Moving forward, a more subtle method will be used to modulate task demand in the N-back battery by modulating the stimulus size and inter-trial intervals rather than the gross steps of N-level, alone.

Collection of the physiological variables was successful. The data collected in the current evaluation were free of any major artifacts. One issue was experienced with the eye tracking calibration for one trial with one subject. Methods to ensure proper calibration and/or recalibration with a standard or previously calibrated model (by the same subject) in post processing will be explored. In some cases, the physiological data appeared to follow the standard changes expected with increasing task demand. The mean pupil size appeared to increase with increasing N-level, an increase that was even noted by experimenters while monitoring the data collection in real-time. The average respiration rate appeared to increase with task demand for both subjects. Trends in the other physiological data collected, including heart rate, heart rate variability, blink duration, and blink count were less clear. However, this is likely due to faults on the small sample size and aspects of the experimental set up that will require changes prior to future iterations of this study. Future analyses using the same approach are warranted.

Conclusion

The long-term goal of this work is to provide the foundation for a system that integrates into the future vertical lift (FVL) platforms (and other systems) to conduct real-time operator state monitoring during flight. There is not currently literature adequate to determine whether a high-accuracy model to predict workload from psychophysiological recordings requires a separate model for each individual or a generalized model that works for most individuals. The purpose of the overarching research program is to systematically explore the entire spectrum of cognitive workload as a function of task demand to generate a large dataset to be used by computer algorithms to objectively quantify cognitive workload. While few conclusions can be drawn from actual data collected in this small data collection effort (due to the small sample size), the picture painted by the collected results indicates we are on the right track in our iterative design efforts. The collected CWL metrics appeared to be sensitive to changes in task demand and will stand as a manufacturer-preferred dataset for later comparison with data synchronization efforts in future evaluations. Most importantly to the iterative design of the larger goals of this study, areas of improvement were identified and potential solutions were suggested.

References

Ames, L. L., & George, E. J. (1993). *Revision and verification of a seven-point workload estimate scale* [Report]. U.S. Air Force. https://apps.dtic.mil/sti/pdfs/ADA269194.pdf

Baddeley, A. D. (1986). *Working Memory* [Book]. Clarendon Press. https://books.google.com/books?id=ZKWbdv__vRMC

Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In *Models of working memory: Mechanisms of active maintenance and executive control.* (pp. 28-61). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.005

Biopac. (2021). *AcqKnowledge.* In (Version 5.1) Biopac. https://www.biopac.com/product/acqknowledge-software/

Cain, B. (2007). *A Review of the Mental Workload Literature* [Report](RTO-TR-HFM-121-Part-II). Defence Research and Development Canada.

Casner, S. M., & Gore, B. (2010). *Measuring and Evaluating Workload: A Primer* [Report](NASA/TM—2010-216395). National Aeronautics and Space Administration. https://matb-files.larc.nasa.gov/Workload_Primer_TM_Final.pdf

Champseix, R. (2018). *hrv-analysis 1.0.4.* In (Version 1.0.4) pypi.org. https://pypi.org/project/hrv-analysis/

de Waard, D. (1997). *The Measurement of Drivers' Mental Workload* [Thesis, University of Groningen]. Groningen, Netherlands. https://pure.rug.nl/ws/portalfiles/portal/13410300/09_thesis.pdf

Hancock, P. A., & Matthews, G. (2019). Workload and Performance: Associations, Insensitivities, and Dissociations. *Human Factors, 61*(3), 374-392. https://doi.org/10.1177/0018720818809590

Hebb, D. O. (1955). Drives and the C.N.S. (conceptual nervous system). *Psychological Review, 62*(4), 243-254. https://doi.org/10.1037/h0041823

Kahneman, D. (1973). *Attention and Effort* [Book]. Prentice-Hall. https://books.google.com/books?id=tkvuAAAAMAAJ

Kassner, M., Patera, W., & Bulling, A. (2014). Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication,

Korbach, A., Brünken, R., & Park, B. (2018). Differentiating different types of cognitive load: A comparison of different measures. *Educational Psychology Review, 30*(2), 503-529. https://doi.org/10.1007/s10648-017-9404-8

Meister, D. (1976). *Behavioral foundations of system development*. John Wiley & Sons.

Naismith, L. M., Cheung, J. J. H., Ringsted, C., & Cavalcanti, R. B. (2015). Limitations of subjective cognitive load measures in simulation-based procedural training. *Medical Education, 49*(8), 805-814. https://doi.org/10.1111/medu.12732

O'Brien, K. J., St. Onge, P. M., & Irvin, W. D. (2020). *A Peak Detection Algorithm for*

Identifying R-Waves in High Noise Electrocardiogram Recordings [Report]. U.S. Army Aeromedical Research Laboratory.

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance.* (pp. 1-49). John Wiley & Sons.

Seeber, K. G. (2007). *Thinking outside the cube: modeling language processing tasks in a multiple resource paradigm* http://www.isca-speech.org/archive/interspeech_2007/i07_1382.html

Shaffer, F., & Ginsberg, J. (2017). An overview of heart rate variability metrics and norms. *Frontiers in public health*, *5*, 258.

Teigen, K. H. (1994). Yerkes-Dodson: A Law for all Seasons. *Theory & Psychology, 4*(4), 525-547. https://doi.org/10.1177/0959354394044004

The pandas development team. (2020). (Version Latest) Zenodo. https://doi.org/10.5281/zenodo.3509134

Vack, N., & Ollinger, J. (2020). *bioread 2.1.2.* In (Version 2.1.2) pypi.org. https://pypi.org/project/bioread/

Wickens, C. D. (1984). Processing resources in attention. *Varieties of attention, 40*, 63-98.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science, 3*(2), 159-177.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*(5), 459-482. https://doi.org/https://doi.org/10.1002/cne.920180503

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*(1), 1-17. https://doi.org/10.1080/00140139.2014.956151

Appendix A. Acronyms and Abbreviations.

3D	Three Dimensional
AS	Auditory Spatial
AV	Auditory Verbal
BPM	Beats Per Minute
CWL	Cognitive workload
DV	Dependent Variable
ECG	Electrocardiogram
EEG	Electroencephalogram
EOG	Electrooculogram
IBI	Inter-beat Interval
ICA	Index of Cognitive Activity
IIR	Infinite Impulse Response
IV	Independent Variable
JND	Just Noticeable Difference
LOS	Line of Sight
MATB	Multi-Attribute Task Battery
MRT	Multiple Resource Theory
USAARL	U.S. Army Aeromedical Research Laboratory
VS	Visual Spatial
VV	Visual Verbal

U.S. Army Aeromedical Research Laboratory
Fort Rucker, Alabama

All of USAARL's science and technical information documents are available for download from the Defense Technical Information Center.

https://discover.dtic.mil/results/?q=USAARL

Army Futures Command
U.S. Army Medical Research and Development Command