Applied Ergonomics 77 (2019) 83–91

Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

The effects of dynamic workload and experience on commercially available EEG cognitive state metrics in a high-fidelity air traffic control environment

Kyle A. Bernhardt[a,*], Dmitri Poltavski[a], Thomas Petros[a], F. Richard Ferraro[a], Terra Jorgenson[b], Craig Carlson[b], Paul Drechsel[b], Colt Iseminger[b]

[a] Department of Psychology University of North Dakota, 501 North Columbia Rd, Stop 8380, Grand Forks, ND, 58202, USA
[b] Department of Aviation University of North Dakota, 4251 University Ave, Grand Forks, ND, 58202, USA

ARTICLE INFO

Keywords:
Air traffic control
Cognitive state monitoring
Electroencephalography
Operator experience

ABSTRACT

The current study evaluated the validity of commercially available electroencephalography (EEG) cognitive state metrics of workload and engagement in differentially experienced air traffic control (ATC) students. EEG and pupil diameter recordings were collected from 47 ATC students (27 more experienced and 20 less experienced) during a high-fidelity, variable workload approach-control scenario. Scenario workload was manipulated by increasing the number of aircraft released and the presence of a divided attention task. Results showed that scenario performance significantly degraded with increased aircraft and the presence of the divided attention task. No scenario performance differences were found between experience groups. The EEG engagement metric significantly differed between experience groups, with less experienced controllers exhibiting higher engagement than more experienced controllers. The EEG workload metric and pupil diameter were sensitive to workload manipulations but did not differentiate experience groups. Commercially available EEG cognitive state metrics may be a viable tool for enhancing ATC training.

1. Introduction

Cognitive workload and engagement are two constructs commonly implicated in human performance research. Cognitive workload refers to the dynamic relationship between the resources that are needed to carry out a task and the ability of the operator to adequately supply those resources (Wickens and Tsang, 2015). Engagement is a state indicating the availability of attentional resources and the mobilization of resources for efficient processing of task-related stimuli typically associated with concentration and sustained attention (Kamzanova et al., 2011). Workload and engagement differ in that workload describes a supply-demand relationship of cognitive resources, while engagement describes more so alertness and the ability to attend to and process relevant task stimuli by being actively orientated toward the task (Freeman et al., 1999; Hockey et al., 2009). In terms of measurement, cognitive workload has been quantified by several hallmark self-report measures, such as the NASA Task Load Index (Hart and Staveland, 1988) and the Subjective Workload Assessment Technique (Reid and Nygren, 1988). Both of these measures contain aspects of time pressure, effort, perceptions of performance, and frustration. Furthermore, engagement has been commonly measured by the Dundee Stress State Questionnaire (DSSQ) developed by Matthews et al. (2002). In contrast to workload scales, the DSSQ task engagement scale contains aspects of energy, motivation, and concentration.

Over the past two decades, a significant amount of research in academia and the applied sector has been devoted to developing neural measures to index these cognitive states. The use of neural measures to infer cognitive states is well-suited for system evaluation and driving adaptive brain-computer interfaces. Neural measures have the key advantage of collecting data in real time without interrupting the operational environment (Parasuraman, 2015). Moreover, these measures assess changes in cognitive states not evident with overt task performance (Parasuraman, 2015; Wickens and Tsang, 2015).

Electroencephalography (EEG) has been widely used to assess cognitive workload (e.g., Borghini et al., 2014; Gevins and Smith, 2003; Matthews et al., 2015) and engagement (e.g., Hopstaken et al., 2015; Wascher et al., 2014; Wilson et al., 2007). Changes in EEG power spectral density (PSD) with different workload and engagement states have also been well documented (e.g., Borghini et al., 2014; Wascher et al., 2014).

* Corresponding author. Department of Psychology, University of North Dakota, 501 North Columbia Rd, Stop 8380, Grand Forks, ND, 58202, USA.
E-mail addresses: kyle.bernhardt@und.edu (K.A. Bernhardt), dmitri.poltavski@und.edu (D. Poltavski), thomas.petros@und.edu (T. Petros), f.richard.ferraro@und.edu (F.R. Ferraro), terra.jorgenson@und.edu (T. Jorgenson), craig.carlson@und.edu (C. Carlson), paul.drechsel@und.edu (P. Drechsel), colt.iseminger@und.edu (C. Iseminger).

https://doi.org/10.1016/j.apergo.2019.01.008
Received 7 April 2018; Received in revised form 7 November 2018; Accepted 18 January 2019
Available online 08 February 2019

Researchers have refined EEG PSD patterns into summary indices to improve cognitive state estimation. For example, Holm et al. (2009) indexed cognitive workload as the ratio of frontal midline theta to midline parietal alpha power. Results from this study indicated that this metric reliably increased with higher task demands. Berka et al. (2007) developed a method for classifying both workload and engagement using discriminant function analysis (DFA) and multiple EEG variables from several different electrode sites. This method utilizes absolute and relative PSD values from multiple EEG bandwidths across different electrode locations. One model generates probabilities of sleep onset, distraction, low engagement, and high engagement states and another model generates probabilities of high and low workload. These algorithms are commercially available through Advanced Brain Monitoring's (ABM) wireless X-series EEG systems (Advanced Brain Monitoring, 2009).

While ABM's metrics were developed under laboratory conditions (Berka et al., 2007; Johnson et al., 2011) with basic cognitive tasks to maximize internal validity, it is still unclear if these metrics are effectively able to be deployed in a variety of ecologically valid, dynamic operational settings. Furthermore, it is also unclear whether these metrics are affected by variables such as differences in experience and follow predictions made by models of cognitive workload and engagement under operational conditions.

One such prediction is that individuals with more expertise or skill have more available resources to supply to a task and thus experience less workload (Ayaz et al., 2012). However, this workload difference may not be apparent in overt task performance but may manifest in neural measures (Ayaz et al., 2012). For example, two operators of differing experience levels could be demonstrating equal task performance; however, the less experienced operator may not have the residual resource capacity to cope with added tasks (e.g., an emergency) compared to the more experienced operator. Thus, the less experienced operator may benefit from more training that would not otherwise be detected with task performance alone. In other words, neural measures provide a means to examine operator cognitive efficiency underlying primary task performance.

Functional neuroimaging studies generally report less neural activation with increased expertise or task practice (Kelly and Garavan, 2005). For instance, Ayaz et al. (2012) reported decreased frontal lobe hemodynamic responses over the course of more practice sessions with an unmanned aerial vehicle task. Similar deactivation patterns have been found while participants repeatedly perform simple laboratory-based tasks (e.g., Beauchamp et al., 2003; McKendrick et al., 2014). EEG variables have also been shown to differentiate between novice and experts. For example, expert pistol shooters have been shown to exhibit increased frontal midline theta power and decreased overall alpha power compared to novices during shooting tasks (Doppelmayr et al., 2008; Johnson et al., 2014). Importantly, Johnson et al. (2014) utilized ABM's cognitive state metrics of engagement and workload while participants performed a deadly force judgment and decision-making shooting task. Their results indicated that expert shooters demonstrated less engagement than novices. No differences between novice and experts were found for the workload metric. However, those identified as having an intermediate expertise level exhibited significantly less workload than novices and experts.

To our knowledge, no studies have explored the use of commercially available EEG cognitive state metrics with concurrent manipulations of task workload and operator expertise. Furthermore, there is a lack of literature using these derived metrics with secondary physiological measures of workload. A better understanding of the utility of such metrics in applied contexts may have important implications for operator training, task assignment, and augmented cognition system development. Moreover, providing empirical validation of these metrics in different operational settings can further improve upon neuroergonomic approaches for practitioners looking to use these commercially available cognitive state metrics in system evaluation protocols.

