Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Multi-Attribute Task Battery configuration to effectively assess pilot performance deterioration during prolonged wakefulness

Youngsun Kong [a,*], Hugo F. Posada-Quintero [a], David Gever [b], Lia Bonacci [b], Ki H. Chon [a], Jeffrey Bolkhovsky [b]

[a] *Biomedical Engineering at the University of Connecticut, Storrs, CT, USA*
[b] *Naval Submarine Medical Research Laboratory, Groton, CT, USA*

ARTICLE INFO

Keywords:
PVT
MATB
NASA-TLX
Sleep deprivation
Performance deterioration, and task difficulty

ABSTRACT

Multi-Attribute Task Battery (MATB) simulates realistic and complex aviation-related tasks a pilot routinely performs during flight. However, past studies using MATB have been unable to consistently elicit the effects of performance degradation during prolonged wakefulness. We surmise that this is because the task difficulty level of the MATB software was set too low. The MATB test that was designed for this protocol includes four tasks: system monitoring (SYSMON), communication (COMM), resource management (RESMAN), and all three tasks performed simultaneously. Twenty subjects performed a "session" consisting of both the psychomotor vigilance task (PVT) and the MATB test every 2 h over 25 h of prolonged wakefulness (i.e., for thirteen sessions). The difficulty levels were set to low for both SYSMON and COMM tasks, medium for RESMAN, and high for completion of the three tasks simultaneously, based on NASA task load index. We then calculated the correlation between PVT and MATB indices. As in previous studies, all PVT performance measures were significantly degraded in the last 2–4 sessions of the task, relative to sessions earlier in the 25-hr period. MATB indices were highly correlated with PVT indices. Moreover, all MATB tasks showed significant performance degradation during prolonged wakefulness unlike previous studies. Finally, the considerably more difficult multitasking required by the three simultaneous tasks led to a more consistent and higher degree of performance degradation with prolonged wakefulness than did most of the single tasks, which were set to either low or medium difficulty level. Thus, our results support the hypothesis that previous work failed to show significant performance degradation with prolonged wakefulness during the MATB due to inadequate levels of task difficulty. Our results provide evidence that the MATB, if set to an appropriate level of difficulty, can be used as an alternative to the PVT to more accurately study the effects of prolonged wakefulness on performance of realistic aviation tasks.

1. Introduction

Prolonged wakefulness has widely been known to cause deterioration of working performance [1,2]. Aircraft crews, notably, often tend to experience prolonged wakefulness due to the demands of their flying schedules and multiple time-zone changes, which has resulted in several aviation accidents [3,4]. Since these types of accidents have resulted in the loss of lives and equipment, there have been many studies that explore the effects of prolonged wakefulness on the working performance of affected personnel using various human performance evaluation tools [3–8].

One of the most commonly used tools is the psychomotor vigilance task (PVT). Developed by Dinges and Powell in 1985, the PVT has become one of the standard methods for measuring working performance during prolonged wakefulness. The PVT measures reaction time (typically via a mouse click or button press) to a visual stimulus on a screen [9]. Its simplicity reduces learning effects and has facilitated the study of working performance during prolonged wakefulness. However, the simplicity of PVT is also its weakness, as it is thought to be inadequate for evaluating other factors, such as working memory and multitasking performance. Moreover, the PVT requires participants to use only visual modality, whereas the MATB provides more sensory modality, including visual and auditory stimuli, allowing expanded interpretation of research. Thus, it is now common for researchers to perform experiments using the PVT alongside other task-performing tools [10, 11].

* Corresponding author. 407 Engineering and Science Building, 67 North Eagleville Road, Storrs, CT, 06269, USA.
E-mail address: youngsun.kong@uconn.edu (Y. Kong).

https://doi.org/10.1016/j.imu.2021.100822
Received 13 August 2021; Received in revised form 6 December 2021; Accepted 14 December 2021
Available online 17 December 2021

In 1992, the National Aeronautics and Space Administration (NASA) created a performance evaluation simulator called the Multi-Attribute Task Battery (MATB) [12]. The MATB is now a widely accepted computer simulator for evaluating the user's performance when simultaneously carrying out multiple, complicated tasks for both pilot and non-pilot subjects. Specifically, the MATB requires simultaneous management of up to four tasks over a short span of time. The MATB has been used to evaluate human performance during prolonged wakefulness for aircraft crews and to assess the effects of medications that cause drowsiness [10,11,13–16]. While the MATB does provide more complexity and varied measurement compared with the PVT, it requires careful configuration of task difficulty levels to avoid learning effects. Furthermore, while the MATB has shown promise for detecting performance degradation due to prolonged wakefulness during complex tasks, several studies have resulted in contradictory or unclear outcomes regarding the number and type of MATB tasks to use for assessment [10,11,16,17]. For example, some of MATB tasks in those studies did not show significant performance degradation during prolonged wakefulness.

Despite the equivocal results regarding the utility of MATB for identifying performance degradation, we hypothesize that some of these conflicting results may be due to variation in how the difficulty levels were configured for each task. Namely, if the difficulty of the tasks was too easy, then learning effects or overall low cognitive effort required by the task may explain the results. In order to shed some light on the conflicting results surrounding the MATB, we configured various parameters (e.g., the frequency of operative tasks) that affect difficulty levels for the various MATB tasks and compared performance on these tasks with PVT performance over a 25-h period of wakefulness. In carefully configuring task difficulty levels, we aimed to show that the MATB can be an appropriate performance evaluation tool for prolonged wakefulness studies that aim to determine the effect of fatigue on more realistic tasks, such as those performed by fighter pilots [3].

2. Methods

2.1. Study protocol

The study protocol was approved by the Institutional Review Board of the University of Connecticut. A total of 20 healthy volunteers were recruited (13 male, 7 female, aged 19–32 years). Our power analysis suggests that 20 subjects allow greater than 95% confidence interval to observe a significant effect ($p < 0.05$) [10,14,18]. Consent forms were collected on the day of each experiment, along with screening questionnaires to confirm that volunteers had no medical issues such as a seizure disorder or any acute illness. The experiment was conducted on the Storrs campus of the University of Connecticut, and all volunteers stayed in the experimental room with experimenters for the duration of the study. Participants were asked to document approximate hours slept and keep their sleep schedules as consistent as possible for a week prior to their experimental sessions to ensure the absence of sleep abnormalities. They were also asked to avoid food or drink containing stimulants such as caffeine for two days prior. Experimenters asked volunteers to confirm that these sleep and dietary constraints were met before the session began. All participants were asked to wake up at 6:00 a.m. on the morning of the study and arrive at the testing site within 4 h. Participants had 6.39 ± 1.16 h of sleep per day before experiments.

Fig. 1 shows a flow chart of our experiment. Participants performed a set of tasks ("session") every 2 h over 25 h, for a total of 13 sessions. For each session, the PVT was performed, followed by three single MATB tasks (the order of which was randomized for each session), and finally the simultaneous three MATB tasks, as shown in Table 1. Participants were given a 30-min MATB practice session within the three days prior to their experimental day, and a 10-min practice session before their experiment began, in order to minimize learning effects. Note that we analyzed MATB performance of thirteen main sessions 1–13 and 3–13 without the practice sessions. The learning effects can still be observed in some tasks. At the beginning of each session, participants were asked to report sleepiness score (i.e., how much they felt sleepiness between 1 and 10 scales). Feedback on performance was given after each task in the form of summary MATB statistics to encourage optimal performance on each session over the 25-h period.

2.2. PVT indices

The 10-min PVT was performed using PC-PVT, a MATLAB-based software [19]. For the PVT experiment, participants were asked to click the left mouse button as fast as possible when a timer displaying the elapsed time from start of stimulus randomly appeared in the center of

Table 1
Study protocol.

