Towards Robust Estimation of Cognitive Workload from Wearable Physiological Sensors

Aaron Novstrup[1], Monica Tynan[2], Joshua Kline[3], Gianluca De Luca[4], James Heaton[5],
Stottler Henke Associates[1], Wellman Center for Photomedicine[2], Delsys/Altec[34],
Massachusetts General Hospital Voice Center[5]
Seattle Washington[1], Boston Massachusetts[2345]
anovstrup@stottlerhenke.com[1], mtynan2@partners.org[2], jkline@delsys.com[3], gdeluca@delsys.com[4],
James.Heaton@mgh.harvard.edu[5]

ABSTRACT

Reliable, objective, and timely estimation of cognitive workload has potential applications in training (e.g., facilitating curriculum development), human performance assessment (e.g., treating workload itself as a performance metric), the design and development of human-automation teaming systems (e.g., evaluating the impact of design choices on operators' cognitive workload), and adaptive automation (i.e., adapting automation behavior based on the cognitive workload of human operators). A wide variety of physiological indicators of cognitive workload have been investigated over the past five decades, including heart rate/variability, respiratory measures, pupil size and other pupillometrics, electrodermal activity (EDA), and indicators extracted from complex sources such as functional near-infrared spectroscopy and electroencephalography. However, individual physiological indicators are non-specific to workload and must be combined with others to derive a useful estimate. The sensitivity and specificity of joint estimates depend on the sensitivities of the individual indicators to variations in cognitive workload and the unique information contributed by each.

This paper explores the utility of *face and neck surface electromyography* (fnsEMG)—non-invasive, skin surface measurement of the motor action potentials that drive muscle activity—as a sensing modality for cognitive workload and its associated emotional responses. The sensitivity of fnsEMG to cognitive workload variations at nine face and neck sensing locations was established in a Defense Advanced Research Projects Agency funded human study in which participants performed multiple concurrent cognitive tasks in a modified version of the NASA Multi-Attribute Task Battery. Task performance and frequent self-reports of task difficulty were compared with multiple physiological signals, including fnsEMG, electrocardiography, EDA, respiration, eye gaze, and pupil size. Non-parametric methods were used to predict task errors and self-reported task difficulty based on these physiological signals. Calibrated predictions on a held-out test set demonstrated the combined sensitivity of these measures, and of fnsEMG in particular, to cognitive workload and overload.

ABOUT THE AUTHORS

Aaron Novstrup is an applied artificial intelligence researcher at Stottler Henke Associates, Inc., where he develops intelligent software systems. His current research focuses on applying machine learning and statistical modeling techniques to automated human state/performance assessment and monitoring.

Monica A. Tynan is a research technologist at the Wellman Center for Photomedicine at Massachusetts General Hospital (MGH) in Boston, MA. She received her B.S. in Biological Sciences from the University of Rhode Island in 2017 and her MSc in Applied Neuroscience from King's College London in 2022.

Joshua Kline is a biomedical engineer with over ten years of R&D experience. After receiving a Ph.D. from Boston University, he joined Delsys, Inc. and Altec, Inc. to advance sensor technology for health and human performance.

Gianluca De Luca is an electrical engineer and Vice President of Product Development at Delsys, Inc. and Altec, Inc., developing wearable sensor technologies for health care and human performance applications.

James T. Heaton, Ph.D., is Director of the Laryngeal Surgery Research Laboratory at MGH, an Adjunct Professor at the MGH Institute of Health Professions, and an Associate Professor of Surgery at Harvard Medical School. His research interests include voice and speech physiology focusing on using face and neck electromyography for automatic speech recognition and cognitive workload assessment.

Towards Robust Estimation of Cognitive Workload from Wearable Physiological Sensors

Aaron Novstrup[1], Monica Tynan[2], Joshua Kline[3], Gianluca De Luca[4], James Heaton[5],
Stottler Henke Associates[1], Wellman Center for Photomedicine[2], Delsys/Altec[34],
Massachusetts General Hospital Voice Center[5]
Seattle Washington[1], Boston Massachusetts[2345]
anovstrup@stottlerhenke.com[1], mtynan2@partners.org[2], jkline@delsys.com[3], gdeluca@delsys.com[4],
James.Heaton@mgh.harvard.edu[5]

INTRODUCTION

The applications for a robust, objective, and timely measure of *cognitive workload*—understood roughly as the portion of a person's limited cognitive resources consumed during the performance of a task in a particular environment—include training, human performance assessment, the design and development of human-automation teaming systems, and adaptive automation (i.e., automation that adapts its behavior to the changing needs of human operators). In training, for example, a measure of cognitive workload could indicate task mastery independent of performance measures, as trainees can achieve a fixed level of performance with decreasing cognitive effort as mastery is acquired. It could also facilitate curriculum development by giving insight into trainees' cognitive workloads as they experience various curriculum designs.

Although the research community continues to debate nuances of the construct, there is general agreement that cognitive workload: a) is multi-dimensional (Matthews et al., 2015), b) relates task demands (i.e., *taskload*) and cognitive task performance (Cain, 2007), and c) is associated with taskload by a function of the cognitive resources available to the individual for performance of the tasks at hand—resources in turn determined by many factors, including the individual's acquired skill and natural aptitude for the tasks. The effect of cognitive workload on task performance is complex and non-linear, making performance metrics poor indicators of cognitive workload in general. Researchers have demonstrated some success using secondary task performance measures as indicators of cognitive workload induced by a primary taskload (Kaber and Riley, 1999). However, this approach has the significant drawback that adding secondary tasks can interfere with the performance of primary tasks. Aside from its influence on task performance, cognitive workload has observable effects on subjective instruments, such as the NASA Task Load Index (NASA-TLX) (Hart, 2006; Hart and Staveland, 1988), and various physiological variables.