The purpose of the present study was to test two commercially available EEG cognitive state metrics in collegiate air traffic control (ATC) students with varying experience levels during a dynamic, high-fidelity Terminal Radar Approach Control (TRACON) scenario. The neuroergonomic approach has a strong foothold in the aviation domain. For example, a recent study utilized the P300 event-related potential to compare two different types of notification system designs for use in the ATC environment (Giraudet et al., 2015). Using a concurrent auditory oddball paradigm, results from this study indicated that P300 amplitude was greater when notifications were delivered via an animated method rather than a stationary blinking alert. Aricò et al. (2016) found that frontal theta and parietal alpha bands could be used in a multivariate algorithm to drive an adaptive automation system to improve ATC performance. Furthermore, Dasari et al. (2017) found that theta power increased in frontal regions and alpha power tended to decrease in the central medial, parietal, motor, and occipital regions during more demanding parts of a simulated ATC task. At the aggregate level, a recent review of neuroergonomic methods in the ATC environment also reported similar patterns of alpha and theta changes during more demanding ATC tasks (Aricò et al., 2017).

Additionally, to serve as an established measure of workload, EEG workload was compared to pupil diameter. The task-evoked pupillary response has also been used as a measure of workload in a variety of perceptual and cognitive tasks (Beatty, 1982). Studies have documented pupil diameter increases with higher difficulty working memory tasks, such as the *n*-back (Mandrick et al., 2016) and secondary auditory driving tasks (Marquart et al. 2015). A recent study evaluating an immersive virtual ATC environment (Truschzinski et al., 2018) found that pupil diameter significantly increased with scenario difficulty. The pupil diameter of ATCs has also been shown to increase following alerts warning of potential air traffic conflicts, indicating an increase in cognitive workload (Kearney et al., 2018). Finally, it has been reported that pupil diameter could discriminate between difficulty levels of a mental multiplication task to a similar degree as the NASA Task Load Index (Marquart and de Winter, 2015).

It was hypothesized that during high workload periods, less experienced students would demonstrate higher EEG workload than more experienced students because of restricted cognitive resources available to perform the scenario (Tsang and Vidulich, 2006). It was also predicted that EEG workload would correspond to the projected scenario workload and correlate with concurrent pupil diameter measurements. Finally, we hypothesized that more experienced controllers would exhibit less EEG engagement because of increased neural efficiency resulting from more practice (Kelly and Garavan, 2005).

This study is novel in that we examined a set of commercially available cognitive state metrics with concurrent manipulations of both task demands and operator experience in a high-fidelity TRACON environment. Previous studies (e.g., Johnson et al., 2014) have only used one of these manipulations. As researchers turn toward the neuroergonomic approach for a more fine-grained analysis of human-system interactions, commercially available algorithms are more likely to be used rather than researchers developing these metrics from scratch. Therefore, this study addresses how well these cognitive state metrics generalize to novel settings across two independent variables that are commonly encountered simultaneously during operational conditions (i.e., fluctuations in workload and different operator experience levels).

2. Method

2.1. Participants

Forty-nine undergraduate ATC students (two women and 47 men, $M_{\text{age}} = 20.97$) participated in the study for course credit or $20. All participants had ongoing experience with TRACON simulator procedures. Participants did not have prior knowledge of the experimental scenario. Two participants were excluded due to simulator recording malfunction, resulting in a final sample size of 47 (two women and 45 men). Participants were split into low (n = 20) and high (n = 27) experience levels based on the current ATC course in which they were enrolled. More experienced students were enrolled in an upper-level capstone course that offered advanced training in topics such as ATC publications, regulations, airspace utility, special operations, and emergencies. Less experienced students were enrolled in a lower-level course in which they received introductory training on the topics presented in the advanced course. Students in the high experienced group must have passed the lower-level course first as well as an additional tower and radar operations course before enrolling in the upper-level course. Prior to enrolling in the classes, students in the low experience group would have acquired an average of 56 h of simulation time, while those in the high experience group would have acquired approximately 97 h of simulation time. It is important to note that both courses provided students with the relevant knowledge to perform all aspects of the experimental scenario.

2.2. Individual difference measures

Measures of mental rotation ability, sleep quality, ATC knowledge, and general intelligence were administered prior to simulator performance. These measures were included to ensure the two experience level groups were equated on variables potentially related to ATC performance.

2.2.1. Mental rotation

Mental rotation ability was measured with the Vandenberg mental rotations test (Vandenberg and Kuse, 1978) administered via paper and pencil method. Mental rotation has been shown to positively correlate with multitasking performance during changing task constraints (Morgan et al., 2013) and is suggested to play an important role in ATC performance (Shorrock and Isaac, 2010). Higher scores indicate better mental rotation ability.

2.2.2. Sleep quality

Sleep quality was assessed with the Pittsburgh Sleep Quality Index (PSQI; Buysse et al., 1989). The PSQI is a self-report measure that evaluates sleep quality over the past 30 days. The PSQI includes seven subscales, which are summed to form a global score ranging from 0 to 21. Higher scores indicate worse sleep quality. Prior research has shown that sleep quality can influence cognitive task performance (Nebes et al., 2009).

2.2.3. ATC knowledge

Participant ATC knowledge was assessed with a 15-item, free-response style test consisting of questions regarding basic TRACON procedures, phraseology, and separation. The knowledge test was developed by university ATC faculty members so both experience levels could adequately answer the questions. Tests were de-identified and blindly scored by ATC faculty members. One point was awarded for each correct answer.

2.2.4. General intelligence

The vocabulary subset of the Wechsler Adult Intelligence Scale Third Edition (WAIS-III; Wechsler, 1997) was used to estimate general intelligence. The vocabulary subset of the WAIS has good test-retest reliability (r = 0.89) and correlates well with overall measures of intelligence (Wechsler, 1997). Studies have found positive relationships between multitasking and general intelligence (Morgan et al., 2013; Salomon et al., 2015).

2.3. Simulator and scenario

A high-fidelity, computerized TRACON simulator running ATCoach Global (V.4.32.5) Air Traffic Software was used in this study. Participants were seated at a station with a keyboard, aircraft strips, touch-screen communications box, trackball mouse, and a computer monitor displaying a RADAR scope (20.1 in diagonal viewable area). Controllers gave verbal commands via a radio headset to two “pseudo-pilots” located at stations similar to the controller's station (out of the controller's sight). The pseudo-pilots inputted the controller's commands for the aircraft and gave verbal readbacks of all controller instructions. Controllers were also required to complete flight strip marking procedures. The experimental setup is displayed in Fig. 1.

Fig. 1. A photo of the TRACON experimental setup.

The airspace used in this study was a fictitious training airspace named Academy. Academy is a model of the airspace utilized by the Federal Aviation Administration (FAA) ATC training academy at the Mike Monroney Aeronautical Center in Oklahoma City, Oklahoma. The Academy airspace consists of centrally located Academy Airport (KAAC) and several satellite airports (one towered military airport and four uncontrolled airports). All simulated aircraft were operating under instrument flight rules (IFR). Arrivals and departures occurred at KAAC and satellite airports.

The TRACON scenario was 1.25 h long and divided into five, 15 min phases. This scenario was developed in-house during prior work to create a variable workload TRACON scenario for research purposes. Two workload-contributing factors distinguished the phases: (1) number of aircraft arriving and departing the airspace within a given 15 min interval and (2) the presence or absence of issuing uncontrolled aircraft IFR clearances. These uncontrolled clearances served as a divided attention task. Air traffic volume fluctuations are a common ATC scenario workload manipulation (Vogt et al., 2006). That is, the more aircraft controllers are required to handle, the more cognitive resources are required to meet demands. Moreover, providing uncontrolled clearances strains working memory such that controllers were required to maintain and integrate aircraft information into a standard clearance phraseology statement for departing aircraft. Manipulations of working memory load show predictable patterns of workload changes (Wickens and Tsang, 2015), such that increasing working memory load corresponds to increases in workload. Phases were designated A, B, C, D, E and consisted of 8, 11, 16, 8, and 3 departing/arriving aircraft, respectively. Uncontrolled aircraft clearances occurred during Phase C. The predicted workload of the scenario was designed to follow a negative quadratic shape, with peak workload occurring at Phase C (see Fig. 2).