Time (Minutes)	10	10	10	10	15
Procedure	PVT	MATB Single Task 1	Single Task 2	Single Task 3	Simultaneous 3 Tasks

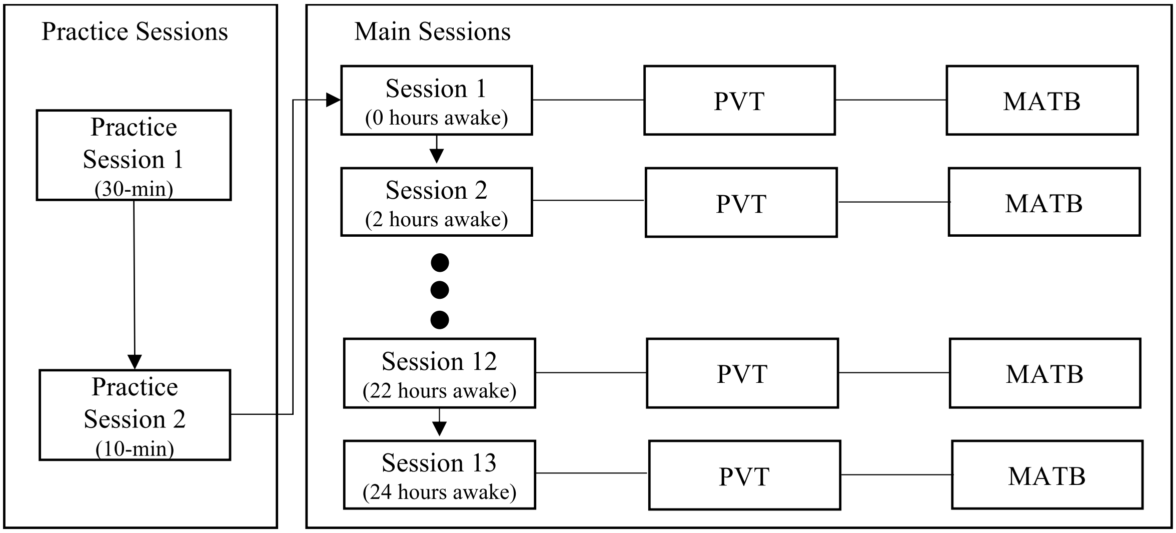

Practice Sessions
Practice Session 1 (30-min)
Practice Session 2 (10-min)
Main Sessions
Session 1 (0 hours awake)
Session 2 (2 hours awake)
Session 12 (22 hours awake)
Session 13 (24 hours awake)
PVT
MATB

Fig. 1. Flow chart of our experiment.

the computer screen. We calculated four PVT indices to measure deterioration in working performance between sessions: average reaction time across all elapsed time in the session; the number of major lapses, defined as the number of responses made at least 1000 ms after stimulus onset; the number of minor lapses, defined as the number of responses made between 500 and 1000 ms after stimulus onset; and the number of false starts (FS), defined as the number of times a participant made a mouse button click before the stimulus appeared.

2.3. Multi-Attribute Task Battery

The MATB includes four tasks: system monitoring (SYSMON), tracking, communications (COMM), and resource management (RESMAN). Since many studies have already shown that the tracking task is sensitive to fatigue [10,11,16], we focused on the three other tasks: SYSMON, COMM, and RESMAN. Participants used their preferred means of interacting with the tasks, either via keyboard or mouse.

2.3.1. System monitoring (SYSMON)

The SYSMON task involves simultaneous interaction with indicators of the system status, which include two buttons and four scales, as shown in Fig. 2-A. The left and right buttons must be pressed within 15 s of the buttons' color changing from green to the background color and from the background color to red, respectively. While the button colors are changing, an indicator bar (a yellow mark inside a dark blue box) on each scale is moving around the middle of the scale. Participants have to adjust the indicator bar back to the center of the scale within 10 s of its moving. We calculated three performance indices for this task—average reaction time (RT), accuracy (ACC), and false alarm (FA) rate. These indices were calculated separately for each button and the scales, as well as for overall performance on the tasks. ACC and FA rate were calculated from Eqs. (1) and (2), respectively.

$$ACC = \frac{Total\ Number\ of\ Events - Number\ of\ Missed\ Events}{Total\ Number\ of\ Events} \times 100(\%) \tag{1}$$

$$FA\ rate = \frac{Number\ of\ Incorrect\ Responses}{Total\ Number\ of\ Events} \times 100(\%) \tag{2}$$

Note that in Eq. (2), an incorrect response is when a subject interacted with a button or slider in the absence of an event.

2.3.2. Communications (COMM)

In the COMM task, shown in Fig. 2-B, random pre-recorded voice messages announced call signs, such as "NASA504," along with one of the four possible radio channels and its frequency (three integer digits and three decimal numbers). Messages were presented through speakers on the experimental computer. Participants were asked to set the correct radio channel and frequency within 15 s of the message being announced. Other call signs were announced throughout the task (non-

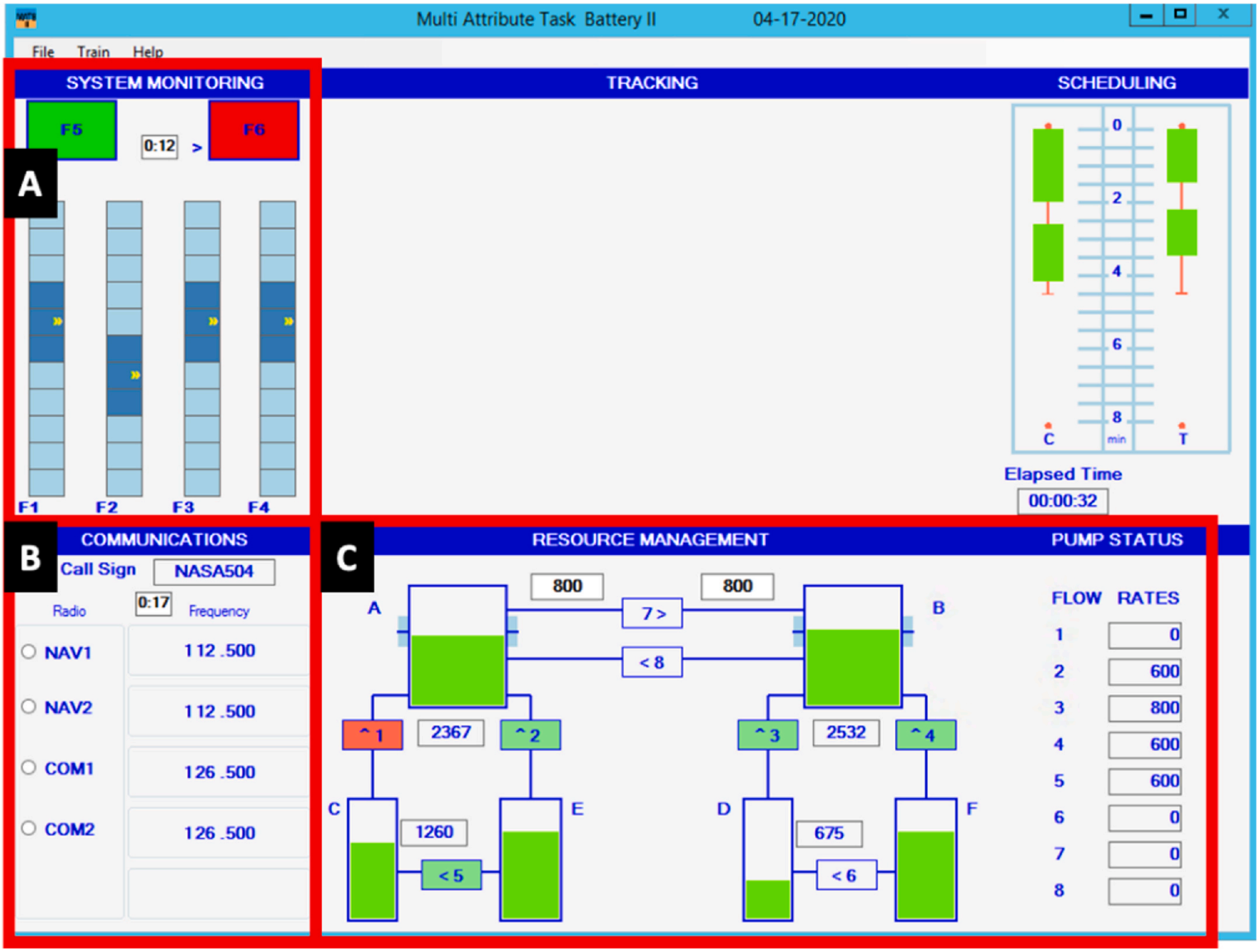


Fig. 2. An example of MATB tasks. A: System Monitoring, B: Communication, and C: Resource Management.

target messages), but participants were supposed to ignore those and respond only to messages that included their own call sign (target messages). Responses were counted as correct if the radio channel and frequency matched those in the target message. We recorded the average RT and calculated ACC and FA rate from Eqs. (3) and (4), respectively.