Prior research has identified several somatic and autonomic nervous system outputs (human behaviors) that change predictably when individuals are placed under different taskloads (Cain, 2007; Rosenberg and Ekman, 2020; Quatieri et al., 2017). Taskload-dependent autonomic responses include an electrodermal response (sweat gland secretion), skin temperature fluctuation (vasculature smooth muscle tone), blood pressure and pupillary aperture (vascular and pupillary smooth muscle tone), and heart rate (cardiac muscle rhythm). Taskload-dependent variation in multiple volitional or reflexive skeletal-muscle behaviors, including changes in facial expression, jaw tension, body posture, voice and speech patterns, and manual task execution (e.g., movement rate, accuracy, coordination, etc.), has also been demonstrated. Aside from our prior retrospective assessment of neck intermuscular beta coherence (Novstrup, Goan, and Heaton, 2019), we are unaware of previous attempts to explicitly connect face and neck surface electromyography (fnsEMG) to cognitive workload or volitional responses to the same.

The human study presented in this paper demonstrates both an explicit connection between cognitive workload and fnsEMG and the ability to predict cognitive workload from a complement of physiological measures derived from multiple sensing modalities (including fnsEMG). The physiological measures considered in this study—heart rate (HR) and heart rate variability (HRV) from electrocardiography (ECG), respiration rate, electrodermal activity (EDA) measures, pupillometrics, and fnsEMG—could, in principle, be collected from relatively unobtrusive wearable sensors, appropriate for practical application in many real-world work/task environments. The sensitivity of fnsEMG to taskload variations is demonstrated, along with the ability to predict both a task performance measure and a subjective self-assessment of task difficulty. The next section presents the methods employed, followed by a section describing the results, and then the paper concludes with discussion of the study's implications and directions for future work.

METHODS

Study Design and Enrollment Overview

All data collection for this study occurred at Massachusetts General Hospital (MGH). Eleven healthy adults of both sexes (5 males and 6 females), ages 21 – 28, were recruited to visit the MGH Center for Laryngeal Surgery and Voice Restoration ("MGH Voice Center") for one study visit per participant, each lasting approximately 5 – 6 hours. All participants provided written informed consent. The study protocol was reviewed and approved by the MGH Institutional Review Board. Study participants were recruited from recipients of MGH broadcast emails describing research subject opportunities and were screened according to the selection criteria described below.

A primary focus of the study was performance on the computer-based Multi-Attribute Task Battery (MATB; see Enhancements section below). Participants were required to be literate and fluent in English since MATB is only available in English. Participants were also required to have adequate vision, hearing, and hand motor control for performing the MATB and to reach a minimal baseline performance on the MATB (see Study Procedures section) for full participation. All pre-screened individuals who visited the Voice Center qualified for full participation.

We developed a modified version of the MATB that delivers noxious shocks to the forearm as punishment for some task errors (see Study Procedures). Individuals with a cardiac pacemaker or other electrically sensitive implants were excluded from participation due to possible interaction with skin-surface electrical stimulation. In addition, individuals particularly fearful of electrical stimulation or with low pain threshold for cutaneous sensation were excluded, along with those reporting skin sensitivity to medical-grade adhesives like those used in stimulation and recording electrode application. Finally, individuals were advised not to enroll if pregnant, and tests were provided if pregnancy status was unknown.

Multi-Attribute Task Battery (MATB) Enhancements

The present research protocol aimed to identify cognitive workload during complex tasks through unobtrusive physiological measures reflecting mental/emotional state. A modified version of the MATB was used to evaluate research participants' performance and degree of cognitive workload through several simultaneous computer-based tasks. The MATB provides a set of tasks analogous to what aircraft crewmembers perform in flight, yet it does not require piloting experience. The MATB involves the simultaneous performance of manual aircraft flight (joystick control), aircraft system monitoring, dynamic resource management, and communication tasks, and it has been used in numerous experiments of cognitive demand and performance since initially developed in the early 1990s (Comstock & Arnegard, 1992; Santiago-Espada et al., 2011). The custom software implementation used in this study, derived from the Air Force MATB software (AF-MATB v4.9), had three principal modifications: 1) The addition of real-time performance feedback and shock punishment, 2) utilization of an automated joystick providing assistive and disruptive steering, and 3) incorporation of a computerized implementation of the Instantaneous Self-Assessment of Workload (ISA) (Tattersall and Foord, 1996; Jordan and Brennen, 1992).

The cognitive tasks provided by the AF-MATB reflect many of the cognitive challenges faced by pilots. However, this testbed does not represent a high-fidelity flight simulation or provide consequences or feedback for task errors, aside from performance scores at the end of testing. Therefore, to increase participant engagement, self-awareness, and motivation, we modified the MATB software to provide real-time performance feedback through auditory cues for correct/incorrect responses. For example, a rewarding "ding" sound played after each correct response to the System Monitoring gauges and lights, and a harsh buzzing sound occurred when the user was out of range in the Tracking task (see Study Procedures section below). In addition, our modified MATB platform brought error consequences under experimental control using methods like those detailed in Lindstrom et al. (2013). Briefly, aversive consequences in the form of individually calibrated unpleasant (but not painful) electrical shocks were administered after some task errors (i.e., limited to one shock per 20 seconds of task engagement). Feedback of correct/incorrect responses and error punishment were expected to reduce the attenuation of physiological responses to cognitive workload observed in simulated work environments (e.g., Angelborg-Thanderz, 1990).

Brief (100 ms) electrical shocks were administered to the forearm skin at a voltage that participants reported as being uncomfortable but not painful, as determined in a shock level calibration procedure conducted prior to the first recorded test trial (approximately 30-70V for 100 ms monopolar pulses; see Lindström et al., 2013). Shocks were delivered using two Ag/AgCl disposable 20mm diameter circular electrodes (Natus Medical) placed on the left dorsolateral forearm skin approximately 5 cm apart. Shock voltage was controlled by a constant-voltage stimulation module (BIOPAC STM200) and delivered current was monitored by a BIOPAC MP160 data acquisition system. Shock discomfort level was assessed using an 11-point (zero-to-ten) pictographic faces pain scale (Figure 1; Hicks et al., 2001) at the time of shock calibration and after each of the 15 trials, with the expectation that participants would report discomfort greater than zero but no greater than 7 (see the dotted vertical line on Figure 1), since levels 8 – 10 would be considered painful and therefore exceed the intended shock discomfort for this study. Participants could adjust the shock voltage up or down between trials if the shocks became ineffective or too uncomfortable, respectively. The average shock level at the start of Trial 1 was 51.9V (SD 11.8V). Throughout data collection, two participants requested a reduction in shock level (average -8%), and five participants asked for one or more increases in shock level (average +11%). Reported shock discomfort levels ranged from two to seven after the initial calibration and averaged 4.6 (SD 1) at the beginning of Trial 1. Throughout the 15 trials, 9 of 11 participants reported changes in discomfort. In most cases, changes were *decreases*, prompting the participants to request an increase in shock voltage, which then increased reported discomfort in subsequent trials.