2.3.1. Simulation scoring

Each scenario run was digitally recorded by the simulator and later scored by a team of university ATC faculty members. Scenario raters were blind to participant experience level. Participant performance was evaluated by tracking the number of errors made in phraseology, violation of aircraft separation minimums, and airspace procedures for each 15 min phase according to the standards set by the FAA and established procedures for the scenario.

2.4. Physiological recordings

2.4.1. EEG recording

EEG was recorded using the Advance Brain Monitoring (ABM) B-Alert X-24 wireless Bluetooth system. This system was selected because it provides easily interpreted cognitive state metrics that are derived from multiple scalp locations and EEG frequencies. The B-Alert cognitive state metrics are less computationally intensive than other algorithms, allowing for a smaller overall system (Johnson et al., 2011). Relatedly, the B-Alert system is wireless, making it suitable for evaluating cognitive states in dynamic, ecologically valid working situations. This system was also selected because it is a widely used and cost-effective EEG platform that provides a user-friendly interface for collecting EEG recordings in operational settings. In a usability test, the B-Alert platform outperformed two other wireless EEG systems in terms of participant comfort and experimenter preference for application (Hairston et al., 2014).

The X-24 incorporates 20 electrodes placed according to the international 10/20 system: Fp1, Fp2, F7, F3, Fz, F4, F8, T3, C3, Cz, C4, T4, T5, P3, Pz, P4, T6, O1, POz, and O2. Reference electrodes were placed at the mastoids. EEG data were sampled at 256 Hz and filtered with 50, 60, 100, and 120 Hz notch filters with Low Pass FIR filters online during data collection. PSD was automatically computed by the acquisition software (B-Alert Live; Advanced Brain Monitoring, 2009) by performing Fast Fourier Transform (FFT) on raw data and calculating the amplitudes of the sinusoidal components for designated frequency bins on a second-by-second basis. FFT analyses were performed after application of a 50% overlay between two consecutive epochs both with and without application of a Kaiser window for data smoothing. Selected 1-Hz bins were then automatically averaged and logged to create conventional EEG bands across five bandwidths: delta (1–2 Hz), theta (3–7 Hz), alpha (8–13 Hz), beta (13–29 Hz), and gamma (25–40 Hz).

Prior to computing the 1-Hz PSD bins, the software automatically processed the raw signals to eliminate known artifacts including electromyographic signals from muscle activity (EMG), electrooculographic signals from eye blinks (EOG), spikes, excursions, and amplifier saturations (See Advanced Brain Monitoring, 2009; Berka et al., 2004 for more details).

The software provided metrics of cognitive states including probabilities of high and low engagement, cognitive workload, distraction, and sleep onset. These metrics are based on algorithms validated by Berka et al. (2007) and Johnson et al. (2011). The metrics are derived using absolute and relative PSD values across multiple differential channel locations and frequency bands in the ranges of 1–4, 5–7, 8–13, 14–24, and 25–40 Hz. The high and low engagement, distraction, and sleep onset metrics were derived for each 1-s epoch using 1 Hz PSDs from differential sites FzPOz and CzPOz (for the number of EEG features for each frequency range and each location, see Berka et al., 2007) using a four-class quadratic discriminant function analysis (DFA). The model coefficients were initially derived from a database of over 100 individuals under both sleep deprived and fully rested conditions (Berka et al., 2007). Individualized discriminant function coefficients were computed during a coefficient adjustment procedure in which participants performed three neurocognitive benchmark tasks: 3-choice vigilance task, visual psychomotor vigilance task, and auditory psychomotor vigilance task, with the sleep onset classification predicted from the baseline PSD values, for a total of 15–17 min of data across tasks. From this individualized model, posterior probabilities of participants being classified in one of the four cognitive states ranging 0.00 to 1.00 were generated for each 1 s epoch online during the simulation by the software. For this study, the probability of high engagement was used as the outcome measure for engagement, with probabilities closer to 1.00 indicating more engagement.

The acquisition software also provided generalized (not individually fit) metrics of cognitive workload for each 1-s epoch using 1 Hz PSDs from differential sites FzPOz, CzPOz, FzC3, C3C4, F3Cz, and F3Cz based on two models – one model was built on a forward digit span (FDS) task and the other on a backward digit span (BDS) task. According to Advanced Brain Monitoring (2009), the FDS model fits 85% of the population, while the BDS model provides a better fit for the other 15% of the population. Probabilities of high and low workload were generated using two-class linear DFAs. Probabilities closer to 1.00 reflect higher workload. The software also provides an average probability of the FDS and BDS models. This mean probability was used as the EEG workload outcome measure.

For further details on these cognitive state metrics, see Berka et al. (2007) and Johnson et al. (2011).

2.4.2. Pupil diameter recording

Binocular pupil diameter was recorded using the display-mounted Tobii X2-60 Eye Tracker (60 Hz sampling rate). The X2-60 was mounted directly below the RADAR display monitor allowing the participants approximately 50 × 36 cm of head movement 70 cm from the screen. Before each participant's run, the eye tracker was calibrated with a 9-point moving target and the participant's head positioning from the screen was checked. Once ideal positioning was achieved, participants were instructed to remain in this position as much as possible during the ATC simulation. Pupil diameter data were then exported using Tobii Studio (version 3.2). Epochs during which both eyes could not be reliability detected were rejected. This was accomplished using left/right eye validity codes inserted into the data stream by Tobii Studio. Room lighting conditions were kept constant across participants.

2.5. Procedure

Participants arrived at the TRACON simulation room, provided informed consent, and then completed a demographics questionnaire. Participants then completed the individual differences measures. Next, participants were fitted with the wireless EEG system and completed the neurocognitive benchmark tests. Then, participants were briefed on the scenario and flight strip marking procedures. The eye tracker was then calibrated and the scenario commenced. The scenario ran continuously for 1.25 h. Since participants had continuous experience with the TRACON simulator, no practice session was given.

2.6. Data processing and analytics

Overall, this study followed a 2 (Experience: low, high) x 5 (Phase: A, B, C, D, E) mixed factorial design with experience serving as the between-subjects factor. EEG data for each phase of the scenario were sectioned off using time-synchronized markers manually inputted by a research technician. Post data collection, EEG markers were transposed from the EEG record to the Tobii X2-60 data sequence to synchronize EEG and eye tracking data. For each 15 min scenario phase, 5% trimmed means were computed for time-locked EEG indices and pupil diameter recordings to eliminate extreme values and improve normality. For pupil diameter data, the results for the left eye are reported. The results for the right eye were identical. EEG, eye-tracking, and performance data were analyzed with linear mixed effects models to determine the effects of TRACON scenario phase (within-subjects variable) and controller experience (between-subjects variable) on EEG cognitive state metrics and simulator performance using the *nlme* package (Pinheiro et al., 2017) for R (R Core Team, 2017). Linear mixed effects models do not require independent observations or the assumption of sphericity as with repeated measures ANOVA analyses. Furthermore, linear mixed models account for individual variation in outcome measures with the inclusion of random intercepts and slopes across participants (West et al., 2015).

Models were specified iteratively using a “step-up” method (West et al., 2015) by first testing the effects of random participant intercepts

Table 1
Summary of individual difference measures by experience group.