$$ACC = \frac{Number\ of\ Correct\ Responses}{\text{Total Number of Messages}} \times 100(\%) \tag{3}$$

$$FA\ rate = \frac{Number\ of\ Responses\ to\ Nontarget\ Messages}{Total\ Number\ of\ Nontarget\ Messages} \times 100\ (\%) \tag{4}$$

2.3.3. Resource management (RESMAN)

Fig. 2-C shows an example RESMAN task. The main goal of the RESMAN task is to maintain fluid levels in the two main tanks (Tank A and Tank B). Fluid in each of the two main tanks is consistently consumed at a rate of 800 units per minute, and must be replenished by other sub-tanks (Tanks C, D, E, and F) via a series of pumps. Tanks C and D can supply fluid directly to Tanks A and B, respectively, but their supply is limited. Tanks E and F have unlimited supply, and can pump fluid to Tanks C and D, respectively. Tanks E and F can also pump directly to Tanks A and B, respectively, but this is slower than pumping directly from limited Tanks C and D. Each pump is unidirectional and has a unique flow rate, represented by a number next to an arrow indicating flow direction. Pumps can be started or stopped by clicking or pressing the appropriate keyboard keys.

Throughout the task, pumps are broken and repaired randomly to increase the difficulty level. At most, three pumps can be broken at the same time. However, only one path to each of the main tanks may be blocked at a time. In order to quantify performance on this task, we calculated the absolute deviation from target fluid levels for Tanks A and B. Since several studies have reported that the task was insensitive to prolonged wakefulness [10,11,17], we aimed to increase the difficulty by requiring participants to maintain the fluid levels in Tanks A and B at *exactly* 2500 units rather than allowing a tolerance range *around* this target level.

2.4. Assessment of difficulty levels of the MATB tasks using NASA-TLX

In adjusting the difficulty level of each MATB task, we aimed to assign a low difficulty for the SYSMON and the COMM tasks, and medium difficulty for the RESMAN task. Moreover, we assumed that the simultaneous task inherently had the highest difficulty level, since it required subjects to perform 3 tasks at once. The difficulty of each task was heuristically determined by adjusting event frequency, as shown in Table 2, as well as appropriately setting task-specific parameters (described in the "Details" column of Table 2). We created MATLAB code to automatically generate extensible markup language (XML) files that appropriately set up these difficulty levels. Event inter-stimulus intervals were defined using a Poisson distribution where lambda was set as 60 s, divided by the total number of events. With the defined intervals, we randomly generated and permutated a total of 52 XML files for MATB tasks for each participant: 4 for each of the 13 sessions. Note that call signs for the COMM task were varied across each session by randomly selecting audio files from a pool of 80 recording files (40 own call signs and 40 other call signs).

Table 2
MATB configuration.

	Event Frequency (per minute)		Details
	Single	***Simultaneous***	
System Monitoring	6	6	
Communications	2	2	1 for NASA 504, 1 for others
Resource Management	8	6	Maximum 3 pumps can be broken at the same time Pumps do not get broken to block more than one path to tank A or B (pumps 1, 2, and 8 & pumps 3, 4, and 7)

We tested our MATB configuration using the NASA Task Load Index (NASA-TLX) in our pilot study, additional to our main study. NASA-TLX has been widely used to evaluate task performance for pilots [20–22]. The NASA-TLX measures six subjective categories—mental demand, physical demand, temporal demand, performance, effort, and frustration—on a scale from 0 to 100. Ten participants were given 4-min practice sessions for each task, and subsequently asked to assess them using the NASA-TLX [23]. Table 3 shows the average NASA-TLX scores for each category in both the individual tasks as well as the simultaneous tasks. The scores range from zero to hundred, which represent "bad" to "good" perceived performance on the task for the "Performance" category and "low" to "high" for all other categories describing the task. Averages across these categories were also calculated for an overall measure of difficulty. SYSMON and COMM showed lower average scores of 14.8 and 16.4, respectively. RESMAN showed a higher average score than the other two individual tasks but a lower score than the simultaneous tasks, with an average score of 26. The combined simultaneous tasks had the highest score with an average of 40.5. Due to the normally distributed scores, a repeated-measures one-way analysis of variance (ANOVA) was used to determine if there were significant differences among the averaged NASA-TLX scores. The Tukey's test was performed for multiple comparison. Normality of samples was determined prior to the ANOVA by the Kolmogorov-Smirnov test. The result of the ANOVA found a significant difference in average score among the four tasks ($p = 0.001$). The multi-comparison testing determined that scores for both the COMM and SYSMON tasks were significantly different than the average score for the simultaneous task ($p = 0.0264$, $p = 0.0161$, respectively, Tukey's test for 6 comparisons).

2.5. Statistics

For the 20 participants, we obtained PVT and MATB performance indices for 13 sessions, which were distributed across 25 h of prolonged wakefulness. Each subject's PVT and MATB indices were first normalized by dividing by the Euclidian norm, calculated across sessions per subject. The four PVT indices were then compared between every pair of sessions to determine when performance degrades significantly due to prolonged wakefulness. Significant differences between sessions were determined using the paired *t*-test, since PVT measures from all sessions were normally distributed. The normality of variables was checked using the Kolmogorov-Smirnov test, and we corrected for multiple comparisons using the Bonferroni procedure. We then computed Pearson and Spearman correlation coefficients between the inter-subject mean of each PVT index and each MATB index, for normally and non-normally distributed variables, respectively, both in single tasks, and simultaneous tasks. We compared these correlation coefficients between indices obtained from two session ranges (main sessions 1–13 and 3–13) to determine which MATB indices were affected by learning. Finally, for the MATB indices found to be significantly correlated with PVT scores, we compared task scores between each pair of sessions to determine when performance degrades significantly due to prolonged wakefulness. Again, since all MATB indices for these analyses were normally distributed, paired t-tests were used for this purpose, with Bonferroni correction for multiple comparisons.

3. Results

Figs. 3, 4, 6 and 7 show the mean ± SEM of sleepiness score, PVT indices, and MATB indices; significant differences between session scores are indicated by the numbers located vertically on top of each session. Participants showed slightly decreasing trends until the 4th

Table 3
Averaged (n = 10) NASA-TLX scores.

Category	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration	Average
COMM	21.8	13	15.7	13.0	23.5	11.7	16.4*
SYSMON	18.2	9.90	12.7	13.5	17.7	17.0	14.8*
RESMAN	33.7	16.5	20.9	26.5	33.1	25.4	26.0
Simultaneous	52.3	29.2	39.6	37.6	53.7	30.7	40.5

Asterisk (*) in Average indicates that the samples are significantly different from the simultaneous one.

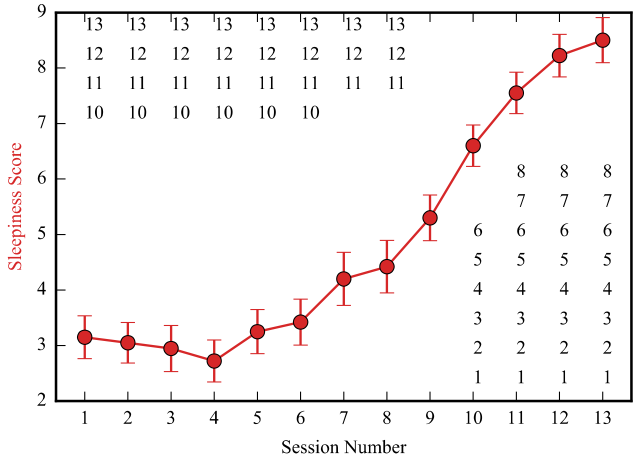


Fig. 3. Sleepiness score (1–10 scales). The column of numbers indicates between which sessions the score was significantly different ($p < 0.05$ for all comparisons, Dunn's test, Bonferroni-corrected for 78 comparisons).

session, followed by increasing trends until the last session. The last four sessions showed significant differences with the rest of sessions. The 10th session was significantly different with sessions 1–6, and the last three sessions showed significant differences with sessions 1–8.

