Evaluating EEG Measures as a Workload Assessment in an Operational Video Game Setup

Lucille Lecoutre[1], Sami Lini[1], Christophe Bey[2], Quentin Lebour[1] and Pierre-Alexandre Favier[2]
[1]*Akiani, 109 avenue Roul, 33400 Talence, France*
[2]*CNRS IMS UMR 5218, 33400 Talence, France*

Keywords: EEG, Mental Workload, Operational Situation, Video Games.

Abstract: We tested the electroencephalography (EEG) B-Alert X10 system (Advance Brain Monitoring, Inc.) mental workload metrics. When we evaluate a human-systems interfaces (HSI), we need to assess the operator's state during a task in order evaluate the systems efficiency at helping the operator. Physiological metrics are of good help when it comes to evaluate the operator's mental workload, and EEG is a promising tool. The B-Alert system includes an internal signal processing algorithm computing a mental workload index. We set up a simple experiment on a video game in order to evaluate the reliability of this index. Participants were asked to play a video game with different levels of goal (easy vs hard) as we measured subjective, behavioral and physiological indices (B-Alert mental workload index, pupillometry) of mental workload. Our results indicate that, although most of the measure point toward the same direction, the B-Alert metrics fails to give a clear indication of the mental workload state of the participants. The use of the B-Alert workload index alone is not precise enough to assess an operator mental workload condition with certainty. Further evaluations of this measure need to be done.

1 INTRODUCTION

The integration of Humans in the design and evaluation of complex systems is an approach that is becoming increasingly important. There now is a real interest in assessing the impact of such systems on operators who need to handle them. Humans have features of their own, with their constraints and limitations that it is necessary to identify in order to correctly adapt the systems.

When evaluating a system, the concept of mental workload is of particular interest to qualify the operator state. According to Cerraga and Chevalier (Cegarra and Chevalier 2008), cognitive load presupposes that cognitive processes have costs drawn from a limited pool of cognitive resources. The cognitive load is then often defined as the ratio between the demand of the task and the human resources available.

An overload situation can have tragic consequences onto the performances of an operator. Disposing of appropriate tools to assess an operator mental state becomes crucial when evaluating a system and the reliability of such tools is an important issue.

Among the physiological indices of mental workload, electroencephalography (EEG) is a good candidate for measuring and monitoring mental workload (Antonenko et al. 2010; Borghini et al. 2012; Tsang and Vidulich 2006).

EEG has some advantages for use in operational environment. In particular, wireless solutions like the B-Alert system (Advance Brain Monitoring, Inc.[1]) (Berka et al. 2004; Berka et al. 2005; Berka et al. 2007; Johnson et al. 2011) seem very promising, as they allow more ecological experimental situations. The implemented classification algorithm allows one to use it without requiring extensive medical expertise. We were particularly interested in the mental workload gauge and decided to evaluate this tool following a "black-box" approach.

Berka and colleagues argue in a previous study that the workload measure of their classifier relates more to "*cognitive processes generally considered more of the domain of executive function*", whereas their engagement measure "*tracks demands for sensory processing and attention resources*". As the

[1]http://www.advancedbrainmonitoring.com/

Lecoutre L., Lini S., Bey C., Lebour Q. and Favier P..
Evaluating EEG Measures as a Workload Assessment in an Operational Video Game Setup.
DOI: 10.5220/0005318901120117
In *Proceedings of the 2nd International Conference on Physiological Computing Systems* (PhyCS-2015), pages 112-117
ISBN: 978-989-758-085-7

task we used required some level of perceptual processing, we also evaluated the engagement metrics. Furthermore, the engagement metrics is based on an individually fitted model, whereas the workload index is based on a group of individuals tested in Berka and colleagues' article from 2007.

We set up an experiment on a video game. It allowed us to put the participants in operational conditions (they sat on an office chair, in front of TV flat screen, with PlayStation controllers in their hands and were able to move freely) making it close to a real life situation.

We used Rayman Origins (Figure 1), developed for PlayStation 3 for it is a 2D platform game. This allows us to reduce the degrees of freedom (compared to a 3D game) and ensure us the scenarios reproducibility despite the ecological environment.