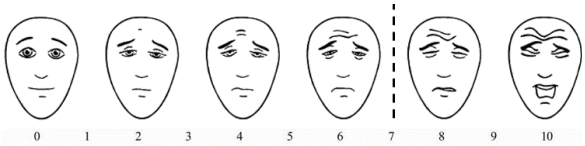
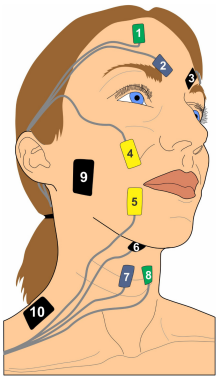


Figure 1. Faces Pain Scale (Hicks et al., 2001)

Physiological Measures

In addition to task performance quantified by the computer software, we hypothesized that physiologic measures of somatic and autonomic nervous system function would reflect the cognitive demands and emotional state of research participants as they performed the MATB. Signal acquisition in this study included sEMG, ECG, EDA (also called galvanic skin response or GSR), respiration, pupillometry/eye gaze, and microphone (audio) recording. Physiological signals were digitized using a BIOPAC MP160 system (ECG, EDA, respiration, microphone, shock current; 20ks/s) synchronized with a Delsys Trigno system (11 channel EMG; 2ks/s) and additional high-fidelity audio (Tascam DR40; 48ks/s) via a headset microphone (Sony ECM-66B) positioned 5cm from the mouth. Eleven reusable Delsys Trigno EMG sensors were applied to the skin surface above targeted muscle groups using double-sided adhesive, and signals were transmitted wirelessly to a Trigno base station. The skin at EMG sensor locations was prepared by alcohol wipe cleaning and 'peeling' exfoliation using common, clear desk tape (Stepp, 2012). Target locations are shown in Figure 2. Eight locations (1 – 8 in Figure 2) used "Mini" Trigno sensors designed for face and neck surface recording, while three locations (9 – 11 in Figure 2) used standard Delsys Trigno EMG sensors. ECG and respiration signals were digitized and transmitted wirelessly to the MP160 using a BioNomadix transmitter.

Three disposable ECG sensors were positioned on the right and left clavicular surfaces and below the left floating ribs on the ventrolateral abdomen (cathode, ground, and anode, respectively) to form Einthoven's triangle (Dupre et al., 2005). Respiration-related thoracic movements were obtained with a transducer band (BIOPAC BN-RESP-XDCR) positioned horizontally across the pectoral region (chest). Two disposable EDA sensors were applied to the distal phalanges of the middle and index fingers of the left hand and digitized using a BIOPAC MP36R module attached to the MP160 (see Braithwaite et al., 2013). These signals enabled skin conductance level and conductance response measurements, capturing both tonic and phasic components of sympathetic neuronal activity, respectively. A Gazepoint GP3 high definition (HD) eye tracker was mounted to the monitor displaying the MATB cognitive task software graphical user interface (GUI). The eye tracker uses infrared (IR) light to illuminate the eyes of the participant and a machine-vision IR camera to resolve the location of the head/eyes, pupil diameter, and eye gaze/fixation coordinates. The eye tracker was individually calibrated early in the data acquisition session.



Sensor	Targeted Muscle(s)
1	Frontalis
2, 3	Corrugator supercilii
4	Zygomaticus major
5	Depressor anguli oris
6	Anterior digastric, genioglossus, mylohyoid
7, 8	Platysma, omohyoid, sternohyoid, sternothyroid
9	Masseter
10	Trapezius
11	Extensor carpi radialis longus (see Fig. 4)

Figure 2. Surface Electromyography (sEMG) Sensor Locations

Study Procedures

After participants provided written consent, sensors and electrodes were attached in preparation for data collection. The disposable electrode pairs for shock delivery were placed first to provide time for equilibration between the forearm skin and conductive adhesive. Next, the sEMG, ECG, and EDA sensors were placed, along with the respiration band and headset microphone. Sensor application took approximately 45 minutes, after which the shock stimulation was calibrated using an adaptive staircase procedure (Treutwein, 1995). Starting from a minimal voltage, an experimenter iteratively administered a 100 ms pulse and asked the subject to rate the sensation using the pain scale (Figure 1) described previously. The calibration process terminated at an estimate of the highest voltage the subject experienced as unpleasant but not painful.

An early step in the data collection protocol was to elicit muscle contractions relating to each sEMG sensor location to confirm each sensor's location and signal fidelity and provide a representative, vigorous contraction for muscles targeted by each sensor. For example, activity was elicited from the sensor on the right forehead (targeting the frontalis muscle) by asking the participant to raise their eyebrows to demonstrate surprise. Although the instructions for eliciting facial expressions and other muscle contractions were intuitive and practical, the authors concluded that having participants mimic modeled facial expressions would increase consistency in elicited contractions and provide a better reference contraction with which to compare spontaneous expressions recorded during MATB Trials. For this purpose, one male and one female model were selected from the Warsaw Set of Emotional Facial Expression Pictures (Olszanowski et al., 2015). This resource provided high-quality photographs of facial expressions receiving the highest recognition accuracy among their study participants, with the greatest intensity and purity of the target emotions expressed. Four participants of the present study were presented with PowerPoint slides of joy, disgust, surprise, anger, and sadness, with each emotional expression separated by a neutral expression. They were asked to mimic these neutral and emotional expressions for approximately 1 – 2 seconds as each slide was presented, thereby generating five emotional expressions separated by neutral expressions. These expressions did not add appreciable time to the recording sessions, and they elicited similar fnsEMG signals across participants. An example of the fnsEMG signals from one study participant (#8) for one set of the five mimicked emotional expressions is provided in Figure 3.