All PVT indices generally increased with the session number; the last two and, in some cases, the last four sessions significantly differed from the previous sessions. AvRTs in the 12th and 13th sessions were significantly different than those in sessions 1–10; scores in session 11 were significantly different from those in sessions 1, 2, 5, and 7 (Fig. 4a). Likewise, for the last two sessions, major lapses (MaL) were significantly different than sessions 1–10; the scores in the 13th session was also significantly different than the 11th sessions (Fig. 4b). For the last three sessions, minor lapses (MiL) were significantly different than those of sessions 1–9 (Fig. 4c). The minor lapses in the 10th session were also significantly different from those of sessions 1–7, except for the 5th session. In the case of false starts (FS), only scores from the last two sessions were significantly different from those of the previous sessions (Fig. 4d).

Table 4 shows correlation coefficients between each of the four PVT indices and the various MATB indices. Most tasks for sessions 3 to 13 showed higher correlation coefficients than for sessions 1 to 13 likely due to learning effects, but the single COMM task did not appear to be largely affected by learning effects. This is likely because of the fact that the stimulus requires a different sensory modality, where COMM and the other two tasks rely on response to auditory and visual stimuli, respectively. Most MATB indices were significantly correlated with PVT indices. Fig. 5 shows scatterplots of the raw data of sessions 3–13 for correlation analysis between PVT AVRT and single and simultaneous MATB indices. Due to potential learning effects in MATB tasks for the first two sessions, all descriptions below are based on sessions 3 to 13.

In terms of RT, SYSMON and COMM scores were significantly correlated with all PVT indices for both single and simultaneous tasks. Likewise, almost all ACC scores for SYSMON and COMM tasks were also significantly correlated with PVT indices for both single and simultaneous tasks. Note, however, that ACC in pressing the green button in the single SYSMON task was not correlated with all PVT indices even though ACC for pressing the red button was, especially in the simultaneous task, which may be due to differences in button salience. In terms of the RESMAN tasks, all tank deviation indices for both single and simultaneous tasks were significantly correlated with all PVT indices.

We also observed that overall FA rates for both single and simultaneous SYSMON tasks were significantly correlated with all PVT indices. However, results were less clear when these measures were taken separately for each button and the scale. In particular, FA rates for the single SYSMON task's green and red buttons showed no correlation with PVT indices, but the FA rate for scale adjustment was significantly correlated with all PVT indices. In the simultaneous SYSMON task, the FA rate for pressing the green button and adjusting the scale showed almost no significant correlation with PVT indices, though correlation was found between FA rate for the green button and the PVT average RT. FA rate for pressing the red button in the simultaneous SYSMON task was correlated with all PVT indices. FA rates for both single and simultaneous COMM tasks also showed no significant correlation with PVT indices.

Fig. 6 shows the performance indices for single MATB tasks that were found to be highly correlated with PVT indices, as indicated in Table 4. Note that measures for the SYSMON task reflect overall scores, rather than scores for individual buttons (i.e., red button, green button, scale). For the SYSMON task, RT did not change between sessions 1–9, but increased in the last three sessions—RT was significantly greater in these last three sessions than in all preceding sessions. (Fig. 6a). Likewise, the ACC of the SYSMON task was stable until the 9th session but decreased by the 11th session until the 13th session, with a high standard error in this last session; however, no significant difference was exhibited between sessions (Fig. 6b), which suggests the task was too easy for ACC measures to reflect the effects of prolonged wakefulness. Fig. 6c shows FA rates for the single SYSMON task. In this task, the FA rate decreased until the 3rd session, likely due to some learning effects; however, by the 13th SYSMON session, FA rates were significantly larger than they were in the 3rd session (Fig. 6c).

Fig. 6d and e shows performance metrics for the single COMM task. Although the RT generally decreased from the first to last session, it showed no significant difference between subsequent sessions throughout the experiment (Fig. 6d). Note that this is in contrast to the PVT and SYSMON tasks, which are inherently visual tasks; therefore, the fact that we see no significant change in RT may be due to the auditory component of the COMM task. Like RT, the ACC in the COMM task also generally decreased from the first to last session, but sessions 2 and 7 had significantly different scores than sessions 11 and 12 (Fig. 6e), suggesting that correctly responding in the COMM task was more affected by prolonged wakefulness than simply reacting to the auditory cue.

Fig. 6f shows overall tank deviations for the single RESMAN task. Deviations exhibited a noticeable drop in the second session (likely due to some learning effects), but then remained stable through session 9. After session 9, tank deviations drastically increased, and the average deviation in the second-to-last session was significantly different from

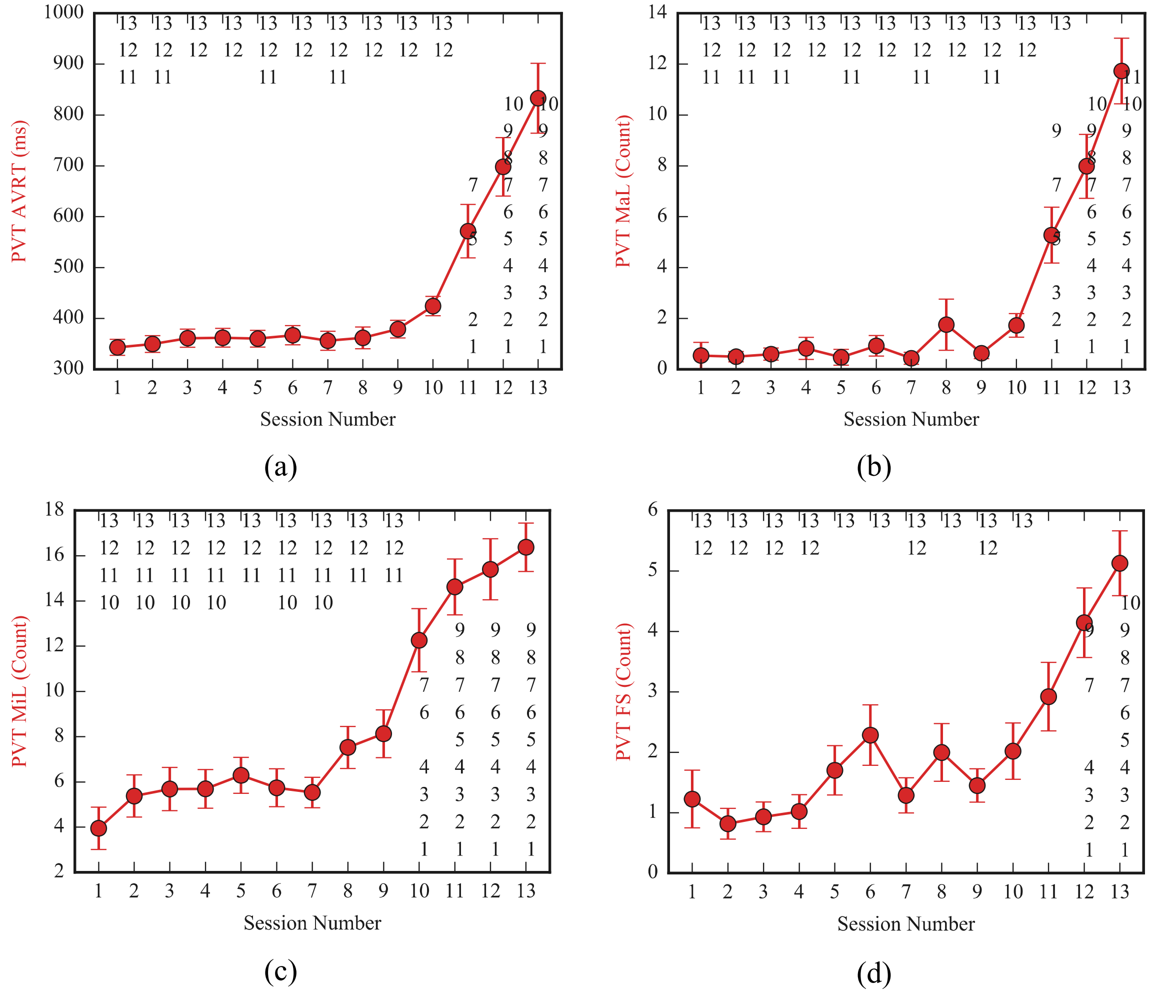


Fig. 4. PVT indices. The column of numbers indicates between which sessions the index was significantly different ($p < 0.05$ for all comparisons, *t*-test, Bonferroni-corrected for 78 comparisons). AvRT: average reaction time, MaL: major lapses, MiL: minor lapses, and FS: false starts.

that in sessions 5 and 9.