Like for most platform games, the player has to collect items along the level, which defines several performance steps as a function of the number of collected items. Moreover, some scrolling levels were particularly suited to our needs. We manipulated the workload conditions by imposing two types of goal, an easy to achieve and a hard to achieve goal.

Figure 1: Screenshot of one of the two levels chosen for the experiment.

It is advised to cross several measures when estimating mental workload. Following this rational, we chose a set of subjective, behavioral and physiological measures to address the issue of the B-Alert workload index reliability.

We had the participants take the Nasa-TLX questionnaire, a standard subjective index of the mental workload (National Aeronautics and Space Administration TaskLoadindeX, Hart and Staveland 1988). This measure operates as a control measure.

Pupillometry is an indirect indicator of cognitive load (Beatty and Lucero-Wagoner 2000; Kahneman and Beatty 1966).We used this measure as a control as well, since it proved its reliability in an operational environment (Lini et al. 2013).

Heart Rate Variability (HRV) is another physiological indirect indicator of cognitive load (Backs et al. 1999; Wilson 2002). It is measured by continuously monitoring heart rate.

Based on the hypothesis that an overload has a negative effect on performances (Wickens 1992), we also collected two performance indicators: the number of collected items and the number of times the participant dies within a level. A higher number of collected items and a lower number of times a participant dies mean better performances.

We had two tests for the EEG workload index: its sensitivity to our experimental manipulation of the mental workload (easy *vs* hard condition), and its confrontation to our control measures.

We hypothesized that:

(H1) Our control measures were sensitive to our task manipulations of the mental workload.

(H2) They were correlated with each other.

(H3) The EEG index of mental workload was sensitive to our task manipulation of the mental workload.

(H4) It was correlated with our control measure.

2 METHODS

2.1 Participants

Eight healthy participants (mean age: 22.1 ± 2 years old, 7 males) took part in the study after signing a consent form. They were informed of the purpose of the study.

2.2 Measures

2.2.1 Subjective Measure

Subjects were asked to evaluate their mental workload after each of the four runs of the experiment with the Nasa-TLX questionnaire. We used an approved French version of this questionnaire (Cegarra and Morgado 2009). In order to avoid any ambiguity, the participants took a first dry questionnaire during the setup. They then took the questionnaire after each run.

2.2.2 Behavioral Measures

Each participant was asked to fill a short questionnaire about information such as his nicotine consumption, sleep deprivation, video games familiarity. We also measured during each run the number of collected items and the number of time

the participant died as a measure of their performance.

2.2.3 Physiological Measures

EEG
A portable sensor headset, the B-Alert X10 System was used to record electrophysiological data from 9 electrodes sites (Fz, F3, F4, Cz, C3, C4, POZ, P3, P4), with left and right mastoid as reference. Each EEG channel was sampled at 256 samples per seconds, with a 50-Hz notch filter applied to remove environmental artifact. We also recorded electrocardiogram (ECG) from two electrodes linked to the B-Alert system.

The signal decontamination procedure and the classification algorithms are already implemented in the B-Alert system. We were interested into two specific metrics: the probability of being in a high mental workload state, and the probability of being in a high engagement state. The mental workload model and algorithm is described in an article from Berka and colleagues (Berka et al. 2007). The model is adjusted using the group of participants tested in their study. For our analysis, we averaged the index for each run of the experiment to derive a mean workload index.1-second samples with error values were taken out of the analysis prior to averaging as advised by the constructor.

The engagement metrics is based on an individual model, adjusted for each subject with three baseline tasks. The algorithm is presented in a Johnson and colleague article from 2011 (Johnson et al. 2011). Four cognitive states are discriminated using a quadratic discriminant functional analysis: drowsiness, sleep onset, low engagement, and high engagement. For this study we were only interested in the high engagement probability metrics. We took out of the analysis the same samples as for the workload index, prior to averaging the high engagement probability for each run.

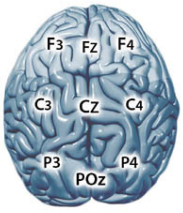

Figure 2 : EEG electrode placement.

HRV
We used the ECG channels of the B-Alert system to derive another measure of the Heart rate variability (HRV). HRV was estimated in the frequency domain by calculating the power spectral density of the Fast Fourier Transform for the tachogram. Calculations were performed in both the 0.05-0.15 Hz band (low frequency, LF), and 0.15-0.3 Hz band (high frequency, HF). Total HRV retained for analysis was the ratio of these two values (LF/HF). HRV varies in the opposite direction to that of the mental workload.