Measure	Low Experience		High Experience		t	df	p
	M	SD	M	SD			
Mental Rotation	8.70	5.32	7.00	3.21	−1.36	45	.180
Global PSQI	6.16	3.48	5.78	2.41	-.44	44[a]	.663
ATC Knowledge Test	5.90	2.40	6.89	2.06	1.51	45	.137
WAIS	32.25	6.87	35.41	8.51	1.36	45	.180

Note. [a]One participant failed to complete all PSQI items and a global score could not be computed. This participant was removed from the PSQI analysis.

and slopes across phases and then adding effects of phase, experience, and phase by experience interaction. Parameters were estimated using the maximum likelihood method to allow for comparisons between models using likelihood ratio tests (i.e., change in -2log likelihood). Fixed effects for final models were then analyzed with F-ratios using Type-III sums of squares. When warranted, pairwise comparisons were conducted using the Bonferroni correction for multiple comparisons to compare outcome measures across scenario phases. Statistical significance was set at $\alpha = 0.05$.

3. Results

3.1. Individual differences

Means and standard deviations for the individual difference outcome measures are displayed in Table 1. Independent samples t-tests comparing the experience groups on these measures revealed no significant differences between experience groups.

3.2. EEG engagement and workload

One participant was excluded from EEG analyses because of poor data quality (< 70%). Mean data quality for sites used in cognitive state metric computation ranged from 89.59% (C3C4) to 99.82% (CzPOz). Both high engagement [$\chi^2(1) = 529.10$, $p < .001$] and workload [$\chi^2(1) = 419.07$, $p < .001$] models were significantly improved by including a random effect for each subject compared to an intercept only model. Moreover, high engagement [$\chi^2(18) = 102.61$, $p < .001$] and workload [$\chi^2(18) = 124.07$, $p < .001$] models were further improved by including the fixed and random effect of phase. For high engagement, the final model was also significantly improved by including the fixed effect of experience, $\chi^2(1) = 5.55$, $p = .019$. Tests of fixed effects for the final model for high engagement revealed a significant effect for experience [$F(1,44) = 6.58$, $p = .014$] and a non-significant effect for phase, $F(4, 180) = 1.42$, $p = .229$. In other words, EEG engagement differentiated the two experience groups but remained relatively constant throughout the scenario for both experience groups (see Fig. 2). Less experienced students ($M = 0.52$, $SD = 0.25$) exhibited higher engagement than more experienced students ($M = 0.43$, $SD = 0.17$).

For the average workload metric, the best fitting model incorporated the previously mentioned random and fixed effect of phase. Tests of fixed effects revealed that the effect of phase was significant, $F(4, 180) = 13.59$, $p < .001$, indicating that the workload metric fluctuated significantly over the different phases (see Fig. 2). Pairwise comparisons using Bonferroni correction revealed that Phases C ($M = 0.73$, $SD = 0.06$) and D ($M = 0.72$, $SD = 0.07$) were significantly higher than Phases A ($M = 0.71$, $SD = 0.07$) and E ($M = 0.71$, $SD = 0.07$), but were not significantly different from one another. Phase C was also significantly higher in workload than Phase B ($M = 0.72$, $SD = 0.07$). All other comparisons were non-significant. Fig. 2 displays these trends.

3.3. Pupil diameter

One participant was excluded from pupil diameter analyses due to technical failure. The initial pupil diameter model benefitted from a random subject effect, $\chi^2(1) = 577.10$, $p < .001$. The final model for pupil diameter was significantly improved with the inclusion the fixed effect of phase with a corresponding random effect, $\chi^2(1) = 121.24$, $p < .001$. Tests of fixed effects revealed a significant main effect of phase, $F(4, 180) = 22.77$, $p < .001$. Pupil diameter varied across phases in a similar pattern to the EEG workload metric (see Fig. 2). Pairwise comparisons showed that participants during Phase A ($M = 5.63$, $SD = 0.65$) had significantly smaller pupil diameters than Phases B ($M = 5.73$, $SD = 0.65$) and C ($M = 5.76$, $SD = 0.63$). Pupil diameter during Phase D ($M = 5.67$, $SD = 0.65$) was significantly smaller than during Phase C. Finally, participants had smaller pupil diameters during Phase E ($M = 5.60$, $SD = 0.65$) than Phases B, C, and D. All other comparisons were non-significant.

3.4. Workload measure correlation

EEG workload and pupil diameter measures were averaged across the five sessions and correlated with one another. A non-significant correlation resulted between the two workload metrics, $r(43) = -0.25$, $p = .098$. Furthermore, intercorrelations between EEG workload and pupil diameter at each phase were non-significant ($ps > .05$).

3.5. Performance

The final models for separation and procedures included the random effect and the fixed effect for phase, indicating that performance in these domains did not vary across experience overall or as a function of experience and phase. F-tests for the fixed effect of phase were significant for separation [$F(4, 184) = 25.61$, $p < .001$] and phraseology, $F(4, 184) = 20.61$, $p < .001$. Procedures performance did not benefit from mixed effect modeling [$\chi^2(1) = 0.53$, $p = .47$]; therefore, linear models using generalized least squares were fitted to these data. Similar to the other models, experience did not have an impact on performance, while only the effect of phase was significant in the final model, $F(4, 180) = 66.26$, $p < .001$. Trends in performance data are displayed in Fig. 3. Performance means and standard deviations are displayed in Table 2 with corresponding pairwise comparisons. In general, errors increased after Phase B, peaked during Phase D, and subsequently decreased during Phase E.

4. Discussion

While the EEG cognitive state metrics examined here have been validated in the laboratory with basic cognitive tasks (e.g., Berka et al., 2004, 2007; Johnson et al., 2011; Sciarini et al., 2014), little to no research has examined how these metrics perform in operational settings with continuous, dynamic workloads and differentially experienced operators. We tested ABM's EEG metrics of workload and engagement in more and less experienced ATC students during a variable workload approach-control scenario. In general, our results indicate that ABM's commercially available EEG workload metric provides a rough indication of ATC workload and that EEG variables indicative of cognitive engagement can be used to differentiate ATC skill levels despite no overt differences in performance.

4.1. Workload and engagement metrics across phases

Results showed that ABM's EEG average workload metric varied across scenario phases and approximated the scenario's predicted workload. Participants experienced peak workload during Phases C and D. Although this metric was able to detect peak workload, several pairwise comparisons were non-significant between phases that were