Fig. 7 shows the performance indices for simultaneous MATB tasks that were significantly correlated with PVT indices, based on the results in Table 4. RT and ACC of the simultaneous SYSMON task showed decreasing and increasing trends, respectively, in the first four sessions (Fig. 7a and b). This is in contrast to the RT and ACC scores in the single SYSMON task, which remained stable in this time interval (see Fig. 6a and b). For the simultaneous SYSMON task, RT in session 12 was significantly longer than that in sessions 2 to 9, and RT in sessions 10 and 11 was significantly longer than that in session 4 and sessions 6 to 8. While the ACC of the single SYSMON task was not significantly different between any pair of sessions (see Fig. 6b), in the simultaneous task, it was significantly lower in session 12 than in sessions 4, 7, and 9, suggesting that accuracy on tasks with increased difficulty is more affected by prolonged wakefulness compared with simpler tasks. FA rates for the simultaneous SYSMON task are shown in Fig. 7c. Note that while the FA rate in the last session of the single SYSMON task was significantly different from an earlier session (Fig. 6c), no significant difference in FA rate was found between sessions in the simultaneous version of the task; this reduction in the effect on FA may be explained by the fact that subjects' attention was more divided during the simultaneous tasks, though identifying such mechanisms is beyond the scope of this study.

RT and ACC for the simultaneous COMM task are shown in Fig. 7d and e, respectively. As in the single COMM task, no significant differences between sessions were observed for RT for the simultaneous task. However, a significant difference in ACC was observed between the 12th session and earlier sessions (3, 5, and 7) for the simultaneous COMM task. The simultaneous RESMAN task had overall higher average tank deviation scores over all sessions than the single RESMAN task (compare Figs. 6f and 7f).

The single RESMAN task results showed significant differences in tank deviation only among three sessions; however, in the simultaneous RESMAN task, the 11th and 12th sessions had significantly higher tank deviation scores compared to a large number of earlier sessions (sessions 3–7, and 9). This again is likely due to increased task difficulty.

4. Discussion

We adjusted task difficulty levels for the MATB test developed by NASA to examine the resulting effect on performance deteriorations in response to prolonged wakefulness. We used the NASA-TLX in order to assess and ensure our MATB parameters led to the desired task difficulty. Our test included four tasks, the system monitoring (SYSMON) task, the communication (COMM) task, the resource management (RESMAN) task, and the three tasks performed simultaneously. In using these tasks during prolonged wakefulness, we found performance deteriorations that were highly correlated with those observed from PVT measures, especially when the task difficulty level was set to high (i.e., simultaneous tasks) compared to those of a lower difficulty (i.e., single SYSMON and COMM tasks). We found that when carefully setting difficulty levels, the resulting MATB tasks are sensitive to the effects of prolonged wakefulness, especially with high task difficulty levels. Therefore, we

Table 4
Correlation coefficients between PVT indices and MATB indices.

				Session 1–13				Session 3–13			
				AvRT	**MaL**	**MiL**	**FS**	**AvRT**	**MaL**	**MiL**	**FS**
Single Task	**SYSMON**	RT	Green	0.92**	0.52	0.89**	0.97**	0.94**	0.94**	0.89**	0.97**
			Red	0.97**	0.75*	0.92**	0.96**	0.97**	0.94**	0.92**	0.96**
			Scale	0.96**	0.63*	0.87**	0.93**	0.97**	0.89**	0.86**	0.92**
			Overall	0.95**	0.66*	0.92**	0.97**	0.96**	0.95**	0.91**	0.97**
		ACC	Green	−0.75*	−0.59*	−0.54	−0.60*	−0.82*	−0.59	−0.56	−0.61*
			Red	−0.50	−0.60*	−0.60*	−0.50	−0.75*	−0.75*	−0.75*	−0.75*
			Scale	−0.84**	−0.73*	−0.82**	−0.85**	−0.85*	−0.82*	−0.82*	−0.84*
			Overall	−0.86**	−0.69*	−0.82**	−0.85**	−0.87**	−0.82*	−0.82*	−0.85*
		FA	Green	0.42	0.36	0.37	0.49	0.33	0.44	0.31	0.46
			Red	−0.03	0.39	0.04	0.02	0.44	0.33	0.36	0.32
			Scale	0.69*	0.45	0.78*	0.84**	0.82*	0.90**	0.86**	0.91**
			Overall	0.58*	0.38	0.68*	0.74*	0.79*	0.88**	0.82*	0.88**
	COMM	RT		−0.78*	−0.72*	−0.66*	−0.62*	−0.76*	−0.64*	−0.67*	−0.61*
		ACC		−0.92**	−0.81**	−0.86**	−0.87**	−0.92**	−0.86**	−0.84*	−0.87**
		FA		0.42	0.30	0.46	0.48	0.37	0.49	0.39	0.46
	RESMAN	TankA		0.51	0.55*	0.47	0.52	0.89**	0.75*	0.73*	0.76*
		TankB		0.66*	0.60*	0.62*	0.67*	0.97**	0.85**	0.83*	0.87**
		Overall		0.60*	0.60*	0.55*	0.60*	0.94**	0.81*	0.79*	0.83*
Simultaneous Tasks	**SYSMON**	RT	Green	0.55*	0.38	0.51	0.63*	0.87**	0.81*	0.76*	0.85**
			Red	0.72*	0.57*	0.62*	0.77*	0.88**	0.83*	0.74*	0.86**
			Scale	0.71*	0.41	0.68*	0.75*	0.89**	0.80*	0.80*	0.85**
			Overall	0.67*	0.37	0.62*	0.72*	0.91**	0.83*	0.79*	0.87**
		ACC	Green	−0.19	−0.18	−0.11	−0.21	−0.60	−0.41	−0.34	−0.47
			Red	−0.41	−0.36	−0.34	−0.40	−0.89**	−0.68*	−0.68*	−0.73*
			Scale	−0.53	−0.48	−0.56*	−0.61*	−0.90**	−0.85**	−0.86**	−0.87**
			Overall	−0.47	−0.41	−0.47	−0.53	−0.90**	−0.80*	−0.80*	−0.83*
		FA	Green	0.23	0.14	0.15	0.32	0.61*	0.56	0.41	0.58
			Red	0.77**	0.56*	0.77**	0.79**	0.94**	0.86**	0.87**	0.89**
			Scale	0.48	0.55	0.61*	0.51	0.47	0.54	0.57	0.49
			Overall	0.61*	0.67*	0.67*	0.63*	0.68*	0.69*	0.69*	0.67*
	COMM	RT		−0.85**	−0.51	−0.74*	−0.79*	−0.82*	−0.76*	−0.74*	−0.81*
		ACC		−0.91**	−0.63*	−0.82**	−0.88**	−0.96**	−0.87**	−0.86**	−0.90**
		FA		−0.41	−0.39	−0.48	−0.37	−0.35	−0.39	−0.41	−0.32
	RESMAN	TankA		0.74*	0.65*	0.67*	0.74*	0.94**	0.85**	0.81*	0.86**
		TankB		0.72*	0.52	0.62*	0.68*	0.91**	0.76*	0.75*	0.79*
		Overall		0.73*	0.65*	0.65*	0.72*	0.94**	0.82*	0.79*	0.84*

*p < 0.05, **p < 0.001. AvRT: average reaction time, MaL: major lapses, MiL: minor lapses, and FS: false starts; RT: reaction time, ACC: accuracy, and FA: false alarm.

have provided evidence that the MATB can be used to assess fatigue-related performance deterioration for pilots.

The PVT is widely used to measure working performance during prolonged wakefulness due to its inherent simplicity. However, unlike the MATB, the PVT cannot simulate more complex aviation environments, since performance metrics are based solely on the ability to react to a visual stimulus. We therefore compared MATB and PVT indices by calculating correlation coefficients as a first step to assess the feasibility of using MATB indices to measure performance deterioration in prolonged wakefulness studies. Our results showed that RT, ACC, and FA of the SYSMON task, RT and ACC of the COMM task, and the absolute deviations of RESMAN task were all significantly correlated with PVT indices, suggesting that like the PVT, MATB tasks are also sensitive to the effects of prolonged wakefulness and may be used similarly to assess fatigue.

