Don't Think Twice, It's All Right? – An Examination of Commonly Used EEG Indices and Their Sensitivity to Mental Workload

Anneke Hamann(✉) and Nils Carstengerdes

Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Flugführung, Lilienthalplatz 7, 38108 Braunschweig, Germany
anneke.hamann@dlr.de

Abstract. Physiological monitoring of the operator's current state has gained much attention in aviation research, especially for the development of adaptive assistance systems. In order to tailor the assistance to the human operator's current needs, these systems need to be informed about their operator's state. Physiological data can provide such information objectively, continuously and almost in real-time. Using electroencephalography (EEG) band power analyses, changing cortical activation can be detected and inferences about cognitive states drawn. In addition, the combination of band powers into indices is sometimes used to enhance sensitivity. In the current work, we compared the sensitivity of two indices commonly used for mental workload (MWL) assessment, the Task Load Index (TLI) and the Engagement Index (EI), against each other and with single band powers. We computed the TLI and EI from the datasets of two flight simulator studies that induced MWL while controlling for mental fatigue (MF) ($N = 35$) and vice versa ($N = 31$). We hypothesized that both TLI and EI would increase with MWL, but would not vary with MF. Additionally, according to the literature, TLI and EI should be more sensitive to changes in MWL than single bands. The TLI increased with increasing MWL, but proved less sensitive than theta band power alone. It did not vary with increasing MF. The EI did not vary with MWL, but decreased slightly with MF. We conclude that the usefulness and sensitivity of EEG indices is not universal, but varies considerably across studies and most likely experimental tasks. Therefore, the choice of an EEG feature should be made carefully. Especially in automated systems developed to monitor the operator's state, EEG features should not be used blindly as a seemingly valid data source, but always empirically validated with respect to their sensitivity.

Keywords: EEG · mental workload · physiological monitoring

1 Introduction

In aviation, there has always been a trend towards higher levels of automation. Right from the beginning of powered flight, there have been approaches to automate functions in order to decrease the demand that is put on the pilot. From simple stabilization

D. Harris and W.-C. Li (Eds.): HCII 2023, LNAI 14017, pp. 65–78, 2023.
https://doi.org/10.1007/978-3-031-35392-5_5

mechanisms that controlled roll and pitch of the aircraft to automated approaches and landings using an Instrument Landing System (ILS) to fly-by-wire systems that keep the aircraft within its safe operation envelope, this trend has continued all throughout the 20th and 21st century [1, 2]. Today, research and industry work on even more advanced systems that can adapt flexibly to the situation or the needs of the pilot. Such adaptive or even artificial-intelligence based systems could one day replace the co-pilot and enable so-called single pilot operations [2].

Even with the help of automation, flying an aircraft is still a complex, cognitively demanding task. Cognitive resources, however, are limited and pose a boundary to the pilot's capabilities [3]. The extent to which the cognitive resources are occupied by the task is defined as mental workload (MWL) [4]. An easy task that does not require many resources elicits only low MWL, whereas a difficult task elicits higher levels of MWL. Both extremes should be avoided to ensure optimal performance [5]: Too high MWL will drain the limited cognitive resources and lead to performance decline and errors. Too low MWL, on the other hand, will lead to boredom and distraction, and will negatively impact performance just as well. Besides MWL, other factors such as mental fatigue (MF) can have an impact on the pilot's performance. MF results from long periods of task execution, and is characterized by reduced alertness and the unwillingness to expend further effort [6]. If not counteracted in time, MF can transition into sleepiness.

An adaptive or intelligent assistance system could help to keep the pilot's MWL at an optimal level, and could intervene before MF increases to an unwanted extent. The assistance system could for example relieve the pilot of some tasks if they are overloaded, or ask them to take a break and take over completely. It could also give tasks back to manual control if the pilot is underloaded and in danger of getting too distracted. However, there is one problem that still needs to be solved. In order to accurately tailor the assistance to the pilot's current needs, the system needs to be informed about their state. Ideally, this information is valid, objective and available in (near-) real-time. Physiological measurement can provide such data – provided that the measures used are indeed valid indications of the pilot's state.

Especially for cognitive factors such as MWL, electroencephalography (EEG) has been the physiological assessment method of choice [7, 8]. EEG records the electrical activity of the brain via electrodes placed on the scalp. From these raw data, information about the frequencies present in the signal can be extracted. Different frequency bands have been defined and associated with different brain functions and cortical activity (from low to high: delta, theta, alpha, beta, and gamma frequencies). By performing spectral analyses on the gathered data and comparing the composition of the EEG signal between tasks or to a baseline measurement, conclusions can be drawn to the changing cortical activation and cognitive states. For example, increasing MWL usually results in increasing activity in the theta band at frontal cortical areas [9–11], and decreasing activity in the alpha and beta bands at parietal areas [12, 13]. Unfortunately, brain activity is complex and the frequencies captured by the EEG are mere approximations of the underlying processes. Moreover, different frequency bands are associated with more than one function. They thus vary with different influencing factors such as MWL or MF, and interactions between these factors are seldom controlled for [11]. As a result, the

sensitivity of the EEG features varies across studies and there is still a lack of consensus on the best EEG features to validly measure each cognitive state.