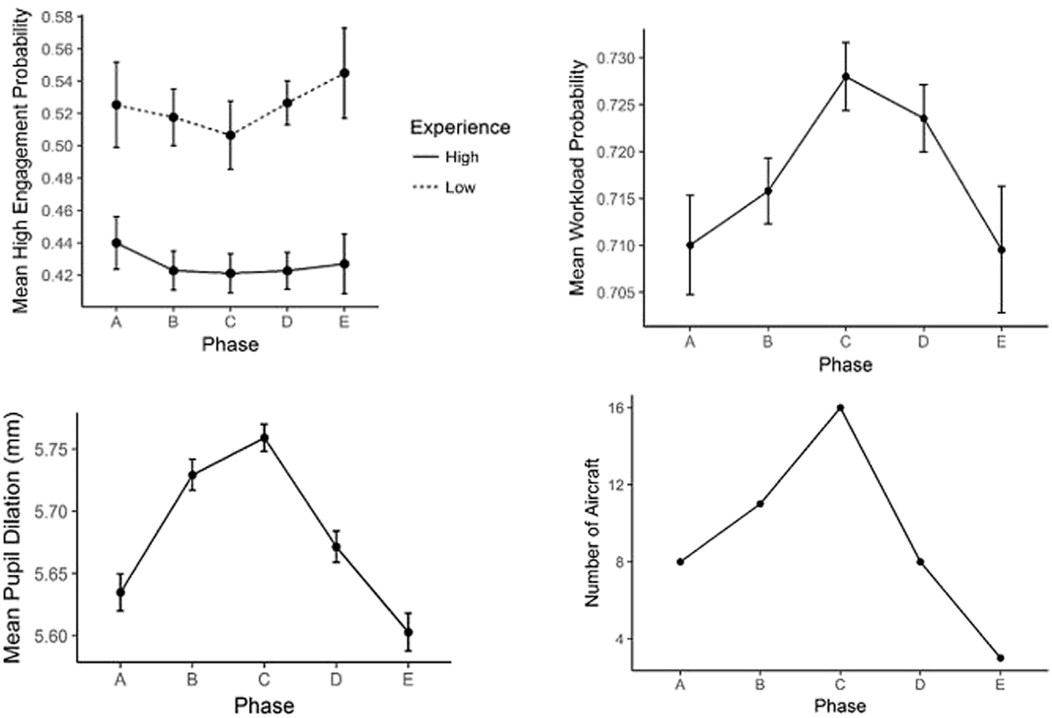

Fig. 2. Mean EEG engagement (top left panel), EEG workload (top right panel), pupil diameter (bottom left panel), and predicted workload (bottom right panel). Error bars represent 95% confidence intervals.

predicted to be higher in workload. Thus, this metric may not be able to discriminate between subtle fluctuations in workload, such as increased numbers of aircraft only. It should be noted that the non-significant comparisons were between phases with small increases in aircraft; for example, Phase A (eight aircraft) and Phase B (11 aircraft). Although Phase D released fewer aircraft than Phase C, the comparison between these two phases was not significant. This was likely the result of "overflow" from Phase C to Phase D. That is, some of the traffic from Phase C may not have been controlled adequately resulting in more traffic and increased workload in Phase D. Thus, the EEG workload metric may have reflected this residual workload.

Noticeable differences across phases in ABM's workload metric were observed when students had to provide uncontrolled clearances. Indeed, ABM's workload metric was built on forward and backward

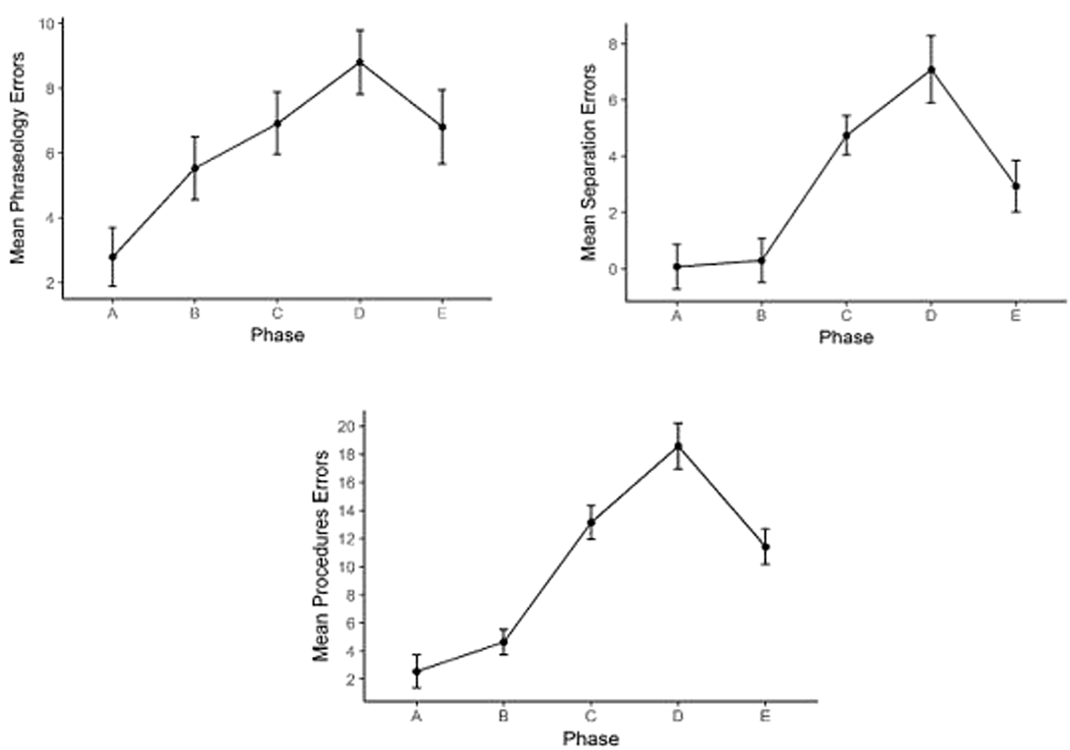

Fig. 3. Mean phraseology (top left panel), separation (top right panel), and procedures (bottom center panel) errors. Error bars represent 95% confidence intervals.

Table 2
TRACON scenario performance means, standard deviations, and pairwise comparisons by phase (N = 47).

Phase	Error Type					
	Separation		Phraseology		Procedures	
	M	*SD*	*M*	*SD*	*M*	*SD*
A	$0.09_{C, D, E}$	0.35	$2.79_{B, C, D, E}$	2.33	$2.55_{B, C, D, E}$	1.74
B	$0.30_{C, D, E}$	0.69	$5.53_{A, D}$	4.50	$4.64_{A, C, D, E}$	3.09
C	$4.74_{A, B, D, E}$	3.77	$6.91_{A, D}$	4.94	$13.15_{A, B, D}$	5.52
D	$7.09_{A, B, C, E}$	5.66	$8.81_{A, B, C}$	5.17	$18.57_{A, B, C, E}$	7.96
E	$2.94_{A, B, C, D}$	4.65	6.81_{A}	5.17	$11.43_{A, B, D}$	6.21

Note. Subscripted letters indicate the phases the specific mean is significantly different from (e.g., $0.09_{C, D, E}$ indicates that the mean is significantly different from phases C, D, and E). The Bonferroni correction was used for multiple comparisons.

digit-span tasks, which are linked with working memory processes (Advanced Brain Monitoring, 2009; Engle, 2002). One crucial component of cognitive load theories (e.g., Mayer, 2009; Sweller et al., 1998) is the concept of working memory based on the original multi-component working memory model by Baddeley (1986, 2000, 2012). According to Baddeley's model, storage components of working memory (i.e. phonological loop and visuospatial sketchpad) are separate from the attentional control system. Most cognitive load theories focus on the capacity limitations of the storage components of working memory. According to this view, if the amount of information an operator has to process at any given moment exceeds the available storage resources, a bottleneck is created and performance is hindered and may break down (Gerjets et al., 2014).

The cognitive processes used during the clearing of uncontrolled departures strains working memory and requires controllers to adequately plan when to execute the clearance in the midst of sequencing other traffic. Therefore, it is not surprising that ABM's workload metric was more sensitive when executive functioning processes were strained.

The engagement metric did not vary across scenario difficulty. The results of this study mirror those found in studies that utilized simple cognitive tasks. In the original validation of these metrics, Berka et al. (2007) reported that the engagement metric generally did not increase linearly across increasing difficulty levels of simple cognitive tasks (forward/backward digit span, grid-recall, and mental addition). Moreover, the results of this study are also consistent with other findings showing ABM's workload, but not engagement metric, to vary across increased Stroop Task demands (Sciarini et al., 2014).