Consistent with previous studies [24,25], we found that for subjects performing the PVT over 25 h of prolonged wakefulness, PVT indices of performance were degraded in the last 2–4 sessions, relative to indices measured in earlier sessions. In addition to performance degradation observed during the PVT, we also found significant performance deterioration based on MATB task indices. In particular, we found that RT for the SYSMON task, ACC of COMM tasks, and the absolute deviation of RESMAN task all degraded with prolonged wakefulness. Interestingly, ACC on the COMM task was deteriorated by prolonged wakefulness, but RT was not, even though both RT and ACC in both single and simultaneous COMM tasks were significantly correlated with all PVT indices. This may be because participants were able to concentrate in the single COMM fully, thus having more cognitive resources available to respond to the correct call sign (i.e., "NASA 504"). In the simultaneous sessions, as humans react faster to auditory stimuli compared to visual cues [26,27], participants responded to auditory stimuli prior to the visual stimuli and did not dedicate cognitive resources to discrimination of the auditory stimulus, which purely shows the learning curve. Therefore, it may be the case that, while still correlated, the response to auditory cues is less affected by prolonged wakefulness than the response to visual PVT stimuli [28]. In other words, while response time to auditory cues may decrease initially due to learning effects and increase slightly over time due to fatigue, the effects of prolonged wakefulness on this measure are largely negligible, with the accuracy of interpreting the instruction (i.e., channel and frequency) being more susceptible. However, further experiments are needed to fully shed light on how vision and audition are differentially affected by prolonged wakefulness. Also, in future studies, the response time limit can be considered, in addition to the frequency of the events, in order to increase the difficulty level.

Previous studies have shown that there are learning effects associated with the MATB tasks [11,17]. In order to examine this further, we compared MATB indices with PVT indices for two time intervals: one that included the first two sessions, and one that did not. As shown in Table 4, correlations between most pairs of indices were higher for the interval that excluded the first two sessions. In particular, the single RESMAN and SYSMON tasks, and the simultaneous MATB tasks, were affected by this learning effect. Note that while we took measures to reduce this learning effect through practice sessions prior to the experiment, the effect was still noticeable, likely due to the fact that no MATB performance threshold was required before beginning the session. Future studies should carefully account for these learning effects when interpreting differences in MATB performance over time. Considering that simultaneous RESMAN was set to be difficult, exhibited significant

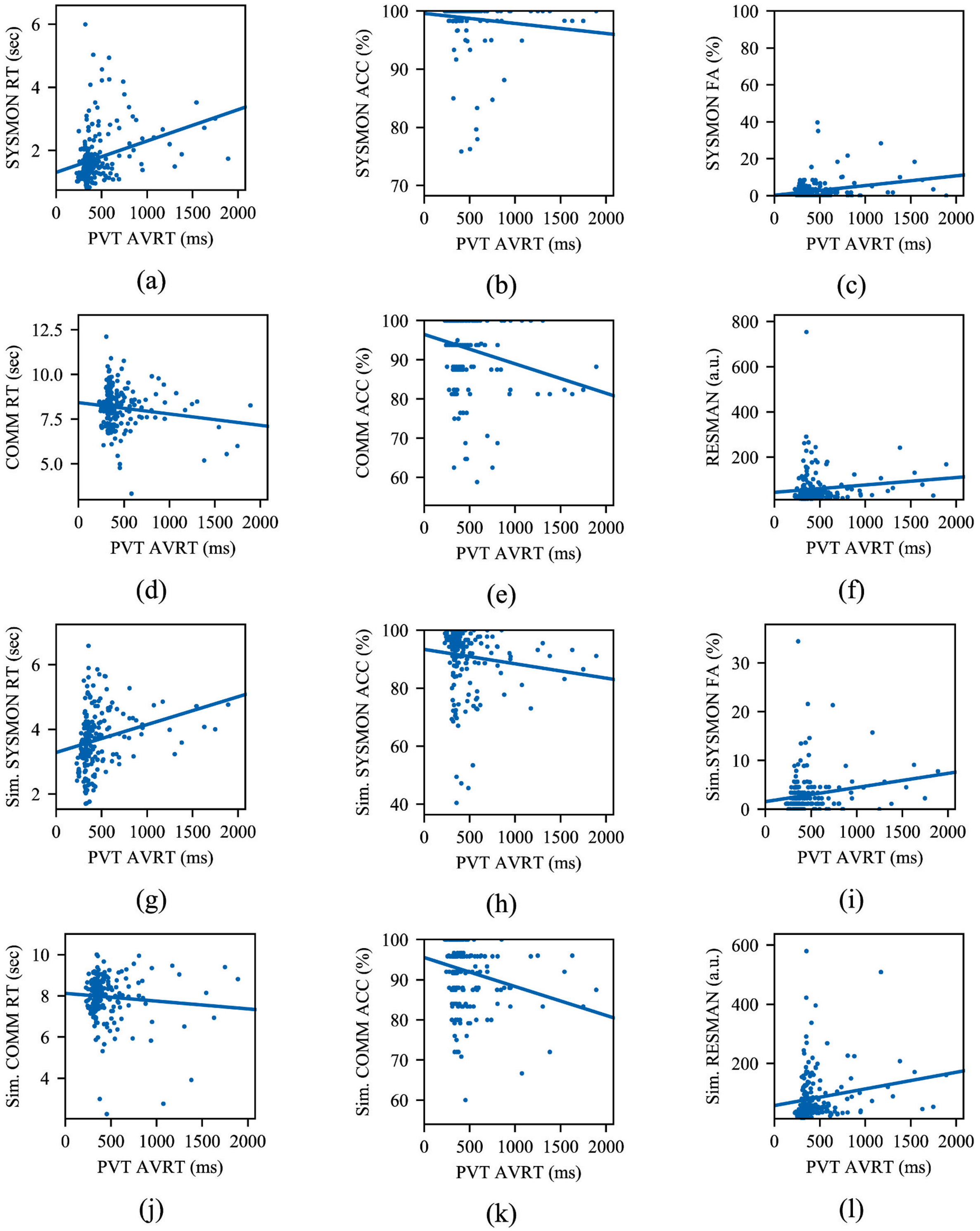


Fig. 5. Scatterplots of the raw data of sessions 3–13 for correlation analysis between PVT AVRT and (a–f) single and (g–l) simultaneous MATB indices. AvRT: average reaction time.

differences between sessions 3–9 and 11–12, we recommend at least two full sessions to minimize learning effects.

A few previous studies have already reported that performance on some MATB tasks was significantly affected during prolonged wakefulness. In particular, researchers have found that performance on the MATB tracking task (not tested here), as well as RT on the SYSMON task, and ACC on the COMM task were significantly affected by fatigue caused by either prolonged wakefulness or medication that induces similar effects [10,11,14–16]. Although aiming their MATB task to be very difficult, these studies did provide MATB parameters that affected the difficulty levels of the tasks. Our results serve to reproduce these findings while also outlining the specific parameters used to induce these effects.

Notably, no previous study has reported that the RESMAN task was affected by fatigue during prolonged wakefulness. Assuming the studies used the default configuration for RESMAN (no pumps broken), we