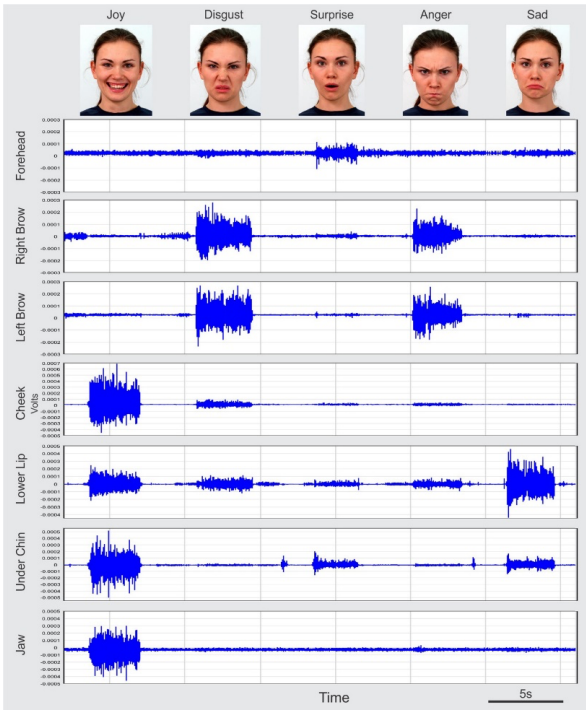


Figure 3. sEMG from 7 face and neck locations while mimicking the 5 emotional facial expression photographs (separated by neutral expressions — not shown). Sensor locations are 1 – 6 and 9 shown in Fig. 2. Photos are from Olszanowski et al., 2015.

Participants were trained on the MATB platform using written instructions and four training scripts (routines) designed to introduce each task. All other tasks were suspended when practicing each task independently. They first practiced the **Tracking task**, learning to steer a target circle toward the center of a reticle (cross-hairs) with the joystick. They experienced the harsh buzzer sound emitted when the circle drifted beyond the outer boundary of the reticle region and felt how the joystick could assist in the steering process via the integrated motors. They were not shown how the joystick could intermittently steer them off course to maintain automation failure surprise. Participants next practiced the **System Monitoring task**, with instructions regarding the normal versus error states for the gauges and lights. They practiced correcting the error states and experienced hearing the rewarding ding sound when providing correct responses and the error sounds associated with missed error states (a buzz followed by the announcement of "lights" or "gauges") or false keystrokes for gauge or light correction (phone off-the-hook sound). Tracking was suspended while practicing the system monitoring task. Participants then practiced the **Communication task**, learning how to change the radio channel and frequency when hearing them announced by the control

tower operator. They learned that correct radio changes resulted in a ding sound, whereas missed radio requests or incorrect inputs resulted in "communication" or "radio error" sounds, respectively. Auditory feedback was identical in training and testing trials. The final training script showed participants how the MATB tasks would pause every 35 seconds for a 15-second period, during which time they were prompted to provide an **instantaneous self-assessment** (ISA) of task difficulty. They would hear a recording of "Please report task difficulty" and would respond by saying, "This is NGT504, task difficulty is _____", choosing among five discrete difficulty levels (Very Low, Low, Fair, High, Very High). They learned to press a button on an ISA Unit (see Figure 4), which sent a signal to the MP160 unit indicating the selected difficulty level. They practiced saying the ISA phrase clearly with moderate amplitude without rushing or yelling, and having it fall within the 15-second pause in tasks. Training script completion took approximately 28 minutes.

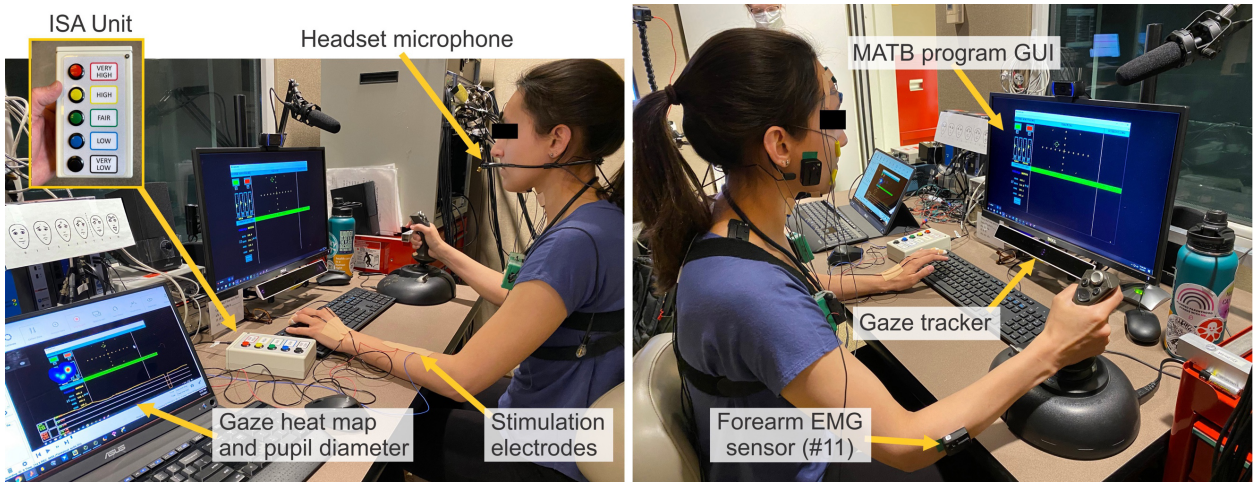


Figure 4. Recording Setup Showing a MATB Trial with Photo Insert of the ISA Reporting Unit

After completing the four training scripts, participants completed one practice trial where all tasks were presented simultaneously in the usual manner. Acoustic feedback of correct responses and task errors was also provided, but without shock punishment for errors (described below). Participants were expected to perform at or above a minimal performance baseline on the practice trial to qualify for full study participation, which all enrolled individuals achieved. This minimum was scoring 92 or above on a 100-point task performance scale in at least one of seven 35-second segments in the practice trial. This threshold was the performance level needed for automatic advancement in segment difficulty across consecutive segments in each trial (as described below). It was therefore an important performance level for experiencing the dynamic, incremental increase of task difficulty intended to manipulate cognitive workload. Practice trial completion took approximately six minutes.