4.2. Workload and engagement metrics across experience levels

Models of mental workload predict that those with less training or experience with a task should exhibit more cognitive workload, or less residual resources, than more experienced individuals (Loft et al., 2007; Tsang and Vidulich, 2006). Partially supporting our hypothesis, the engagement, not workload EEG metric, differentiated the student experience groups. Specifically, less experienced students demonstrated more engagement than their more experienced counterparts. One explanation for this finding may relate to the automaticity more experienced students have developed over the course of their training with respect to certain TRACON procedures. At the physiological level, studies have demonstrated changes in brain responses resulting from increased training (e.g., Ayaz et al., 2012; McKendrick et al., 2014). Furthermore, Hockey (1997, 2011) proposed that increased effort to maintain performance may result in a greater physiological cost. In this case, the physiological cost for less experienced individuals may be reflected in the EEG engagement metric. TRACON performance and ATC knowledge did not differ between experience groups, but EEG engagement did, suggesting an underlying cost for less experienced controllers to maintain adequate performance. Furthermore, Tsang and Vidulich (2006) suggested that the differences between novices and experts in terms of knowledge may be evident, but experts use knowledge more efficiently than novices (p. 261). This implies reduced neural activation resulting from increased neural efficiency (Kelly and Garavan, 2005).

These results are consistent with those found by Johnson et al. (2014). In their study, ABM's engagement metric distinguished between novice and expert shooters, but the workload metric mostly failed to distinguish these groups. As with the current study, novices exhibited more engagement than experts.

4.3. EEG workload and pupil diameter

Pupil diameter results were similar to the EEG workload metric but demonstrated more sensitivity to the number of aircraft released within a particular phase. As the scenario progressed, pupil diameter increased until Phase C and subsequently decreased. This suggests that ocular measures may be more sensitive to increased visual load than ABM's EEG workload metric. Previous research has suggested ocular measures to be more sensitive to workload from visual rather than strictly cognitive sources (Ryu and Myung, 2005). However, other studies have shown pupil diameter to be sensitive to other qualitatively different sources of workload including *n*-back performance (Mandrick et al., 2016) and secondary auditory driving tasks (Marquart et al., 2015), as well as possibly reflecting an overall workload construct (O'Donnell and Eggemeir, 1986). More research is needed to delineate how different psychophysiological measures relate to qualitatively different sources of workload (Tsang and Vidulich, 2006) and if these measures are related to differentiated resource theories (e.g., Wickens, 2002).

EEG workload and pupil diameter did not correlate. This result is consistent with the findings obtained by Matthews et al. (2015) who measured several physiological responses, including EEG and eye-tracking, while participants completed a variable workload unmanned ground vehicle task. Although the physiological measures were sensitive to changes in task demands, the measures demonstrated poor intercorrelations with one another, suggesting that workload is not a unitary construct. Due to our study's scenario design, it is unclear if increases in traffic alone or the presence of uncontrolled departures influenced EEG workload and pupil diameter to a similar degree. Perhaps a scenario with several phasic fluctuations in workload contributing factors in isolation could clarify this interpretation. Furthermore, although the use of pupil diameter has been established as a measure of cognitive workload (e.g., Mandrick et al., 2016; Marquart et al., 2015), pupil diameter is also related to cognitive engagement. Hopstaken et al. (2015) reported changes in pupil diameter over prolonged task performance that correspond with changes in subjective task engagement ratings. This confounded relationship between pupil diameter and other psychological constructs may explain the non-significant correlation between pupil diameter and EEG workload. In other words, pupil diameter may not necessarily be a “pure” measure of cognitive workload and could consist of other physiological responses corresponding to alertness.

4.4. Practical implications

The present findings have implications for operator training. While performance may remain similar across individuals with differing experience levels, the underlying state of the operator may indicate additional training is required to reduce the probability of inefficient stimulus processing during situations with added sources of workload (Ayaz et al., 2012). From a practitioner's perspective, “turn-key” EEG systems, such as those produced by ABM, offer cognitive state monitoring algorithms that can be used to ascertain a holistic view of a trainee's functional state. However, this study underscores the importance of practitioners understanding the limitations of the cognitive state metrics to be used (Matthews et al., 2015). For example, as shown here, two different workload metrics seemed to differentiate workload phases based on the qualitative nature of the scenario demands (e.g., visual vs. increased cognitive processing). Therefore, practitioners should initially conduct a careful evaluation of the system/scenario to be tested to select the appropriate workload metric or consider developing classification algorithms tailored to the environment of interest.

Additionally, this study advances the field of neruoergonomics by addressing the ecological validity of using commercially available cognitive state metrics for system design and evaluation. For instance, ABM's cognitive state metrics have been used in military studies conducting system evaluations on different flight system configurations (Feltman et al., 2018; McAtee et al., 2017) as well as training programs for unmanned aerial vehicle pilots (Sibley et al., 2010) and robot-assisted surgery (Guru et al., 2015). Therefore, research regarding how these metrics react to different operational conditions including task load and operator experience are important considerations for practitioners using these metrics for system evaluations. Here, we showed that in the ATC environment, ABM's workload metric is the most sensitive to greater increases in working memory demand and less so to smaller increases in workload due to other factors like increased air traffic. We also showed that ABM's engagement metric can differentiate experience levels in ATC students.

4.5. Limitations and future research

One limitation to this study was the trade-off between simulator fidelity and experimental control. Because the scenario ran continuously and aircraft were released at specified times regardless of the participant's completion of the previous phase, there may have been “workload overflow” into the next segment. That is, any decisions made by the participant in one phase likely impacted how traffic was controlled in the next phase. Evidence of this stems from notable performance decrements in the phase following projected peak workload combined with no significant differences between Phases C and D in terms of EEG workload. Additionally, strip marking performance was not evaluated. Analyzing strip marking could have revealed task offloading by some participants to allocate more resources to primary controlling. This offloading could have influenced EEG workload results.

Another limitation involved using class designations to differentiate experience levels. The differences between these two groups may not have been strong enough to elicit certain group differences (e.g., performance). Despite this limitation, a moderate to large mean difference was found between groups for the engagement metric. Thus, future studies examining cognitive state monitoring should utilize participants with a wider range of experience (e.g., students and professional controllers). This study also did not include measures of subjective workload and engagement. The rationale for this was to keep simulation fidelity at a relatively high level. Using subjective measures for each 15 min phase would have required the simulation to be stopped several times, thus interrupting the operational flow. We acknowledge that this sacrificed a level of criterion validity. However, we used pupil diameter in an attempt to compensate for this. The inclusion of subjective measures may have revealed important metric dissociations and should be considered in future studies using commercially available cognitive state metrics.

Artifacts in EEG and pupil diameter recordings could have also impacted findings. However, we took steps to minimize the effects of artifacts. For pupil diameter, ambient and screen lighting conditions in the experimental room were kept at a standardized level for all participants. We also rejected epochs during which both eyes could not be reliably detected. For EEG data, ABM's system employs algorithms that decontaminate the EEG signal from artifacts. Average data quality for the channels used to compute posterior workload and engagement probabilities was high (> 89% good signal). ABM's cognitive state metrics also use bi-polar channels in order to reduce the potential for movement artifacts (Berka et al., 2007). The pupil diameter and EEG workload results also converged upon a similar pattern and closely matched predicted air traffic control scenario workload. Additionally, EEG workload and engagement showed different patterns, indicating that the results were most likely not artifact driven.

Future research regarding physiological workload metrics should focus on defining the relative diagnosticity of these metrics in response to specific channels of workload (i.e., qualitatively different sources) under controlled manipulations that correspond to differentiated resource theories (e.g., multiple resource theory). While this interpretation of psychophysiological measures dissociating from one another due to different workload channels has been suggested (e.g., Ryu and Myung, 2005), to our knowledge, no studies have comprehensibly and experimentally evaluated this.