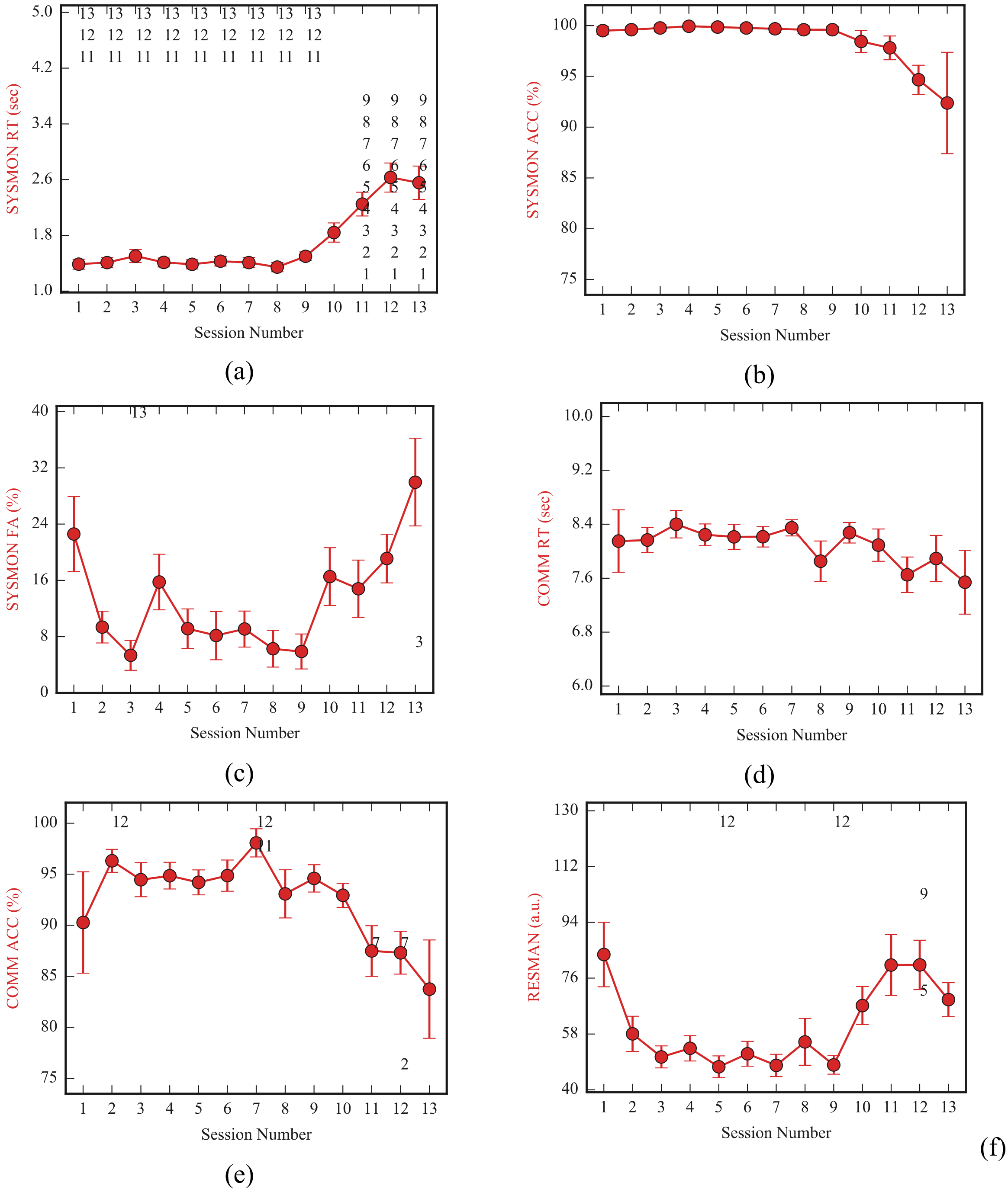
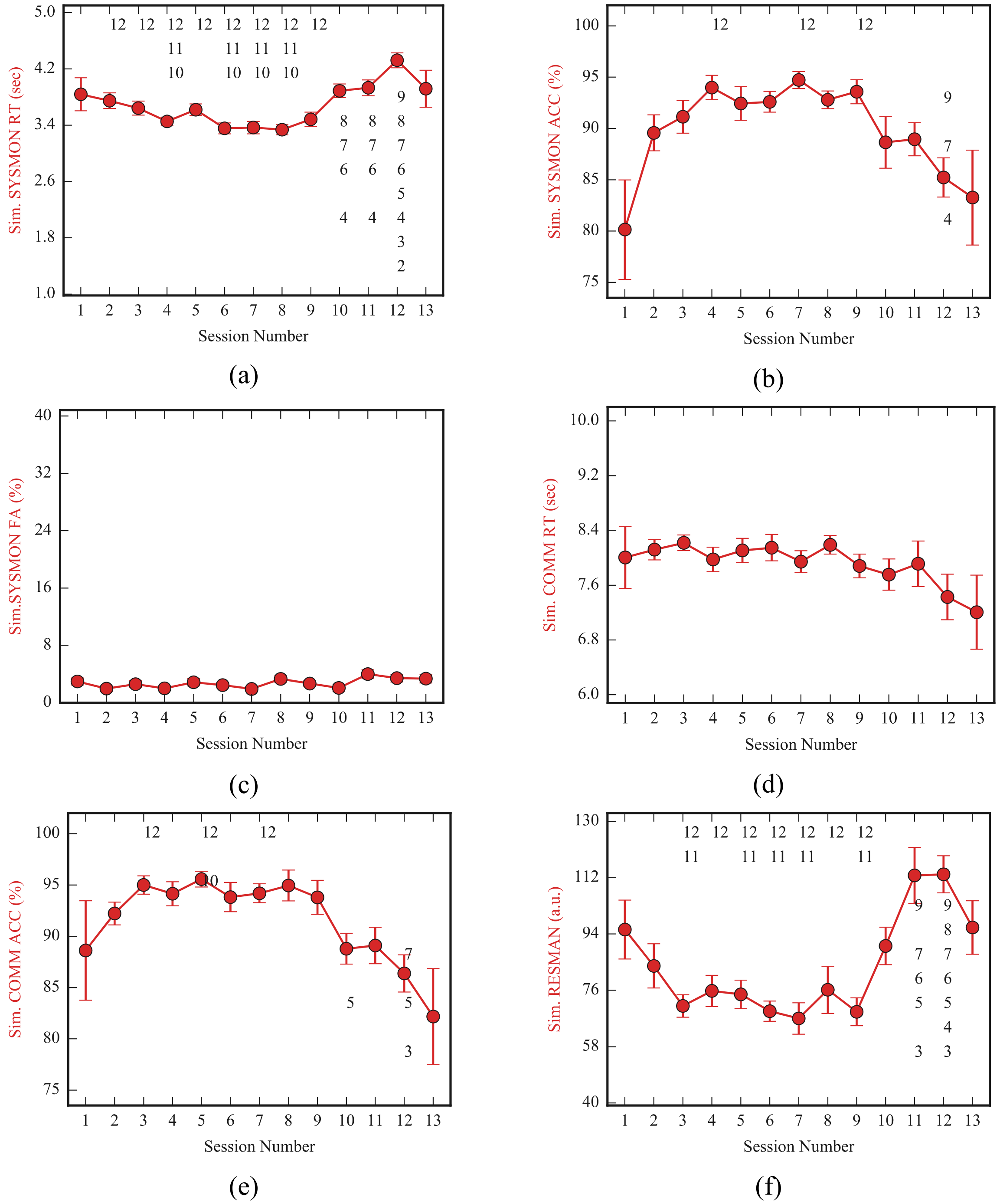


Fig. 6. Single MATB task indices significantly correlated with PVT indices. The column of numbers indicates between which sessions the index was significantly different ($p < 0.05$ for all comparisons, *t*-test, Bonferroni-corrected for 78 comparisons) (a) reaction time, (b) accuracy, (c) false alarm rate of SYSMON, (d) reaction time, (e) accuracy of COMM, and (f) tank deviation of RESMAN.

presume that the tasks did not reach a level of difficulty to elicit performance deterioration. Only Caldwell and Ramspott revealed their configuration for RESMAN: 1) only pumps 2 and 4 failed once every 2 min, and 2) target main tank levels were 2500 within a range of approximately 300 [17]. Based on our findings, their configuration for the RESMAN task was likely not difficult enough to induce significant deterioration in performance over the sessions performed during prolonged wakefulness. Therefore, when using the MATB to study the effects of prolonged wakefulness, future experiments should carefully account for these settings.

Circadian rhythm has also been shown to affect cognitive performance [29]. In this study, we observed a slight decrease in RT for both single and simultaneous SYSMON tasks at the last session compared with the second-to-last session, though this effect was not statistically significant. We also observed a slight decrease in tank deviation for both single and simultaneous RESMAN tasks at the final session. While these performance enhancements may reflect effects of circadian phase, they may also be explained by subjects anticipating the end of the study because the participants were able to access the time. In future studies, external factors such as light control and access to time must be controlled to further examine the circadian phase. We did not observe any effects of circadian phase in PVT indices, which may be



(a)

(b)

(c)

(d)

(e)

(f)

Fig. 7. Simultaneous MATB task indices significantly correlated with PVT indices. The column of numbers indicates between which sessions the index was significantly different ($p < 0.05$ for all comparisons, *t*-test, Bonferroni-corrected for 78 comparisons). (a) Reaction time, (b) accuracy, (c) false alarm rate of SYSMON, (d) reaction time, (e) accuracy of COMM, and (f) tank deviation of RESMAN, during simultaneous MATB tasks.

because 25-h prolonged wakefulness is not enough to observe circadian rhythm in PVT indices with our experimental environment.