In addition to the investigation of single band powers, the combination of multiple frequency bands into indices (e.g. by adding or dividing powers of multiple frequency bands and electrodes) is used to enhance the sensitivity of the single band powers. This way, indices are built to emphasize certain trends in the single bands. Although guided by theoretical considerations about the relevant frequency bands and electrode positions, the exact way the EEG features are combined into an index depends entirely on the researchers' choices of frequency bands, electrode positions and formulae to achieve certain value ranges. This leads to a wide variety of different indices [14–19] whose relationships and validity are seldom compared.

In the following, we describe two indices that are widely used to assess MWL and engagement during manual and automated tasks and that therefore, in theory, present two viable candidates for the assessment of pilots' cognitive states: The Task Load Index (TLI) and the Engagement Index (EI).

1.1 Task Load Index (TLI) and Engagement Index (EI)

The TLI is defined as *theta Fz/alpha Pz*. It was originally developed by Smith et al. (2001) to assess changes in EEG with changing task load [17]. In their study, the authors found the most prominent activity in the theta band at 6–7 Hz at electrode Fz, and in the alpha band at 8–12 Hz, further divided into "slow" (8–10 Hz) and "high" (10–12 Hz), at electrode Pz. Thus, the TLI is often computed from these bands and electrodes. It is noteworthy, however, that in the original publication, the authors did not assign a fixed index to all participants. They tested different combinations of alpha and theta power in varying bandwidths and at varying electrode sites, and chose the best combination for each participant. This way they accounted for interindividual variance, and recommended using participant-specific indices instead of a "one size fits all" index [17]. In following studies this approach has largely been abandoned, and the same index is used for all participants. While most researchers now use frontal theta power at Fz and parietal alpha power at Pz to compute the TLI, the definition of the theta and alpha band vary considerably across studies, see Table 1 for an overview of selected studies.

The TLI usually increases with increasing task demands and can therefore be used to assess MWL [12, 19–23]. In some studies, it was more sensitive to task demands than single bands [12, 21, 22], while others find no effect of task demands on the TLI [24]. Finally, in a series of studies using a simulated air traffic control task, the TLI decreased instead of increased with increasing demand [25–27]. Therefore, it is possible that the TLI is task-specific and responds differently to certain aspects of a task.

The EI is defined as *beta/(alpha + theta)*, and each band is averaged over four electrode sites, Cz, Pz, P3 and P4. The EI was originally developed by Pope et al. (1995) to assess an operator's engagement (i.e. alertness/attention) during manual and automated tasks [16]. The idea is that when the index is high, the human operator is engaged in the task and able to monitor it. Thus, more tasks can be automated without the risk that the operator is inattentive and out of the loop. If the index decreases and thus engagement wanes, tasks can be shifted from automated to manual control so that the human operator needs to engage again [16]. The authors initially computed various

Table 1. Overview of studies using the TLI and EI, with respective definitions of the theta, alpha and beta bands. Studies using the inverse EI are marked *.

Studies using TLI and/or EI	TLI (theta Fz/alpha Pz)		EI (beta/(alpha + theta)) over Cz, Pz, P3, P4		
	theta (Hz)	alpha (Hz)	theta (Hz)	alpha (Hz)	beta (Hz)
*D'Anna et al., 2016 [28]	–	–	4–8	8–13	13–30
Figalová et al., 2022 [24]	4–8	8–13	–	–	–
Freeman et al., 1999 [29]	–	–	4–8	8–13	13–22
Hockey et al., 2009 [20]	6–7	8–12	4–8	8–13	13–22
Holm et al., 2009 [12]	4–8	8–12	–	–	–
Jaquess et al., 2018 [21]	4–7	8–13	–	–	–
Kamzanova et al., 2011 [26]	4–8	8–14	4–8	8–14	14–30
Kamzanova et al., 2012 [25]	4–8	8–14	4–8	8–14	14–30
Kamzanova et al., 2014 [27]	4–8	8–14	4–8	8–14	14–30
Matthews et al., 2015 [22]	4–8	9–13	–	–	–
Nickel et al., 2006 [19]	6–7	10–12	4–8	8–13	13–22
Nickel et al., 2007 [30]	6–7	10–12	4–8	8–13	13–22
Nuamah et a., 2020 [23]	4–8	8–12	–	–	–

indices and found the EI to be most sensitive to engagement. However, it was only built upon the data of six participants which may limit the generalizability of the findings. Like the TLI, the EI is subject to varying definitions of the theta, alpha and beta bands used to compute it. In the original version, the bands were set as follows: theta (4–8 Hz), alpha (8–13 Hz) and beta (13–22 Hz). An overview of selected studies and the used bandwidths can be found in Table 1. In addition to the EI, there is an inverse EI that is computed as *(alpha + theta)/beta.* It was proposed by Brookhuis and de Waard (1993) to assess driving performance [31], even before Pope et al.'s EI. It originally consisted only of the electrodes Pz and Oz. There is a recent study [28] that used the inverse EI, but computed over the four electrodes proposed by Pope et al. Therefore, it is only a mathematical inverse of the EI and thus comparable. Hence, we decided to include this study in our work. It is important to keep in mind, however, that the inverse EI decreases with increasing attention, while the EI increases.