Acknowledgements

We would like to thank Dr. Karen Vanderzanden for her help editing this manuscript. We would also like to thank Andrew Sellner, Charles White, and Jordan Clark for being pseudo-pilots for this study. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Advanced Brain Monitoring, 2009. B-alert Live: User Manual. Carlsbad, CA.
Aricò, P., Borghini, G., Di Flumeri, G., Colosimo, A., Bonelli, S., Golfetti, A., et al., 2016. Adaptive automation triggered by EEG-Based mental workload index: a passive brain-computer interface application in realistic air traffic control environment. Front. Hum. Neurosci. 10. https://doi.org/10.3389/fnhum.2016.00539.
Aricò, P., Borghini, G., Di Flumeri, G., Bonelli, S., Golfetti, A., Graziani, I., et al., 2017. Human factors and neurophysiological metrics in air traffic control: a critical review. IEEE Rev. Biomed. Eng. 10, 250–263. https://doi.org/10.1109/rbme.2017.2694142.
Ayaz, H., Shewokis, P.A., Bunce, S., Izzetoglue, K., Willems, B., Onaral, B., 2012. Optical brain monitoring for operator training and mental workload assessment. Neuroimage 59, 36–47. https://doi.org/10.1016/j.neuroimage.2011.06.023.
Baddeley, A., 1986. Working Memory. Clarendon Press, Oxford.
Baddeley, A., 2000. The episodic buffer: a new component of working memory? Trends Cognit. Sci. 4, 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2.
Baddeley, A., 2012. Working memory: theories, models, and controversies. Annu. Rev. Psychol. 63, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422.
Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol. Bull. 91 (2), 279–292. https://doi.org/10.1037/0033-2909.91.2.276.
Beauchamp, M.H., Dagher, A., Aston, J.A.D., Doyon, J., 2003. Dynamic functional changes associated with cognitive skill learning of adapted version of the Tower of London task. Neuroimage 20 (3), 1649–1660. https://doi.org/10.1016/j.neuroimage.2003.07.003.
Berka, C., Levendowski, D.J., Cvetinovic, M.M., Petrovic, M.M., Davis, G., Lumicao, M.N., et al., 2004. Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. Int. J. Hum. Comput. Interact. 17 (2), 151–170. https://doi.org/10.1207/s15327590ijhc1702_3.
Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Olmstead, R.E., et al., 2007. EEG correlates of task engagement and mental workload in vigilance, learning, and memory. Aviat. Space Environ. Med. 78 (5), B231–B243.
Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F., 2014. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neurosci. Biobehav. Rev. 44, 58–75. https://doi.org/10.1016/j.neubiorev.2012.10.003.
Buysse, D.J., Reynolds III., C.F., Monk, T.H., Berman, S.R., Kupfer, D.J., 1989. The Pittsburgh sleep quality index: a new instrument for psychiatric practice and research. Psychiatr. Res. 28 (2), 193–213. https://doi.org/10.1016/0165-1781(89)90047-4.
Dasari, D., Shou, G., Ding, L., 2017. ICA-derived EEG correlates to mental fatigue, effort, and workload in a realistically simulated air traffic control task. Front. Neurosci. 11. https://doi.org/10.3389/fnins.2017.00297.
Doppelmayr, M., Finkenzeller, T., Sauseng, P., 2008. Frontal midline theta in the pre-shot phase of rifle shooting: differences between experts and novices. Neuropsychologia 46 (5), 1463–1467. https://doi.org/10.1016/j.neuropsychologia.2007.12.026.
Engle, R.W., 2002. Working memory capacity as executive attention. Curr. Dir. Psychol. Sci. 11 (1), 19–23. https://doi.org/10.1111/1467-8721.00160.
Feltman, K.A., Curry, I., Bernhardt, K.A., Hayes, A., Kirby, S., Davis, B., et al., 2018. Integrated Cueing Environment: Simulation Event Four. USAARL Tech Report No. 2018-16.
Freeman, F.G., Mikulka, P.J., Prinzel, L.J., Scerbo, M.W., 1999. Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. Biol. Psychol. 50 (1), 61–76. https://doi.org/10.1016/S0301-0511(99)00002-2.

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., Zander, T.O., 2014. Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. Front. Neurosci. 8, 1–21. https://doi.org/10.3389/fnins.2014.00385.

Gevins, A., Smith, M.E., 2003. Neurophysiological measures of cognitive workload during human-computer interaction. Theor. Issues Ergon. Sci. 4 (1–2), 113–131. https://doi.org/10.1080/14639220210159717.

Giraudet, L., Imbert, J.-P., Bérenger, M., Tremblay, S., Causse, M., 2015. The neuroergonomic evaluation of human machine interface design in air traffic control using behavioral and EEG/ERP measures. Behav. Brain Res. 294, 246–253. https://doi.org/10.1016/j.bbr.2015.07.041.

Guru, K.A., Esfahani, E.T., Raza, S.J., Bhat, R., Wang, K., Hannomd, Y., et al., 2015. Cognitive skills assessment during robot-assisted surgery: separating the wheat from the chaff. Br. J. Urol. 115, 166–174. https://doi.org/10.1111/bju.12657.

Hairston, W.D., Whitaker, K.W., Ries, A.J., Vettel, J.M., Bradford, J.C., Kerick, S.E., McDowell, K., 2014. Usability of four commercially-oriented EEG systems. J. Neural. Eng. 11, 1–14. https://doi.org/10.1088/1741-2560/11/4/046018.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9.

Hockey, G.R.J., 1997. Compensatory control in the regulation of human performance under stress and high workload: a cognitive-energetical framework. Biol. Psychol. 45, 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4.

Hockey, G.R.J., 2011. A motivational control theory of cognitive fatigue. In: Ackerman, P.L. (Ed.), Cognitive Fatigue: Multidisciplinary Perspectives on Current Research and Future Applications. American Psychological Association, Washington, DC, pp. 167–187.

Hockey, G.R.J., Nickel, P., Roberts, A.C., Roberts, M.H., 2009. Sensitivity of candidate markers of psychophysiological strain to cyclical changes in manual control load during simulated process control. Appl. Ergon. 40, 1011–1018. https://doi.org/10.1016/j.apergo.2009.04.008.

Holm, A., Lukander, K., Korpela, J., Sallinen, M., Müller, K.M., 2009. Estimating brain load from the EEG. Sci. World J. 9, 639–651. https://doi.org/10.1100/tsw.2009.83.

Hopstaken, J.F., van der Linden, D., Bakker, A.B., Kompier, M.A.J., 2015. The window of my eyes: task disengagement and mental fatigue covary with pupil dynamics. Biol. Psychol. 110, 100–106. https://doi.org/10.1016/j.biopsycho.2015.06.013.

Johnson, R.R., Popovic, D.P., Olmstead, R.E., Stikic, M., Levendowski, D.J., Berka, C., 2011. Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. Biol. Psychol. 87 (2), 241–250. https://doi.org/10.1016/j.biopsycho.2011.03.003.

Johnson, R.R., Stone, B.T., Miranda, C.M., Vila, B., James, L., James, S.M., et al., 2014. Identifying psychophysiological indices of expert vs. novice performance in deadly force judgment and decision making. Front Hum. Neurosci 8, 1–13. https://doi.org/10.3389/fnhum.2014.00512.

Kamzanova, A.T., Matthews, G., Kustubayeva, A.M., Jakupov, S.M., 2011. EEG indices to time-on-task effects and to a workload manipulation (cueing). *World Academy of Science, Engineering, and Technology*. Int. J. Psychol. Behav. Sci. 5 (8), 928–931.

Kearney, P., Li, W., Yu, C., Braithwaite, G., 2018. The impact of alerting designs on air traffic controller's eye movement patterns and situation awareness. Ergonomics. https://doi.org/10.1080/00140139.2018.1493151.

Kelly, A.M.C., Garavan, H., 2005. Human functional neuroimaging of brain changes associated with practice. Cerebr. Cortex 15 (8), 1089–1102. https://doi.org/10.1093/cercor/bhi005.