After the practice trial, participants completed three testing *blocks*, each consisting of five 350-second *trials* (see Figure 5). Participants were given short breaks (2 – 4 minutes) between trials (while the data acquisition hardware/software and the MATB software were reset) and longer breaks (10 – 22 minutes) between blocks to stretch, use the restroom, etc. Trials were further subdivided into seven *segments*, each consisting of a 35-second *task segment* and a 15-second *ISA pause*. Completion of the three blocks with breaks took approximately 2.5 hours.

Task difficulty was manipulated across segments with the MATB Tracking task difficulty parameters, which determine the rate of the target's (apparently) random drift and the frequency with which the drift direction changes. These parameters were held constant throughout each segment. They were incremented from one segment to the next within a trial *provided the participant had achieved at least 92 points on a 100-point task performance scale in the preceding segment*, to provide a progressively challenging set of tasks. (The performance score was based on the percentage of time in which the Tracking target was within a pre-determined "in-range" threshold of the targeting reticle and was adjusted downward for discrete errors on the other tasks, such as failing to respond to a gauge event within a pre-determined time limit or entering an incorrect frequency in response to a communications prompt.) These difficulty parameters were "partially reset" between trials, to the highest difficulty level at which the participant had achieved a task performance score of at least 96 in any previous segment in the block, to ensure that the participant

started each trial at a manageable level of difficulty without underloading the participant. The trial difficulty was reset to the easiest condition at the beginning of each of the three testing blocks (see "Full difficulty reset" in Figure 5).

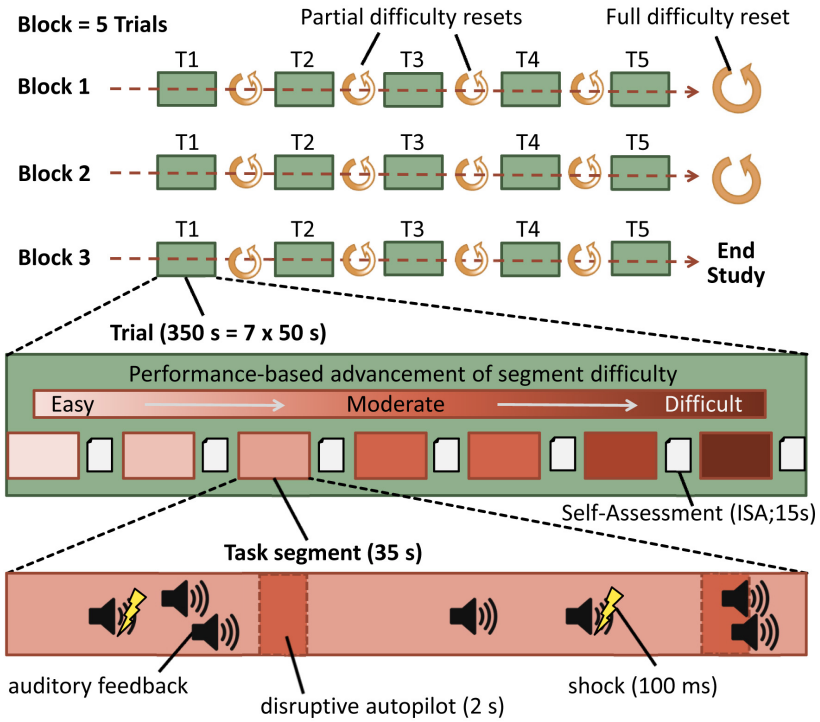


Figure 5. Recording Session Design

Participants simultaneously performed the three MATB tasks described above within each task segment. The fourth MATB task, Resource Management, was disabled and hidden. A modified version of the Tracking task used a Microsoft Sidewinder Force Feedback 2 joystick to deliver mildly assistive haptic feedback—automatically pushing the joystick to steer toward the center of the targeting window when the target drifted beyond a pre-determined threshold. During the testing blocks (i.e., non-training/practice trials), force feedback was also used to simulate disruptive automation failures in which the joystick was forced in the general direction of the random Tracking drift while shaking violently. One Communication event, two System Monitoring gauge fault events, one System Monitoring light fault event, and two 2-second disruptive autopilot events were scripted to occur in each segment. MATB event scripts were generated randomly and differed for all trials and segments but were identical across subjects. These shock windows were independent from trial segments and ISA pause periods and are not represented in Figure 5.

During testing blocks, participants were provided with real-time feedback on their task performance through auditory feedback *and noxious electrical stimuli (i.e., shocks)*. Aversive electrical stimuli were delivered 200 ms after harsh buzzer (error) sounds. In order to limit the number and frequency of shocks, each 350-second trial was subdivided into 17.5 20-second periods and shocks were limited to occur at most once in each such period.

Data Analysis

Feature Extraction

Two broad categories of potential fnsEMG indicators of cognitive workload were considered: *baseline muscle tension* and *coordinated motor activity*. Baseline muscle tension was investigated based on the observation that, even at rest, individuals typically maintain a minimum level of tension in many muscles to maintain posture. For example, minimum tension is required on the jaw muscle and lip elevators to prevent the mouth and lips from falling open. Previous research and the authors' qualitative observations in exploratory research with the AF-MATB noted that some participants appeared to visibly clench their jaws or purse their lips under increased task pressure, suggesting that slight variations in this baseline muscle tension are associated with cognitive workload (and perhaps especially

with cognitive overload induced by automation surprise/frustration). This hypothesis was tested by operationalizing the notion of baseline muscle tension as *short-run minimum EMG amplitude* (i.e., the minimum value of the EMG envelope over two-second intervals). Spearman's rank correlations were computed between this feature (measured on each fnsEMG channel) and taskload, task performance, and ISA. Statistically significant correlations for all subjects suggested that short-run minimum EMG amplitude *is* sensitive to variations in cognitive workload but with substantial variations across subjects in terms of which muscles manifested the hypothesized effect. Put plainly, some individuals under increased task pressure tended to clench their jaws, while others furrowed their brows, pursed their lips, or increased the muscle tension measured under their chins (possibly pressing their tongues against their teeth).