While not specifically designed to assess MWL, the EI is often used as an approximation of task demand. It increases in tasks which require more attention, such as manual as compared to automated tasks [28, 29]. However, other studies could not find effects of task demand on the EI [19, 20]. It has been argued that the TLI and EI represent two different aspects of task demand: While the TLI is associated with more or less effort to perform the task, the EI indicates increasing or waning attention and thus fatigue [30]. However, this would imply that the EI is also affected by time on task, which could not be confirmed [25–27].

While TLI and EI are often used together, in the same studies and therefore on the same tasks, no systematic investigation has been made yet of the sensitivity and specificity of the indices to MWL. If the indices indeed are good indications of MWL and attention in demanding tasks, they should only vary with increasing MWL (task difficulty), and not with increasing MF (time on task). If, on the other hand, they indeed represent two different facets of task demand [30], then the TLI would only respond to changing MWL, and the EI only to increasing MF. With this study, we address this problem with an empirical investigation.

1.2 The Current Study

We used data from two previously conducted experiments in which we induced MWL while controlling MF, and vice versa. In both experiments we used the same task, a simulated flight task. In the experiment on MWL, we used four levels of task difficulty, and used randomization and a task duration of max. 45 min to prevent confounding effects of MF. In the experiment on MF, we used only one difficulty level of the simulated flight task to keep MWL constant, but prolonged the task to 90 min to induce MF. This way, we not only controlled for unwanted influences of other cognitive factors, but also for effects of task characteristics.

We computed the TLI and EI for both datasets and compared their behavior with increasing MWL and MF. In order to build the indices, we used the definitions of frequency bands and electrode positions provided in the original publications. We hypothesized that both TLI and EI would increase with increasing MWL, but would not vary with MF. Following the literature on these indices, we also hypothesized that TLI and EI would be more sensitive to changes in MWL than single bands, i.e. differentiate more MWL levels than single band powers could.

2 Method

In this paper, we present a re-analysis of already gathered EEG data from two studies. Here, we give only a brief overview of the experimental tasks and procedures as an explanation of how the data were obtained. Please note that in our previous studies, more data including self-report, performance and functional near-infrared spectroscopy (fNIRS) data have been gathered as well. We performed analyses to ensure that our manipulation of MWL and MF had worked. Results of these analyses as well as further information on the studies can be found in the respective publications [11, 32].

2.1 Sample

The MWL dataset contains data from 35 participants (24 male, 11 female) between 19 and 30 years ($M = 23.7, SD = 2.1$). The MF dataset contains data from 31 participants (20 male, 11 female) between 19 and 33 years ($M = 24.1$, $SD = 3.4$). All participants were native German-speakers, currently enrolled at a university, right-handed, had normal hearing and normal or corrected-to-normal vision, no pre-existing neurological conditions, no flying experience and held no pilot's or radio telephony licence. All provided

written, informed consent and received monetary compensation for participation. Both studies were approved by the ethics commission of the German Psychological Society (DGPs) and conducted in accordance with the declaration of Helsinki.

2.2 Experimental Tasks and Material

Both studies were conducted in an A321 cockpit simulator at the Institute of Flight Guidance, German Aerospace Center (DLR), Braunschweig. The experimental task for both experiments was a simulated flight task that was simplified in a way that it could be learned by novices. Most of the flight parameters were controlled by the autopilot. The participants only had to monitor the altitude of the aircraft and react to deviations (monitoring task), and to change the heading of the aircraft (adapted n-back task). Both tasks were performed in parallel.

In the monitoring task, the altitude of the aircraft was initially set to 20,000 ft. The experimenter could trigger deviations from this altitude, which would lead to slow altitude increases or decreases. The participants were instructed to react if a deviation of more than 40 ft was reached, and to correct the altitude back to the 20,000 ft by setting a vertical speed. In the adapted n-back task, the participants had to memorize and reproduce headings, i.e. courses in degree (for example, an eastward course equals 90°, i.e. a heading of 090). They heard a series of auditory heading commands which they had to follow in line with an instruction adapted from the classical n-back paradigm. In a 0-back condition, the command had to be put in at once and no memorization was necessary. In a 1-, 2-, or 3-back condition, the participants had to memorize one, two or three headings at a time and thus put in the heading one, two or three prior to the one just heard. Thus, the task had four difficulty levels. More information on the flight task can be found in the original publication [11].