Loft, S., Sanderson, P., Neal, A., Mooij, M., 2007. Modeling and predicting mental workload in en route air traffic control: critical review and broader implications. Hum. Factors 49 (3), 376–399. https://doi.org/10.1518/001872007X197017.

Mandrick, K., Peysakhovich, V., Rémy, F., Lepron, E., Causse, M., 2016. Neural and psychophysiological correlates of human performance under stress and high mental workload. Biol. Psychol. 121, 62–73. https://doi.org/10.1016/j.biopsycho.2016.10.002.

Marquart, G., de Winter, J., 2015. Workload assessment for mental arithmetic tasks using the task-evoked pupillary response. Peer. J. Comput. Sci. 1 (e16), 1–20. https://doi.org/10.7287/peerj.preprints.1105v2.

Marquart, G., Cabrall, C., de Winter, J., 2015. Review of eye-related measures of drivers' mental workload. Procedia Manufact. 3, 2854–2861. https://doi.org/10.1016/j.promfg.2015.07.783.

Matthews, G., Campbell, S.E., Falconer, S., Joyner, L.A., Huggins, J., Gilliland, K., et al., 2002. Fundamental dimensions of subjective state in performance settings: task engagement, distress, and worry. Emotion 2 (4), 315–340. https://doi.org/10.1037/1528-3542.2.4.315.

Matthews, G., Reinerman-Jones, L.E., Barber, D.J., Abich, J., 2015. The psychometrics of mental workload: multiple measures are sensitive but divergent. Hum. Factors 57 (1), 125–143. https://doi.org/10.1177/0018720814539505.

Mayer, R.E., 2009. Multimedia Learning, second ed. Cambridge University Press, New York, NY.

McAtee, A., Russell, D., Feltman, K., Swanberg, D.E., Statz, J.K., Ramiccio, J., Harding, T.H., 2017. Integrated Cueing Environment Testing: Pilot Cueing Synergies for Degraded Visual Environments. USAARL Tech Report No. 2017-.4.

McKendrick, R., Ayaz, H., Olmstead, R., Parasuraman, R., 2014. Enhancing dual-task performance with verbal and spatial working memory training: continuous monitoring of cerebral hemodynamics with NIRS. Neuroimage 85 (Part 3), 1014–1026. https://doi.org/10.1016/j.neuroimage.2013.05.103.

Morgan, B., D'Mello, S., Abbott, R., Radvansky, G., Haass, M., Tamplin, A., 2013. Individual differences in multitasking ability and adaptability. Hum. Factors 55 (4), 776–788. https://doi.org/10.1177/0018720812470842.

Nebes, R.D., Buysse, D.J., Halligan, E.M., Houck, P.R., Monk, T.H., 2009. Self-reported sleep quality predicts poor cognitive performance in healthy older adults. J. Gerontol. B Psychol. Sci. Soc. Sci. 64 (2), 180–187. https://doi.org/10.1093/geronb/gbn037.

O'Donnell, R.D., Eggemeri, F.T., 1986. Workload assessment methodology. In: Boff, K.R., Thomas, J.P. (Eds.), Handbook of Perception and Human Performance: Vol. 2. Cognitive Processes and Performance. Wiley, Oxford, UK, pp. 1–49.

Parasuraman, R., 2015. Neuroergonomic perspectives on human systems integration: mental workload, vigilance, adaptive automation, and training. In: Boehm-Davis, D.A., Durso, F.T., Lee, J.D. (Eds.), APA Handbook of Human Systems Integration. American Psychological Association, Washington, DC, pp. 163–176.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team, 2017. *nlme: Linear and nonlinear mixed effects models*. R package version 3.31-131. Retrieved from. https://CRAN.R-project.org/package = nlme.

R Core Team, 2017. R: A Language and Environment for Statistical Computing (Version 3.4.0). R Foundation for Statistical Computing, Vienna, Austria.

Reid, G.B., Nygren, T.E., 1988. The subjective workload assessment technique: a scaling procedure for measuring mental workload. Adv. Psychol. 52, 185–218. https://doi.org/10.1016/S0166-4115(08)62387-0.

Ryu, K., Myung, R., 2005. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. Int. J. Ind. Ergon. 35 (11), 991–1009. https://doi.org/10.1016/j.ergon.2005.04.005.

Salomon, K.A., Ferraro, F.R., Petros, T., Bernhardt, K., Rhyner, K., 2015. Individual differences in multitasking performance. Proc. Human Factor. Ergon. Soc. 59, 887–891. https://doi.org/10.1177/1541931215591263.

Sciarini, L.W., Grubb, J.D., Fatolitis, P.G., 2014. Cognitive state assessment: examination of EEG-based measures on a Stroop task. Proc. Human Factor. Ergon. Soc. 58, 215–219. https://doi.org/10.1177/1541931214581045.

Shorrock, S.T., Isaac, A., 2010. Mental imagery in air traffic control. Int. J. Aviat. Psychol. 20 (4), 309–324. https://doi.org/10.1080/10508414.2010.487008.

Sibley, C., Cole, A., Gibson, G., Roberts, D., Barrow, J., Baldwin, C., Coyne, J., 2010. Adaptive Training in an Unmanned Aerial Vehicle: Examination of Several Candidate Real- Time Metrics. Naval Research Laboratory, Washington, DC.

Sweller, J., van Merriënboer, J.J., Paas, F.F., 1998. Cognitive architecture and instructional design. Educ. Psychol. Rev. 10, 251–296. https://doi.org/10.1023/A:1022193728205.

Truschzinski, M., Betella, A., Brunnett, G., Verschure, P.F.M.J., 2018. Emotional and cognitive influences in air traffic controller tasks: an investigation using a virtual environment. Appl. Ergon. 69, 1–9. https://doi.org/10.1016/j.apergo.2017.12.019.

Tsang, P.S., Vidulich, M.A., 2006. Mental workload and situation awareness. In: Gavriel, S. (Ed.), Handbook of Human Factors and Ergonomics. John Wiley & Sons, Inc, Hoboken, NJ, pp. 243–268.

Vandenberg, S.G., Kuse, A.R., 1978. Mental rotations, a group test of three-dimensional spatial visualization. Percept. Mot. Skills 47 (2), 599–604. https://doi.org/10.2466/pms.1978.47.2.599.

Vogt, J., Hagemann, T., Kastner, M., 2006. The impact of workload on heart rate and blood pressure in en-route and tower air traffic control. Int. J. Psychophysiol 20 (4), 297–314. https://doi.org/10.1027/0269-8803.20.4.297.

Wascher, E., Rasch, B., Sanger, J., Hoffmann, S., Schneider, D., Rinkenauer, G., et al., 2014. Frontal theta activity reflects distinct aspects of mental fatigue. Biol. Psychol. 96, 57–65. https://doi.org/10.1016/j.biopsycho.2013.11.010.

Wechsler, D., 1997. WAIS-III: Wechsler Adult Intelligence Scale. Psychological Corporation, San Antonio, TX.

West, B.T., Welch, K.B., Gałecki, A.T., 2015. Linear Mixed Models: A Practical Guide Using Statistical Software. CRC Press, Boca Raton, FL.

Wickens, C.D., 2002. Multiple resources and performance prediction. Theor. Issues Ergon. Sci. 3 (2), 159–177. https://doi.org/10.1080/14639220210123806.

Wickens, C., Tsang, P., 2015. Workload. In: Boehm-Davis, D.A., Durso, F.T., Lee, J.D. (Eds.), APA Handbook of Human Systems Integration. American Psychological Association, Washington, DC, pp. 277–292.

Wilson, G.F., Caldwell, J.A., Russell, C.A., 2007. Performance and psychophysiological measures of fatigue effects on aviation related tasks of varying difficulty. Int. J. Aviat. Psychol. 17 (2), 219–247. https://doi.org/10.1080/10508410701328839.