Facial expressions, such as those associated with emotional responses (including, potentially, emotional responses to cognitive workload and automation surprise/frustration), are characterized by idiosyncratic patterns of coordinated motor activity across multiple facial and neck muscles. Other activities, such as chewing and speech, exhibit their own idiosyncratic patterns of activity. Simple, static patterns associated with common emotional facial expressions are evident in the fnsEMG envelopes across multiple sensing locations (see Figure 3). Short-run summary statistics (e.g., minimum, maximum, mean, standard deviation, median) were extracted from the fnsEMG amplitude signals (i.e., the envelopes of the "raw" fnsEMG signals) to facilitate recognition of such coordinated activity.

Apart from fnsEMG, the following ten variables were derived from the physiological signals: respiration rate, the Index of Cognitive Activity (measured in each eye), heart rate (as measured by the mean inter-beat interval), heart rate variability (as measured by inter-beat interval standard deviation), EDA mean, EDA standard deviation, and the mean, standard deviation, and 85th percentile of the first derivative of EDA. The Index of Cognitive Activity is a pupillometric indicator based on using wavelet analysis to tease apart light reflex (regular oscillations in pupil size caused by minute variations in the amount of light entering the pupil) and dilation reflex (large, rapid dilations associated with cognitive activity, usually followed by more gradual contractions) (Marshall, 2002). These ten physiological variables and the short-run fnsEMG statistics described above were then used to train a data-driven cognitive workload model, as described in the following subsection.

Predictive Cognitive Workload Modeling

Machine learning was used to fit a non-parametric model to data collected from each subject in Trials 1 – 10, and then each of the resulting models was evaluated on the remaining held-out portion of each subject's dataset (Trials 11 – 15). The machine learning problem was cast as one of supervised learning. Labeled exemplars consisted of feature vectors extracted from the physiological signals, labeled as either "in-range" or "out-of-range," depending on the status of the participant's Tracking task targeting reticle at the corresponding time point. These class labels were chosen because Tracking task performance provided a suitable high-resolution proxy for cognitive workload in our experimental setup. Exemplars covered task segments at an 8 Hz temporal resolution. The features were the physiological variables described in the previous section. Models were trained separately for each participant to account for individual variation, particularly idiosyncratic emotional responses manifested in fnsEMG. Models were also trained without the fnsEMG variables to assess the benefits of adding fnsEMG.

The eXtreme Gradient Boosting (XGBoost) algorithm (Chen et al., 2015), a derivative of gradient-boosted decision trees (Friedman, 2002), was chosen as the machine learning algorithm for this experiment. This algorithm builds an ensemble of shallow decision trees over a series of iterations. The ensemble's trees are an intuitive choice for representing coordinated motor activity, corresponding to the simple decision rules that humans might postulate based on a handful of amplitude thresholds (see the example in Figure 6). They also have the advantage of being easy to interpret and naturally estimating the posterior class probability distribution.

XGBoost adds a new decision tree to the ensemble at each iteration, fitting it to the (pseudo-) residuals of the ensemble constructed throughout previous iterations. The pseudo-residuals are based on the gradients of a "loss function", chosen so that each iteration (greedily) moves the ensemble closer to a local optimum in the space of models. In this case, the model's "in-range" and "out-of-range" class probabilities were interpreted as estimates of conditional task error probabilities; therefore, log loss (also known

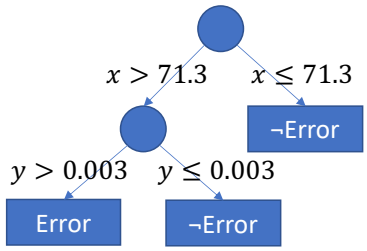


Figure 6. Example Decision Tree (simplified)

as binary cross-entropy; see Equation 1) was chosen as the loss function.

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1 - y_i) \log(1 - p_i) \quad (1)$$

Since task difficulty only increased when participants met a 92% performance threshold, exemplars of the "out-of-range" class were relatively scarce. This class imbalance was addressed using the Synthetic Minority Oversampling Technique (SMOTE) to avoid model bias. That is, exemplars of the minority "out-of-range" class were synthesized by interpolating between randomly selected minority exemplars and their nearest (minority class) neighbors (Chawla et al., 2002). Augmenting (or *over-sampling*) the training dataset in this way results in a balanced dataset and counteracts the tendency for the fitted model to exhibit a preference for the majority class that would otherwise occur with imbalanced training data. Over-sampling was *only* performed during model fitting—the model was evaluated against all observations in the test dataset (i.e., Trials 11 – 15), with their natural, unbalanced distribution of class labels.

Two primary techniques were used to avoid excessive model variance (or "over-fitting"), in which a fitted model effectively "memorizes" the training data and fails to generalize well to new data. First, the XGBoost "shrinkage" parameter (also called the "learning rate") was set to a low value; this parameter scales the weights in the decision tree added in each boosting iteration, thereby making learning more conservative with lower values of the shrinkage parameter. Second, sub-sampling (also known as "bagging" or *stochastic* gradient boosting) was employed—each training iteration was given only a small random sample of the (resampled, balanced) training set. The parameters associated with these techniques (i.e., shrinkage and sub-sample ratio) and the number of training iterations were tuned with grouped *k*-fold cross-validation. The first 10 Trials were first split into a "training dataset" (Trials 1 – 5 and 7 – 9) and a "calibration dataset" (Trials 6 & 10). The training dataset was sub-divided into two roughly equal "folds" so that samples from a given trial never appeared in both folds. Then an XGBoost model was fitted on each of the folds and evaluated on the other, left-out fold. The parameters that led to the best average performance on the left-out folds were then used to train a model on the subject's full training dataset. Finally, the class probability estimates for each subject were calibrated based on the calibration dataset with Kull et al.'s beta calibration method (2017), resulting in a calibrated predictive model for each subject.