2.3 Procedure

Both the MWL and MF study had a within-subject design. The experiments were divided into task blocks (approx. 3–3.5 min) and breaks (2 min). In each block, the monitoring and n-back task were performed in parallel. Per block, the n-back level was kept constant (i.e. instructions did not vary within one block) and one altitude deviation was triggered by the experimenter. In the MWL study, the participants performed each n-back difficulty level (0-, 1-, 2-, 3-back) twice, resulting in eight blocks, and the whole experiment lasted approx. 45 min. In the MF study, only the 1-back level of the task was used, but the experiment was prolonged to 16 blocks, approx. 90 min.

2.4 EEG Data Recording and Pre-processing

EEG data were recorded at 500 Hz with a LiveAmp-32 in BrainVision Recorder 1.24 (Brain Products GmbH, Gilching, Germany). 28 Ag/AgCl active electrodes were positioned according to the 10–20 system with online reference at FCz, see Fig. 1.

Pre-processing was done in BrainVision Analyzer 2.2 (Brain Products GmbH, Gilching, Germany). The data were down-sampled to 256 Hz, re-referenced to average and

bandpass-filtered between 0.5–40 Hz using a 4th order IIR filter with an additional 50 Hz notch filter to remove remaining line noise from the unshielded simulator. Motion artefacts were removed semi-automatically and an independent component analysis was performed for ocular correction. Then, the data were divided into blocks, beginning with the first reaction per block. In the MWL dataset, this resulted in eight blocks (two presentations of 0-back, 1-back, 2-back, 3-back). In the MF dataset, 16 blocks (16 x 1-back, over time) were built. The blocks were segmented into epochs of 2 s with 0.5 s overlap. Power Spectral Density was computed using Fast Fourier Transformation with a Hanning window with 10% overlap. The average per block was exported as raw sum (μV^2/Hz) for the bands and electrodes of interest in both datasets. Following the original publications, for the TLI, theta power (6–7 Hz) at electrode Fz and alpha power (8–12 Hz) at electrode Pz were exported [17]. For the EI, theta power (4–8 Hz), alpha power (8–13 Hz) and beta power (13–22 Hz) were exported from the four electrodes Cz, Pz, P3 and P4 [16].

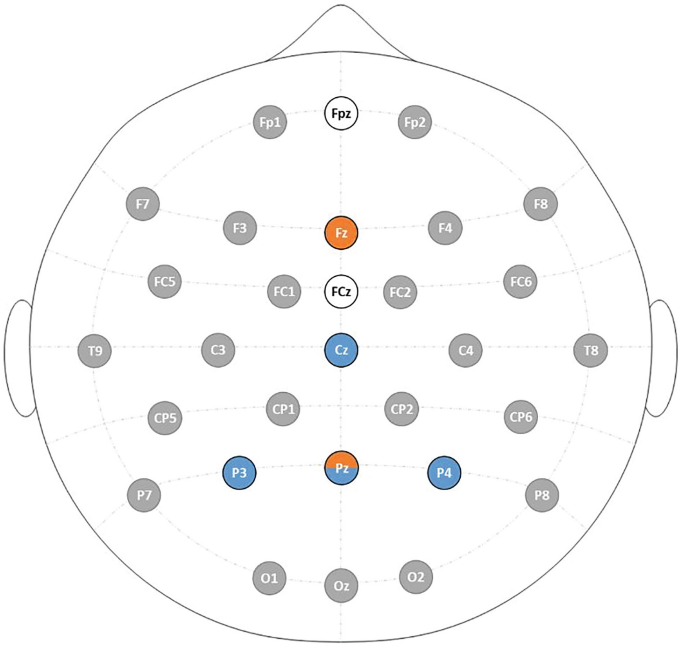

Fig. 1. EEG montage with electrode positions according to the 10–20 system. White = ground and reference electrodes; orange = electrodes used to compute the TLI; blue = electrodes used to compute the EI; grey = all other electrodes. Electrode Pz is part of both TLI and EI.

2.5 Data Analysis

All statistical analyses were conducted in SPSS 26 (IBM Corp., Armonk, NY, USA). Due to skewness, the EEG data were ln-transformed. In the MWL dataset, the two

presentations of each n-back level were averaged. In both datasets, missing values were replaced by the mean of the respective variable if necessary. The TLI was computed as *theta Fz/alpha Pz*. The EI was built by first averaging over the four electrodes Cz, Pz, P3 and P4 for each band, then computing *beta/(alpha + theta)*.

For the two indices TLI and EI, separate analyses of variance (ANOVAs) were conducted in each dataset. The sphericity assumption was tested and, in case of violation, corrected using the Greenhouse-Geisser correction. In the MWL dataset, a one-way (4 n-back levels) repeated-measures ANOVA with Bonferroni-corrected post-hoc pairwise comparisons (two-tailed) was computed for each EEG index, i.e. six comparisons. In the MF dataset, a one-way (time on task as 16 blocks) repeated-measures ANOVA was computed for each EEG index. Due to the large number of possible pairwise comparisons, a significant result was followed up with Bonferroni-Holm corrected paired two-tailed t-tests only for the 1st, 4th, 8th, 12th and 16th block (start, ¼, ½, ¾, end of the experiment), i.e. ten comparisons.