Eye Tracking
Subjects were fitted with a mobile eye tracker, Tobii's glasses (Tobii Technology [2]), a head-mounted eye tracking system resembling a pair of glasses. The tracker is monocular (right eye only), sampling at 30 Hz with 56° × 40° recording visual angle. Tobii studio, the data processing software, allows dealing with mean pupil dilation, i.e. the percentage of dilation compared to the mean dilation measured during the calibration phase.

2.3 Procedure

We had the participants take the B-Alert baseline tasks after the EEG was set-up. The choice vigilance task, and standard eyes open and eyes closed vigilance tasks lasted for 5 minutes each. The eye tracker was calibrated after the baseline acquisition.

The experimental room brightness was maintained constant, and the participant was in the room long enough before the beginning of the experiment so that we can assume he was accommodated. Moreover, mental load related variations of the pupil diameter are much faster than luminosity related variations. Thus, the measured variation of the pupil diameter clearly reflects a cognitive process and not an automatic adaptation to the change of brightness.

Participants were asked to play on two particular levels of the game, chosen for their duration and similar in difficulty. For these levels, the characters are travelling on a mosquito able to shoot at obstacles and villains. The display of the levels moves on automatically, forcing the player to move forward. These levels are scripted, events always occur at the same times independently of the actions of the player. This ensured reproducible scenarios from one condition to another and from one participant to another. The participants could train on some other levels with the same game play while they were being set-up.

Each of the two levels was played twice, either

[2] http://www.tobii.com/

with an easy-to-achieve goal, or a hard-to-achieve goal. The easy-to-achieve goal was to collect at least 150 items. The hard-to-achieve goal was to collect at least 300 items. The order of presentation of the levels and goals were assigned for each participant following a latin square procedure to avoid learning effects on the group level analysis.

In the game we chose for the experiment, the player has to start over the level from the last checkpoint when he dies. We had one of the experimenter take part in the study as player 2 to avoid these events to happen too often. The role of the experimenter was to resurrect the player 1 when he had died (due to contact with a villain or when being caught up by the automatic progress of the level). In any case the experiment helped the participant to collect items, or guided him through the level. We kept a record of the number of time the participant (player 1) died as a measure of performance. Each run lasted 4'53'' on average. The sessions were recorded in order to help synchronizing all the measures.

3 RESULTS

3.1 Sensitivity of the Measures

Error probability was set at 5%.

We first tested the sensitivity of each measure regarding the change in mental workload. We separated the data into two groups: easy and hard goal, regardless of the level of the game. One-tailed Wilcoxon's tests were used to analyse inter-individual variability between both conditions (paired samples). The side of the test was given by hypothesis.

For the Nasa-TLX score, the higher the workload is felt, the higher is the overall score. We then expected the score to be higher in the hard condition than in the easy condition.

For the performance measures, we expected that the higher the workload, the worse the performances. For the number of collected items performance measure, we thus expected a higher number of collected items in the easy condition than in the hard condition. For the number of deaths, we expected a higher number of deaths in the hard condition than in the easy condition.

For the B-Alert metrics and the pupil dilation metrics, we expected the measures to be higher in the hard condition than in the easy condition.

For the heart rate variability measure, we expected the index to be higher in the easy condition than in the hard condition.

As displayed in Figure 3, the Nasa-TLX score(**p=.025**) and the pupil dilation ratio (**p=.022**) were sensitive to our task manipulation of the mental