After model training, each subject's fitted, calibrated predictive model was evaluated on the five held-out trials of the subject's data (Trials 11 – 15). Since the purpose of the predictive model is to produce an estimate of cognitive workload and not actually to predict task errors, the model was not evaluated based on *accuracy* or related metrics such as *balanced accuracy*, *sensitivity*, or *specificity* on the "in-range" / "out-of-range" classification task. Instead, the model was evaluated based on two criteria: 1) the correlations between the model's segment-level mean estimated class probabilities (i.e., the average of the model's probabilistic outputs for all feature vectors in the 8 Hz time series for each segment) and task difficulty (the independent variable), self-reported task difficulty (ISA), and overall task performance; and 2) the quality of the model calibration. The model was also compared to one without fnsEMG features.

Model calibration measures how well a probabilistic model's estimates accord with the observed frequency of events at each estimated probability level—in this case, the model's "out-of-range" probability estimates accorded with the observed frequency of "out-of-range" events at each estimated probability level. When estimated event probabilities are plotted against the observed frequencies of events, as in the calibration plots below (Figure 7), a perfectly calibrated estimate would appear as a line along the y=x diagonal. A well-calibrated probabilistic estimate of performance errors in the Tracking task can be interpreted as an estimate of cognitive workload since increases in cognitive workload ultimately drive the physiological indications that errors are more likely to occur. These physiological indications are present under heightened cognitive workload, even when errors do not occur.

RESULTS

Segment-level mean model estimates were weakly rank-correlated (Spearman's ρ) with self-reported task difficulty and were weakly negatively rank-correlated with performance; these correlations were statistically significant (α=0.05) (see Table 1). Model estimates also exhibited a weak positive correlation (Pearson's r) with the Tracking

task difficulty parameter, a moderate positive correlation with self-reported task difficulty, and a weak negative correlation with performance; all correlations were statistically significant (α=0.05) (see Table 2). Notably, the correlations were stronger between mean model estimates and self-reported task difficulty than with task performance scores. This relationship is consistent with the model functioning as an estimator of cognitive workload rather than a predictor of task performance despite being fitted to data labeled based on performance. However, it is not surprising and is indeed by design, as the model's physiological input features have a much more direct causal relationship with cognitive state than with task performance. The segment-level mean model estimates of identically trained models without access to the fnsEMG features exhibited weaker correlations (e.g., ρ=0.114 and r=0.124 for self-reported task difficulty), suggesting that fnsEMG features improved the model estimates.

Table 1. Spearman's Rank Correlations (ρ) between Predictive Model Estimates and Other Variables

Variable	ρ	p value (one-tailed)
Tracking task difficulty	0.0710	0.095
self-reported task difficulty	0.269	**2.0×10^{-7}**
performance score	-0.210	**4.5×10^{-5}**

Table 2. Pearson's Correlations (r) Between Predictive Model Estimates and Other Variables

Variable	r	p value (one-tailed)
Tracking task difficulty	0.105	**0.026**
self-reported task difficulty	0.303	**5.2×10^{-9}**
performance score	-0.116	**1.4×10^{-3}**

The calibration plots in Figure 7 show non-parametric estimates of model calibration on the calibration data (Trials 6 & 10; left) and the testing data (Trials 11 – 15; right). Since the models' estimates were calibrated to the calibration data, the quality of model calibration on the calibration data is nearly perfect, as expected. The test set calibration plot shows that the fitted models were also well-calibrated on the held-out test data—the observed frequencies of Tracking "out-of-range" events for each range of model estimates were entirely consistent with the probabilistic estimates themselves, resulting in a plot near the ideal y=x diagonal. The models were somewhat over-confident in probability estimates below approximately 8% (i.e., "out-of-range" errors occurred slightly more frequently than predicted by the model estimates in that range), represented visually by the points plotted above the ideal line, and were somewhat under-confident in estimates over 8% before diverging more significantly for the relatively small number of model estimates over 15%, represented visually by the deviation away from the ideal line for those estimates. Histogram bars superimposed on the calibration plot show the binned count of predicted probabilities.

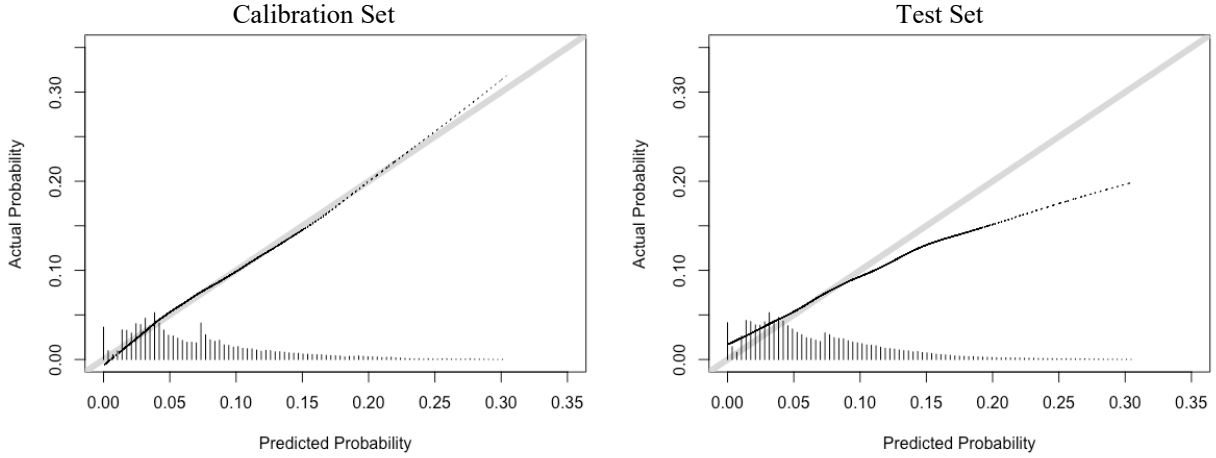


Figure 7. Predictive Model Calibration Plots

DISCUSSION

Taken as a whole, the results of this study support the technical merits of an approach to cognitive workload estimation based on a complement of physiological measures that can, at least in principle, be collected from unobtrusive wearable sensors in a variety of practical applications. The study confirmed the sensitivity of fnsEMG to taskload variations in a complex set of laboratory tasks modeled after real-world piloting tasks. It also established the technical feasibility of predicting cognitive workload from a combination of the physiological measures studied as taskload varied within the same set of laboratory tasks. Since all measures studied can be computed from physiological signals in real time, this approach to cognitive workload estimation has the potential to be used in real-time applications, such as adaptive automation (e.g., intelligent tutoring systems that adapt their training based on trainees' estimated cognitive workloads).