Finally, for the comparison of the indices to single bands, we compared the results to those of our previous MWL study [11]. We compared how many MWL levels could be differentiated with which measure, and how large the effect sizes of the respective ANOVAs were. In order to foster comparability with our previous study, in which we had defined the frequency bands slightly differently, we also computed an alternative TLI and alternative EI with the frequency bands from the previous publication: theta (4–8 Hz), alpha (8–12 Hz) and beta (12–30 Hz).

3 Results

3.1 Task Load Index (TLI)

The ANOVA for the MWL dataset showed a significant increase with increasing n-back level, $F(2.26, 76.86) = 15.91$, $p < .001$, $\eta^2_p = .32$, see Fig. 2. Four of the six pairwise comparisons showed a significant difference: 0- vs. 2-back ($p = .004$), 0- vs. 3-back ($p < .001$), 1- vs. 2-back ($p = .033$), 1- vs. 3-back ($p < .001$). The two lowest difficulty levels (0- vs. 1-back) and the two highest difficulty levels (2- vs. 3-back) did not show significant differences, both ps $> .05$.

The ANOVA for the MF dataset showed a significant increase over time, $F(3.52, 105.53) = 2.63$, $p = .045$, $\eta^2_p = .08$, see Fig. 3. The subsequent t-tests did not show any significant difference between any of the tested blocks, all ps $> .05$.

3.2 Engagement Index (EI)

The ANOVA for the MWL dataset did not become significant, $F(3, 103) = 0.51$, $p > .05$, $\eta^2_p = .02$, see Fig. 2. No further tests were carried out.

The ANOVA for the MF dataset showed a significant decrease over time, $F(7.65, 229.52) = 7.05$, $p < .001$, $\eta^2_p = .19$, see Fig. 3. The subsequent t-tests showed significant differences between blocks only for four of the ten comparisons: There were only significant differences between block 1 and the subsequent blocks 4, 8, 12 and 16, all ps $< .001$. No significant differences between any of the later blocks were found.

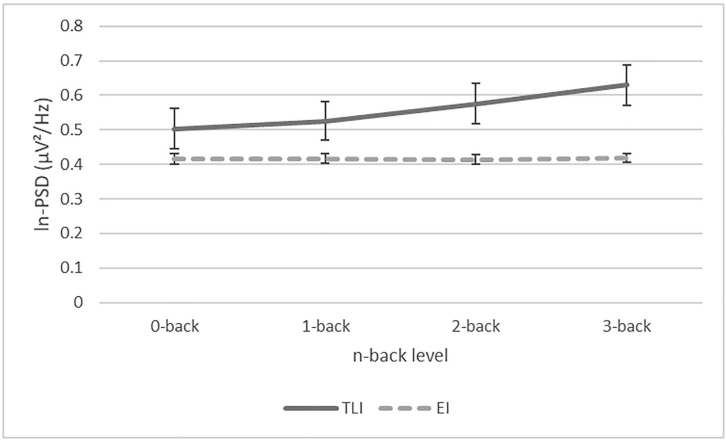

Fig. 2. EEG indices across n-back levels in the MWL dataset. Mean values are shown. Error bars indicate *SE*. EEG data are represented as Power Spectral Density (ln-transformed).

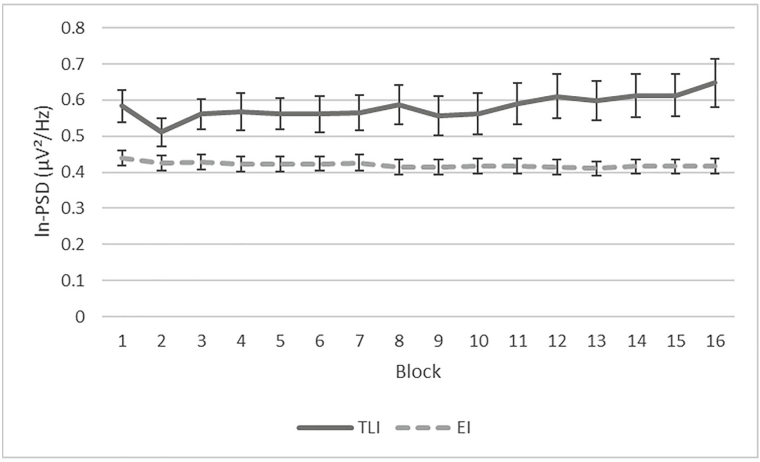

Fig. 3. EEG indices across blocks in the MF dataset. Mean values are shown. Error bars indicate *SE*. EEG data are represented as Power Spectral Density (ln-transformed).

3.3 Comparison to Single EEG Bands

We compared the sensitivity to MWL changes between the indices and single EEG bands. Therefore, we used the results of our previous MWL study on the single band powers for theta Fz, alpha Pz and beta Pz as well as the results from the TLI and EI (computed with the original bandwidths) and alternative TLI and EI (computed with the bandwidths from our previous study). This way, we could analyze if differences were due to the indices per se or variations in the definition of the theta, alpha and beta bands.