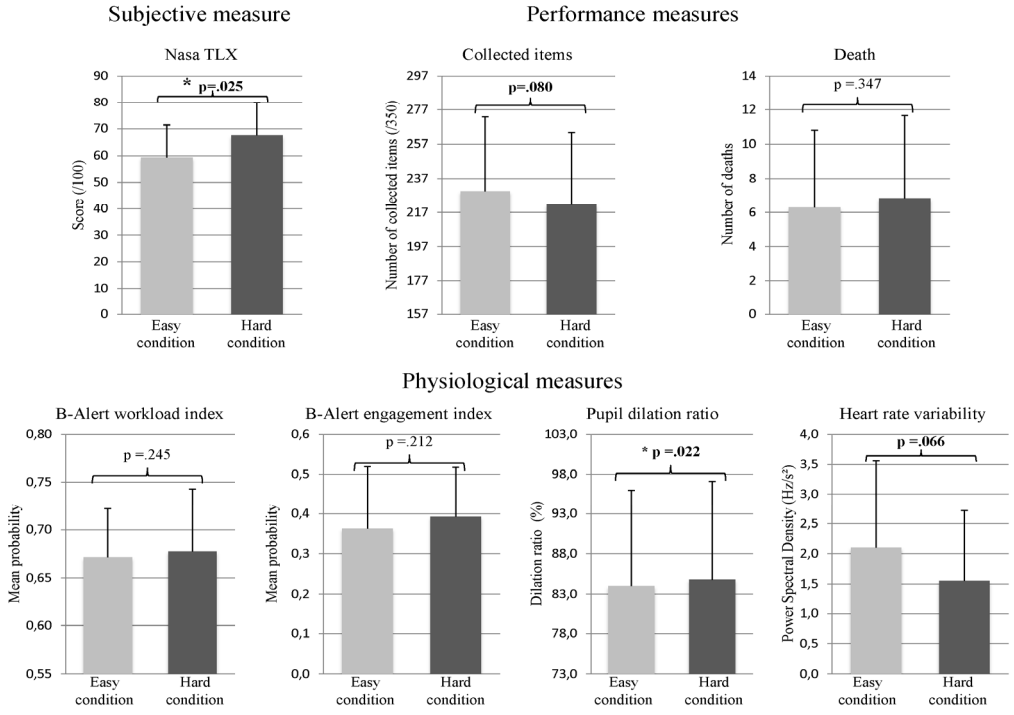

Figure 3: Across subject mean value for each measurement. The easy and hard conditions are compared. Error bars represent one standard deviation from the mean. P-value results from one-tailed Wilcoxon's tests are displayed.

workload, confirming the validity as control measures. The HRV measure (**p=.066**) and the collected item (**p=.080**) measure of performance showed a trend in the direction of our hypothesis. The other measures did not show a significant result.

3.2 Consistency of the Measures

We then analyzed the relations between the measures by computing Spearman correlations (non parametric) between the measures. Only significant values are reported.

The Nasa-TLX score is correlated with both performance measures. As hypothesized, the higher the Nasa-TLX score, the lower the number of collected items (R=-0.432, p=0.025) and the higher the number of death (R=0.461, p=0.016).

The pupil dilation measure is also correlated with the Nasa-TLX score (R=0.486, p=0.042) and both performance measure (collected items: R=-0.663, p=0.004; death: R=0.609, p=0.007), confirming its validity as a control measure. It is also strongly correlated with the EEG workload index (R=0.623, p=0.005), but not with the EEG engagement index.

The EEG workload index is also correlated with the Nasa-TLX score (R=0.430, p=0.026) and the EEG engagement index (R=0.442, p=0.019), although the EEG engagement index is correlated neither with the Nasa-TLX score nor with the pupil dilation measure. The HRV measure is not correlated with any of the other measures.

4 DISCUSSION

Our experiment was designed to address the B-Alert reliability issue in an operational environment: small number of subjects, hard to replicate experimental conditions.

We hypothesized that our control measures would reflect the task manipulation (mental workload, H1 and H2). Then, if the B-Alert system is sensitive to mental workload variations, its metric should also be affected by the task manipulation we set up (H3) and be correlated with the control measures (H4).

4.1 Validation of the Experimental Design

The Nasa-TLX subjective index and the pupil dilation measure were both significantly affected by our task manipulation of the mental workload, as previously reported in the literature (Beatty 1982; Klingner 2010). Hence, we can confirm that the experimental manipulation we used was successful at eliciting differential mental workload conditions (**H1**). The small number of subjects could explain the non-significant HRV result.

The performance measures we chose failed to show a significant variation relative to the easy and hard conditions, although the number of collected items showed a significant trend.

The pupil dilation metrics correlates with the Nasa-TLX score, showing that this index is sensitive to both objective and subjective estimates of mental workload (**H2**) as reported before (Palinko et al. 2010).