The present study has several limitations that must be acknowledged and addressed before fnsEMG can support cognitive workload estimation in real-world settings; these limitations point the way to potentially fruitful directions for future inquiry. First, the study focused on a particular set of artificial (though realistic) tasks. To ensure its generality, the cognitive workload estimation approach must be validated in a broader range of tasks, including real-world tasks and work environments. Likewise, although the shock consequence for task errors likely raised participants' engagement and concern for performance, as we have seen in our prior experiments with the enhanced MATB, the impact of performance feedback and error punishment must be systematically studied to appreciate its impact on physiological responses to cognitive workload. Second, the laboratory study employed a set of relatively encumbering, high-fidelity physiological sensors that required approximately 45 minutes for their application. Further data analysis could indicate options for reducing sensor count without loss of predictive model strength, and new flexible sensor technologies (for sEMG, EDA, ECG, etc.) could hasten the application process and reduce sensor encumbrance without sacrificing fidelity (Gao, Parida, and Lee, 2020). Even less obtrusive sensors may be necessary for many practical applications, and future research is needed to explore whether robust estimates of cognitive workload can be derived from commercially available wearable sensors (e.g., wearable pulsometers for sensing HR/HRV) despite lower signal information content and/or fidelity. Alternative measures that capture similar information should also be considered; for instance, in many settings, it may be more practical to infer emotional responses from imagery than it is to measure these somatic responses directly with fnsEMG—facial (micro-) expressions, pupillometry, skin temperature changes, and even heart/respiration rates can be captured from visible and infrared imagery (including high-speed motion imagery). Future research should also investigate whether this limitation can be offset by exploiting additional objective predictors, including measures of behaviors and/or work context that may predict cognitive state regardless of the nature or direction of any specific causal relationship.

ACKNOWLEDGEMENTS

The authors wish to thank the Air Force Research Laboratory (AFRL) and, in particular, Dr. Justin Estepp for their role in developing the AF-MATB software implementation and for providing its source code to DARPA for use in this research. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. 140D6318C0039[1]. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

REFERENCES

Angelborg-Thanderz, M. (1990). Military flight training at a reasonable price and risk. Economics Research Institute, Stockholm School of Economics and FOA Report C, 50083–50085.

Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. *Psychophysiology*, 49(1), 1017-1034.

Cain, B. (2007). A review of the mental workload literature.

[1] Approved for Public Release, Distribution Unlimited

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., & Chen, K. (2015). Xgboost: Extreme gradient boosting. R package version 0.4-2, 1(4), 1-4.

Comstock, J., & Arnegard, R. (1992). The multi-attribute task battery for human operator workload and strategic behavior research.

Dupre, A., Vincent, S., & Iaizzo, P. A. (2005). Basic ECG Theory, Recordings, and Interpretation. In *Handbook of Cardiac Anatomy, Physiology, and Devices*, 191–201. Springer.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), 367-378.

Gao, D., Parida, K., & Lee, P. S. (2020). Emerging soft conductors for bioelectronic interfaces. *Advanced Functional Materials*, 30(29), 1907184.

Hart, S. (2006). NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage Publications.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in Psychology* (Vol. 52, pp. 139-183). North-Holland.

Hicks, C. L., von Baeyer, C. L., Spafford, P. A., van Korlaar, I., & Goodenough, B. (2001). The Faces Pain Scale–Revised: Toward a common metric in pediatric pain measurement. *Pain*, 93(2), 173-183.

Jordan, C. S., & Brennen, S. D. (1992). Instantaneous self-assessment of workload technique (ISA). Defense Research Agency, Portsmouth.

Kaber, D. B., & Riley, J. M. (1999). Adaptive automation of a dynamic control task based on secondary task workload measurement. *International Journal of Cognitive Ergonomics*, 3(3), 169-187.

Kull, M., Silva Filho, T., & Flach, P. (2017, April). Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics* (pp. 623-631). PMLR.

Marshall, S. P. (2002, September). The index of cognitive activity: Measuring cognitive workload. In *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants* (pp. 7-7). IEEE.

Matthews, G., Reinerman-Jones, L., Wohleber, R., Lin, J., Mercado, J., & Abich, J. (2015, August). Workload is multidimensional, not unitary: What now?. In *International Conference on Augmented Cognition* (pp. 44-55). Springer, Cham.

Novstrup, A., Goan, T., & Heaton, J. (2019). Workload assessment using speech-related neck surface electromyography. In *Human Mental Workload: Models and Applications: Second International Symposium, H-WORKLOAD 2018, Amsterdam, The Netherlands, September 20-21, 2018, Revised Selected Papers 2* (pp. 72-91). Springer International Publishing.

Olszanowski, M., Pochwatko, G., Kuklinski, K., Scibor-Rylski, M., Lewinski, P., & Ohme, R. K. (2015). Warsaw set of emotional facial expression pictures: A validation study of facial display photographs. *Frontiers in Psychology*, 5, 1516.

Quatieri, T., Williamson, J., Smalt, C., Perricone, J., Patel, T., Brattain, L., … Eddy, M. (2017). Multimodal Biomarkers to Discriminate Cognitive State. *The Role of Technology in Clinical Neurophysiology*, 409.

Rosenberg, E. L., & Ekman, P. (Eds.) (2020). What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press.

Santiago-Espada, Y., Myer, R., Latorella, K., & Comstock, J. (2011). The Multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload Research: A User's Guide. Chicago.

Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39(5), 740-748. doi: 10.1080/00140139608964495. PMID: 8635447.

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, 35(17), 2503-2522.