For the alternative TLI, the ANOVA showed a significant increase with increasing task difficulty, $F(2.03, 69.04) = 7.57$, $p = .001$, $\eta^2_p = .18$. Two of the six pairwise

comparisons showed a significant difference: 0- vs. 3-back ($p = .015$) and 1- vs. 3-back ($p = .012$). No other comparison was significant, all ps $> .05$. For the alternative EI, the ANOVA was not significant, $F(2.05, 69.71) = 0.21$, $p > .05$, $\eta^2_p = .06$.

In our previous research [11], theta Fz was most sensitive to changing MWL. It increased with increasing n-back level, showing a large effect, $F(1.56, 52.87) = 23.91$, $p < .001$, $\eta^2_p = .41$, and differentiated all but the two lowest difficulty levels. We did not find significant changes in alpha Pz and beta Pz. In Table 2 we have contrasted the results of the previous publication and the current analysis.

Table 2. Overview of different EEG features and their ability to discriminate the four n-back levels, including effect sizes of the ANOVAs with significant outcome. Significant differences ($p < .05$, Bonferroni-corrected) marked with ✔.

Pairwise comparisons of different n-back levels		Single band powers (previous study [11])			Indices (current study)			
		theta Fz	alpha Pz	beta Pz	TLI	alternative TLI	EI	alternative EI
0 vs.	1	–	–	–	–	–	–	–
	2	✔	–	–	✔	–	–	–
	3	✔	–	–	✔	✔	–	–
1 vs.	2	✔	–	–	✔	–	–	–
	3	✔	–	–	✔	✔	–	–
2 vs.	3	✔	–	–	–	–	–	–
Effect size η^2_p		.41	–	–	.32	.18	–	–

4 Discussion

In this paper, we tested the sensitivity and specificity of two well-known EEG indices, TLI and EI, to changes in MWL. We analyzed their behavior with increasing MWL and increasing MF, and compared them to single alpha, beta and theta band powers.

The TLI increased substantially with increasing MWL. Using it, we could differentiate all induced MWL levels apart from the two lowest (0- vs. 1-back) and the two highest (2- vs. 3-back). The TLI also increased slightly yet substantially with MF, even if the increase was so weak that no discrimination between blocks was possible. We therefore conclude that the TLI is sensitive to changing MWL, but lacks specificity as it also varies with MF. Moreover, the direction of the variation was the same, i.e. the TLI increased in both datasets. It is therefore not possible to conclude without doubt if the person experiences increasing MWL or MF if only the TLI is used.

Contrary to our expectations, the EI did not vary with MWL and could not be used to discriminate any MWL levels. It did show a slight decrease in the MF dataset, but

only between the first and all later blocks. It is unlikely that this is due to MF because of the early onset of the change and the lacking gradual decrease over the time course of the experiment. The observed decrease could be an indication of a learning effect during the first block. With no variation with MWL, we consider the EI not to be sensitive to MWL changes, yet also not sensitive to gradually increasing MF.

When comparing the indices with the single band powers from our previous study regarding sensitivity to MWL, we found that the TLI was more sensitive to MWL changes than only parietal alpha or beta power (which did not vary with MWL at all), but less sensitive than frontal theta band power. This can be seen from both explained variance in the ANOVA and the number of n-back levels it could differentiate. Only single theta band power at Fz was able to differentiate the two highest difficulty levels (2- vs. 3-back). This difference in sensitivity cannot be attributed to varying definitions of the theta band, as can be seen from the alternative TLI: When computed with the same bandwidths as the single band powers, it could only differentiate the highest difficulty level from the two lowest. It was thus less sensitive to MWL changes than both the original TLI and single theta band power. The EI and alternative EI could not differentiate any changes in MWL. In sum, single theta band power at Fz proved most sensitive to changing MWL, followed by the original TLI.

Taken together, when investigated systematically, neither TLI nor EI can be considered ideal measures of MWL. The TLI was sensitive to MWL changes, which aligns with the literature [12, 19–23], but less so than single frontal theta band power. Moreover, the TLI was not specific to MWL as it also varied with MF. The direction of the change, however, was the same: It increased with MWL and with MF. It can therefore be seen as a measure of increasing cognitive demand, regardless of its source. If one wants to pinpoint exactly what causes this increasing demand, the TLI might not be the measure of choice. The EI, to our surprise, was neither sensitive to MWL nor to MF, even though it was designed to capture fluctuations in attention and alertness. We are not the first to conclude this [19, 20, 27], but the first to test it in comparable and controlled experiments on MWL and MF. It is therefore questionable if the EI is a useful measure for any kind of assessment of cognitive demand.

Contrary to earlier work [30], it seems that task demands cannot so easily be split in effort and fatigue, or at least that TLI and EI are not mutually exclusive indices of these facets. When looking at the way the indices are built, however, this is not surprising. Both incorporate parietal alpha and beta band power which vary with a multitude of cognitive factors. They have been shown to decrease with increasing MWL [12, 13] and increase with MF [33–35]. Furthermore, parietal alpha power increases with frequent task switching [10]. Frontal theta band power, which is a part of the TLI, is not exclusive to MWL either. It increases with task demands, both in terms of task difficulty (MWL) [9, 10] and time on task (MF) [33, 36]. If single band powers are subject to variations with different cognitive factors and task characteristics, so are indices that combine them. And the more complex the relationships between bands and influencing factors, the more difficult is the interpretation of the index.