Pupil dilation and Nasa-TLX scores are both correlated with the performance measures. They are negatively correlated with the number of collected items and positively correlated with the number of deaths, confirming again that the participants were indeed in a situation of overload, which impaired their performances.

4.2 Evaluation of the EEG Indices

As revealed by the Wilcoxon's tests, the EEG workload and engagement indices were not sensitive enough to our manipulation of mental workload. The lack of results might be due an important inter-individual variability. We were able to recruit few participants only, partly because of the amount of time needed to set-up each experimental recording.

For the workload index, the model is based on a group of individuals independently of our study and is not adjusted to each participant. It then might be more sensitive to inter-individual factors (for example caffeine consumption), potentially leading to ceiling effects masking the B-Alert metric sensitivity. For the engagement index, we interpret the lack of difference between the easy and hard condition as the fact that both conditions require a similar amount of perceptual and attention related processing.

When confronting the measures, however, we can observe that the EEG workload index is correlated with the pupil dilation metrics and the Nasa-TLX score (**H4**), suggesting that it is sensitive to the same underlying phenomena.

5 CONCLUSIONS

We conducted an experiment using an ecological gaming situation to reproduce an operational

environment. Subjective and physiological measures of mental workload suggest that we succeeded in placing the subjects in an overload situation.

Our analysis indicates that the B-Alert system does not capture the variation of mental workload as can be observed with the pupil dilation and the subjective measures. However, the B-Alert workload index is correlated with both of these measures, suggesting that it varied consistently with our hypothesis, but was not sensitive enough to capture our mental workload variation.

The B-Alert system might be a useful option depending on the environmental conditions: the workload index is not sensitive to a change of brightness ($p=1$), whereas pupil dilation is ($p=.009$). Nonetheless, our results show that the B-Alert metrics are not precise enough to offer a reliable option in operational conditions.

ACKNOWLEDGEMENTS

The authors would like to thank two anonymous reviewers for their comments on a previous version of this manuscript and the subjects who kindly took part to this experiment.

The authors declare no conflict of interest.

REFERENCES

Antonenko, P. et al., 2010. Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4), pp.425-438.

Backs, R.W., Lenneman, J.K. and Sicard, J.L., 1999. The Use of Autonomic Components to Improve Cardiovascular Assessment of Mental Workload in Flight.. *The International Journal of Aviation Psychology*, 9(1), pp.33-47.

Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, 91(2), p.276.

Beatty, J. and Lucero-Wagoner, B., 2000. The pupillary system. *Handbook of psychophysiology*, 2, pp.142-162.

Berka, C. et al., 2007. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78(Supplement 1), p.B231-B244.

Berka, C. et al., 2005. Evaluation of an EEG workload model in an Aegis simulation environment. In *Defense and Security*. International Society for Optics and Photonics, pp. 90-99.

Berka, C. et al., 2004. Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction*, 17(2), pp.151-170.

Borghini, G. et al., 2012. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience and Biobehavioral Reviews*.

Cegarra, J. and Chevalier, A., 2008. The use of Tholos software for combining measures of mental workload: Toward theoretical and methodological improvements. *Behavior Research Methods*, 40(4), pp.988-1000.

Cegarra, J. and Morgado, N., 2009. Étude des propriétés de la version francophone du NASATLX. In *Communication présentée à la cinquième édition du colloque de psychologie ergonomique (Epique)*.

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

Johnson, R.R. et al., 2011. Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biological psychology*, 87(2), pp.241-250.

Kahneman, D. and Beatty, J., 1966. Pupil diameter and load on memory. *Science*, 154(3756), pp.1583-1585.

Klingner, J.M., 2010. Measuring cognitive load during visual tasks by combining pupillometry and eye tracking.

Lini, S. et al., 2013. Evaluating ASAP (Anticipation Support for Aeronauticial Planning): a user-centered case study. In *Proceedings of the 17th International Symposium on Aviation Psychology.*

Palinko, O. et al., 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications*. ACM, pp. 141-144.

Tsang, P.S. and Vidulich, M.A., 2006. Mental workload and situation awareness. *Handbook of Human Factors and Ergonomics, Third Edition*, pp.243-268.

Wickens, C.D., 1992. *Engineering psychology and human performance*, HarperCollins Publishers.

Wilson, G.F., 2002. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, 12(1), pp.3-18.