5. Conclusion

In this paper, we performed and analyzed experiments using both the PVT and the MATB during 25-h prolonged wakefulness. Unlike other studies, we configured the MATB to have varying difficulty levels and found that most MATB indices were significantly correlated with PVT indices. Many MATB and PVT indices showed performance deterioration over time; however, SYSMON RT, PVT RT, and RESMAN tank deviation most noticeably exhibited significant deterioration over the course of the sessions. We also found that performing the three tasks simultaneously resulted in many additional indices indicating performance degradation during prolonged wakefulness when compared to single tasks. We therefore conclude that the MATB is an effective tool to analyze performance deterioration during prolonged wakefulness as well as PVT, but with the added benefit of providing more realistic aviation cockpit simulation scenarios. With appropriate difficulty levels set for the MATB, it can be used as a good alternative simulation tool to study the effects of prolonged wakefulness on aviation pilots.

Author contributions

Conceived and designed the analysis Y.K., H.P., K.C., J.B.; Collected the data, Y.K. and H.P.; Contributed data or analysis tools: Y.K., D.G.; Performed the analysis: Y.K., H.P. L.B. Original draft preparation: Y.K. Review and Editing: H.P., L.B. K.C., J.B. All authors have read and agreed to the published version of the manuscript.

Disclaimers

The views expressed in this article reflect the results of research conducted by the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the United States Government. Institutional Review Board in compliance with all applicable Federal regulations governing the protection of human subjects. The authors are federal and contracted employees of the United States government. This work was prepared as a part of official duties. Title 17 U.S C. 105 provides that copyright protection under this title is not available for any work of the United States Government. Title 17 U.S C. 101 defines a U.S. Government work as work prepared by a military service member or employee of the U.S. Government as part of that person's official duties. The study protocol was approved by the Naval Submarine Medical Research Laboratory.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Military Operational Medicine Research Program work Unit Number F1710.

References

[1] Alhola P, Polo-Kantola P. Sleep deprivation: impact on cognitive performance. Neuropsychiatric Dis Treat 2007;3:553–67.

[2] Basner M, Mollicone D, Dinges DF. Validity and sensitivity of a brief psychomotor vigilance test (PVT-B) to total and partial sleep deprivation. Acta Astronaut 2011; 69:949–59. https://doi.org/10.1016/j.actaastro.2011.07.015.

[3] Caldwell JA. Crew schedules, sleep deprivation, and aviation performance. Curr Dir Psychol Sci 2012;21:85–9. https://doi.org/10.1177/0963721411435842.

[4] Caldwell JA, Ramspott S. Effects of task duration on sensitivity to sleep deprivation using the multi-attribute task battery. Behav Res Methods Instrum Comput 1998; 30:651–60. https://doi.org/10.3758/BF03209483.

[5] Caldwell Jr JA, Caldwell JL, Brown DL, Smith JK. The effects of 37 hours of continuous wakefulness on the physiological arousal, cognitive performance, self-reported mood, and simulator flight performance of F-117A pilots. Mil Psychol 2004;16:163–81.

[6] Comstock JR, Arnegard RJ. The multi-attribute task battery for human operator workload and strategic behavior research. 1992.

[7] Daley MS, Gever D, Posada-Quintero HF, Kong Y, Chon K, Bolkhovsky JB. Machine learning models for the classification of sleep deprivation induced performance impairment during a psychomotor vigilance task using indices of eye and face tracking. Front. Artif. Intell. 2020;3:17.

[8] Dijk D-J, Duffy JF, Czeisler CA. Circadian and sleep/wake dependent aspects of subjective alertness and cognitive performance. J Sleep Res 1992;1:112–7. https://doi.org/10.1111/j.1365-2869.1992.tb00021.x.

[9] Dinges DF, Powell JW. Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations. Behav Res Methods Instrum Comput 1985;17:652–5.

[10] Faul F, Erdfelder E, Lang A-G, Buchner A. G* Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 2007;39:175–91.

[11] Griffith CD, Mahadevan S. Human reliability under sleep deprivation: derivation of performance shaping factor multipliers from empirical data. Reliab Eng Syst Saf 2015;144:23–34. https://doi.org/10.1016/j.ress.2015.05.004.

[12] Hart SG, Staveland LE. Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Hancock PA, Meshkati N, editors. Advances in psychology, human mental workload; 1988. p. 139–83. https://doi.org/10.1016/S0166-4115(08)62386-9. North-Holland.

[13] Hartzler BM. Fatigue on the flight deck: the consequences of sleep loss and the benefits of napping. Accid Anal Prev 2014;62:309–18. https://doi.org/10.1016/j.aap.2013.10.010.

[14] Jain A, Bansal R, Kumar A, Singh KD. A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. Int. J. Appl. Basic Med. Res. 2015;5:124.

[15] Jose S, Gideon Praveen K. Comparison between auditory and visual simple reaction times. Neurosci Med 2010. 2010.

[16] Jung CM, Ronda JM, Czeisler CA, Wright Jr KP. Comparison of sustained attention assessed by auditory and visual psychomotor vigilance tasks prior to and during sleep deprivation. J Sleep Res 2011;20:348–55.

[17] Khitrov MY, Laxminarayan S, Thorsley D, Ramakrishnan S, Rajaraman S, Wesensten NJ, Reifman J. PC-PVT: a platform for psychomotor vigilance task testing, analysis, and prediction. Behav Res Methods 2014;46:140–7. https://doi.org/10.3758/s13428-013-0339-9.

[18] Kong Y, Posada-Quintero HF, Daley MS, Chon KH, Bolkhovsky J. Facial features and head movements obtained with a webcam correlate with performance deterioration during prolonged wakefulness. Atten Percept Psychophys 2021;83:525–40.

[19] Lopez N, Previc FH, Fischer J, Heitz RP, Engle RW. Effects of sleep deprivation on cognitive performance by United States Air Force pilots. J. Appl. Res. Mem. Cogn. 2012;1:27–33. https://doi.org/10.1016/j.jarmac.2011.10.002.

[20] Othman N, Romli FI. Mental workload evaluation of pilots using pupil dilation. Int. Rev. Aerosp. Eng. 2016;9:80–4.

[21] Posada-Quintero HF, Bolkhovsky Jeffrey B, Qin Michael, Chon Ki H. Human performance deterioration due to prolonged wakefulness can Be accurately detected using time-varying spectral analysis of electrodermal activity. Hum Factors 2018;60:1035–47.

[22] Posada-Quintero HF, Reljin N, Bolkhovsky J, Orjuela-Cañón AD, Chon K. Brain activity correlates with cognitive performance deterioration during sleep deprivation. Front Neurosci 2019;13:1001.

[23] Rathakrishnan E, Al-Garni AZ, Nebylov A, Sinha AK, Bhattacharyya D, Drikakis D, Nitzsche F, Fricke H, Gursul I, Rohacs J. International review of Aerospace Engineering (IREASE). 2010.

[24] Rizzo L, Dondio P, Delany SJ, Longo L. Modeling mental workload via rule-based expert system: a comparison with NASA-TLX and workload profile. In: IFIP international conference on artificial intelligence applications and innovations. Springer; 2016. p. 215–29.

[25] Saksvik-Lehouillier I, Saksvik SB, Dahlberg J, Tanum TK, Ringen H, Karlsen HR, Smedbøl T, Sørengaard TA, Stople M, Kallestad H. Mild to moderate partial sleep deprivation is associated with increased impulsivity and decreased positive affect in young adults. Sleep 2020;43:zsaa078.

[26] Santiago-Espada Y. The multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload research: A User's Guide. 2011.

[27] Valk PJ, Simons RM, Struyvenberg PA, Kruit H, van Berge Henegouwen MT. Effects of a single dose of loratadine on flying ability under conditions of simulated cabin pressure. Am J Rhinol 1997;11:27–36.

[28] Valk PJL, Simons M. Effects of loratadine/montelukast on vigilance and alertness task Performance in a simulated cabin environment. Adv Ther 2008;26:89. https://doi.org/10.1007/s12325-008-0127-6.

[29] Wilson GF, Caldwell JA, Russell CA. Performance and psychophysiological measures of fatigue effects on aviation related tasks of varying difficulty. Int J Aviat Psychol 2007;17:219–47. https://doi.org/10.1080/10508410701328839.