In conclusion, the TLI and EI are two of the most widely known and used EEG indices for assessment of MWL and task engagement. And yet, upon further investigation, their sensitivity and specificity to MWL are not as high as commonly thought, with large

variations across studies and experimental tasks. We would therefore like to emphasize the importance of choosing an EEG feature (be it an index or single band) carefully. Ideally, it should be validated for the specific task, application and if possible even tailored to single participants or operators. Especially if the EEG data are meant to be used as a data source for an adaptive or intelligent system that tailors its assistance to the operator's current needs, EEG features should not be trusted blindly. They should always be validated in their sensitivity – and specificity – to indicate the operator's state.

References

1. Billings, C.E.: Toward a human-centered aircraft automation philosophy. Int. J. Aviat. Psychol. **1**, 261–270 (1991). https://doi.org/10.1207/s15327108ijap0104_1
2. Chartered Institute of Ergonomics & Human Factors: The human dimension in tomorrow's aviation system. White Paper (2020)
3. Endsley, M.R.: Situation awareness in aviation systems. In: Garland, D.J. (ed.) Handbook of Aviation Human Factors. Human Factors in Transportation. Erlbaum, Mahwah (1999)
4. O'Donnell, R.D., Eggemeier, F.T.: Workload assessment methodology. In: Boff, K.R., Kaufman, L., Thomas, J.P. (eds.) Handbook of Perception and Human Performance. John Wiley & Sons, New York (1986)
5. Martins, A.P.G.: A review of important cognitive concepts in aviation. Aviation **20**, 65–84 (2016). https://doi.org/10.3846/16487788.2016.1196559
6. Grandjean, E.: Fatigue in industry. Brit. J. Ind. Med. (1979).https://doi.org/10.1136/oem.36.3.175
7. Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F.: Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neurosci. Biobehav. Rev. **44**, 58–75 (2014). https://doi.org/10.1016/j.neubiorev.2012.10.003
8. Charles, R.L., Nixon, J.: Measuring mental workload using physiological measures: a systematic review. Appl. Ergon. **74**, 221–232 (2019). https://doi.org/10.1016/j.apergo.2018.08.028
9. Dussault, C., Jouanin, J.-C., Guezennec, C.-Y.: EEG and ECG changes during selected flight sequences. Aviat. Space Environ. Med. **75**, 889–897 (2004)
10. Puma, S., Matton, N., Paubel, P.-V., Raufaste, É., El-Yagoubi, R.: Using theta and alpha band power to assess cognitive workload in multitasking environments. Int. J. Psychophysiol. Off. J. Int. Organ. Psychophysiol. **123**, 111–120 (2018). https://doi.org/10.1016/j.ijpsycho.2017.10.004
11. Hamann, A., Carstengerdes, N.: Investigating mental workload-induced changes in cortical oxygenation and frontal theta activity during simulated flights. Sci. Rep. **12**, 6449 (2022). https://doi.org/10.1038/s41598-022-10044-y
12. Holm, A., Lukander, K., Korpela, J., Sallinen, M., Müller, K.M.I.: Estimating brain load from the EEG. Sci. World J. **9**, 639–651 (2009). https://doi.org/10.1100/tsw.2009.83
13. Dehais, F., Duprès, A., Blum, S., Drougard, N., Scannella, S., Roy, R.N., Lotte, F.: Monitoring pilot's mental workload using ERPs and spectral power with a six-dry-electrode EEG system in real flight conditions. Sensors (Basel, Switzerland) (2019). https://doi.org/10.3390/s19061324
14. Choi, M.K., Lee, S.M., Ha, J.S., Seong, P.H.: Development of an EEG-based workload measurement method in nuclear power plants. Ann. Nucl. Energy **111**, 595–607 (2018). https://doi.org/10.1016/j.anucene.2017.08.032

15. Freeman, F.G., Mikulka, P.J., Scerbo, M.W., Scott, L.: An evaluation of an adaptive automation system using a cognitive vigilance task. Biol. Psychol. **67**, 283–297 (2004). https://doi.org/10.1016/j.biopsycho.2004.01.002
16. Pope, A.T., Bogart, E.H., Bartolome, D.S.: Biocybernetic system evaluates indices of operator engagement in automated task. Biol. Psychol. **40**, 187–195 (1995). https://doi.org/10.1016/0301-0511(95)05116-3
17. Smith, M.E., Gevins, A., Brown, H., Karnik, A., Du, R.: Monitoring task loading with multivariate EEG measures during complex forms of human-computer interaction. Hum. Fact. **43**, 366–380 (2001). https://doi.org/10.1518/001872001775898287
18. McMahan, T., Parberry, I., Parsons, T.D.: Evaluating electroencephalography engagement indices during video game play. In: Proceedings of the 10th International Conference on the Foundations of Digital Games (FDG 2015). Foundations of Digital Games 2015, Pacific Grove, CA, USA, 22–25 June 2015 (2015)
19. Nickel, P., Hockey, G.R.J., Roberts, A.C., Roberts, M.H.: Markers of high risk operator functional state in adaptive control of process automation. In: Proceedings of IEA 2006, pp. 304–312 (2006)
20. Hockey, G.R.J., Nickel, P., Roberts, A.C., Roberts, M.H.: Sensitivity of candidate markers of psychophysiological strain to cyclical changes in manual control load during simulated process control. Appl. Ergon. **40**, 1011–1018 (2009). https://doi.org/10.1016/j.apergo.2009.04.008
21. Jaquess, K.J., et al.: Changes in mental workload and motor performance throughout multiple practice sessions under various levels of task difficulty. Neuroscience **393**, 305–318 (2018). https://doi.org/10.1016/j.neuroscience.2018.09.019
22. Matthews, G., Reinerman-Jones, L.E., Barber, D.J., Abich, J.: The psychometrics of mental workload: multiple measures are sensitive but divergent. Hum. Fact. **57**, 125–143 (2015). https://doi.org/10.1177/0018720814539505
23. Nuamah, J.K., Seong, Y., Jiang, S., Park, E., Mountjoy, D.: Evaluating effectiveness of information visualizations using cognitive fit theory: a neuroergonomics approach. Appl. Ergon. **88**, 103173 (2020). https://doi.org/10.1016/j.apergo.2020.103173
24. Figalová, N., Chuang, L.L., Pichen, J., Baumann, M., Pollatos, O.: Ambient light conveying reliability improves drivers' takeover performance without increasing mental workload. MTI **6**, 73 (2022). https://doi.org/10.3390/mti6090073
25. Kamzanova, A., Kustubayeva, A., Matthews, G.: Diagnostic monitoring of vigilance decrement using EEG workload indices. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting (2012).https://doi.org/10.1177/1071181312561019
26. Kamzanova, A.T., Kustubayeva, A.M., Jakupov, S.M.: EEG indices to time-on-task effects and to a workload manipulation (cueing). In: World Academy of Science, Engineering and Technology (2011). https://doi.org/10.5281/zenodo.1071802
27. Kamzanova, A.T., Kustubayeva, A.M., Matthews, G.: Use of EEG workload indices for diagnostic monitoring of vigilance decrement. Hum. Fact. **56**, 1136–1149 (2014). https://doi.org/10.1177/0018720814526617
28. Georgiadis, D., et al.: A robotic cloud ecosystem for elderly care and ageing well: the growmeup approach. In: Kyriacou, E., Christofides, S., Pattichis, C.S. (eds.) XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016. IP, vol. 57, pp. 913–918. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-32703-7_178
29. Freeman, F.G., Mikulka, P.J., Prinzel, L.J., Scerbo, M.W.: Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. Biol. Psychol. **50**, 61–76 (1999). https://doi.org/10.1016/S0301-0511(99)00002-2

30. Nickel, P., Roberts, A.C., Roberts, M.H., Hockey, G.R.J.: Development of a cyclic loading method for the study of patterns of breakdown in complex performance under high load. In: de Waard, D. (ed.) Human factors issues in complex system performance. Europe Chapter of the Human Factors and Ergonomics Society, Shaker, Maastricht, pp. 325–338 (2007)
31. Brookhuis, K.A., de Waard, D.: The use of psychophysiology to assess driver status. Ergonomics **36**, 1099–1110 (1993). https://doi.org/10.1080/00140139308967981
32. Hamann, A., Carstengerdes, N.: Assessing the development of mental fatigue during simulated flights with concurrent EEG-fNIRS measurement. Sci. Rep. **13**, 4738 (2023). https://doi.org/10.1038/s41598-023-31264-w
33. Dasari, D., Crowe, C., Ling, C., Zhu, M., Ding, L.: EEG pattern analysis for physiological indicators of mental fatigue in simulated air traffic control tasks. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting (2010).https://doi.org/10.1177/154193121005400304
34. Käthner, I., Wriessnegger, S.C., Müller-Putz, G.R., Kübler, A., Halder, S.: Effects of mental workload and fatigue on the P300, alpha and theta band power during operation of an ERP (P300) brain-computer interface. Biol. Psychol. **102**, 118–129 (2014). https://doi.org/10.1016/j.biopsycho.2014.07.014
35. Nguyen, T., Ahn, S., Jang, H., Jun, S.C., Kim, J.G.: Utilization of a combined EEG/NIRS system to predict driver drowsiness. Sci. Rep. **7**, 43933 (2017). https://doi.org/10.1038/srep43933
36. Roy, R.N., Bonnet, S., Charbonnier, S., Campagne, A.: Mental fatigue and working memory load estimation: Interaction and implications for EEG-based passive BCI. In: Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2013).https://doi.org/10.1109/EMBC.2013.6611070