2015

Compared Evaluation of B-Alert's Encephalographic Workload Metrics Using an Operational Video Game Setup

Sami Lini

Christophe Bey

Lucille Lecoutre

Quentin Lebour

COMPARED EVALUATION OF B-ALERT'S ENCEPHALOGRAPHIC WORKLOAD METRICS USING AN OPERATIONAL VIDEO GAME SETUP

Sami LINI[1]
Christophe BEY[2]
Lucille LECOUTRE[1]
Quentin LEBOUR[1]
Pierre-Alexandre FAVIER[2]
[1] Akiani, 109 avenue Roul, 33400 Talence, France {first.last@akiani.fr}
[2] ENSC, 109 avenue Roul, 33400 Talence, France {first.last@ensc.fr}

When it comes to operational human factors studies, the use of a number of different means (psychophysiological, questionnaires, performance indexes) to complete expert behavioral observations allows specialists to issue practical recommendations despite of the variability of the few operators involved. When it comes to mental workload, literature has identified several different physiological ways to assess it. We used Heart Rate Variability (HRV) and pupillometry for previous works (ISAP'11, '13) and both have strong limitations: HRV can only be analyzed over 5-minutes time periods and pupil dilation is subject to light variability.
During this study, we tested the electroencephalography B-Alert X10 system (Advance Brain Monitoring, Inc.) mental workload metrics. We set up an experiment on a video game in real life conditions 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 indexes (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 accurate enough to assess an operator mental workload condition with certainty. Further evaluations of this measure need to be done. As we observed in a previous study, pupil dilation is a reliable index of mental workload as it correlates significantly with most measures.

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. 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.

Electroencephalography (EEG) is a good candidate for measuring and monitoring mental workload (Antonenko et al. 2010; Tsang & Vidulich 2006). EEG has some advantages for use in operational environment. In particular, wireless solutions like the B-Alert system (Advance Brain Monitoring, Inc.) (Berka et al., 2005, 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.

We set up an experiment on a video game. We chose this particular set-up because 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 an ecological situation.

To get closer from operational constraints, we made the choice of constraining our recruitment: very few subjects, recruited on the basis of their availability.

We used Rayman Origins, 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.

(Cegarra & Chevalier, 2008) advise 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 (National Aeronautics and Space Administration Task Load indeX, Hart & Staveland 1988). Physiological measures are also used

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 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 measure.

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.

Methods

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.

Measures

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 & 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.

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.

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). 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 as advised by the constructor prior to averaging.
- HRV : we measured Heart Rate Variability (HRV) as another physiological indirect indicator of cognitive load (Backs, Lenneman, & Sicard, 1999; Wilson, 2002). We used the ECG channels of the B-Alert system to derive another measure of the Heart rate variability (HRV). 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).
- Eye Tracking: subjects were fitted with a mobile eye tracker, Tobii's glasses, a head-mounted eye tracking system resembling a pair of glasses. 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.

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 each lasted for 5 minutes. 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.

Participants were asked to play on two particular levels of the game, chosen for their duration and similarity in difficulty. 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 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. We kept a record of the number of time the participant died as a measure of performance.

Results

Sensitivity of the measures

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

As displayed in Figure 1, the Nasa-TLX score (**p=.025**) and the pupil dilation ratio (**p=.022**) were sensitive to our task manipulation of the mental 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.

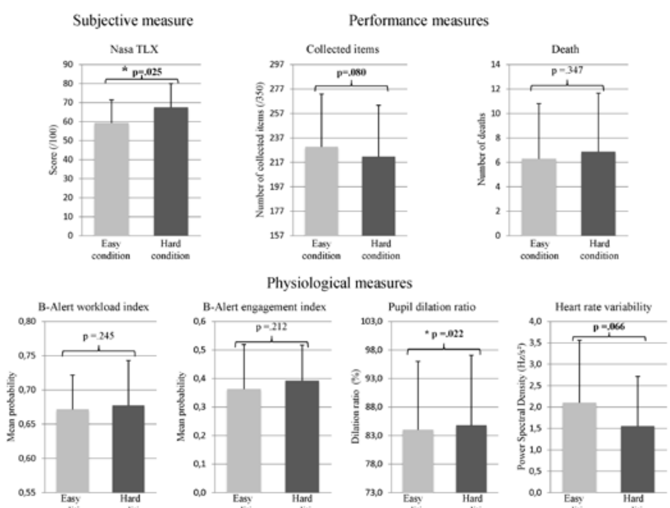

Figure 1. Across subject mean value for each measurement. The easy and hard conditions are compared. Error bars represent one standard deviation from the mean. P-value results from one-tailed Wilcoxon’s tests are displayed.

Consistency of the measures

We then analyzed the relations between the measures by computing Spearman correlations (non parametric) between the measures.

The NasaTLX score is correlated with both performance measures. As hypothesized, the higher the NasaTLX 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 NasaTLX score (R=0,486, p=0,042) and both performance measure (items: R=-0,663, p=0,004 ; deaths: 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 NasaTLX 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 NasaTLX score nor with the pupil dilation measure. The HRV measure is not correlated with any of the other measures.

Discussion

Validation of the experimental design

The NasaTLX subjective index and the pupil dilation measure were both significantly affected by our task manipulation of the mental workload, whereas the HRV did not. Hence, we can confirm that the experimental manipulation we used was successful at eliciting differential mental workload conditions (**H1**).

The performance measures showed a trend for the number of collected items.

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**).

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.

Evaluation of the EEG indices

As revealed by the Wilcoxon's tests, the EEG workload and engagement indices were not sensitive enough to our task manipulation of mental workload. The lack of results might be due an important inter-individual variability. We made the choice of recruiting few participants only.

For the workload index, the model is based on a group of individuals independently of our study and is not adjusted to each participant. The consumption of nicotine or caffeine thus has even more impact, 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 NasaTLX score (**H4**), suggesting that it is sensitive to the same underlying phenomena.

Conclusion

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 this index 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.

References

Antonenko, P., Paas, F., Grabner, R., & van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, *22*(4), 425–438.

Backs, R. W., Lenneman, J. K., & 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), 33–47.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., … Craven, P. L. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, Space, and Environmental Medicine*, *78*(Supplement 1), B231–B244.

Berka, C., Levendowski, D. J., Ramsey, C. K., Davis, G., Lumicao, M. N., Stanney, K., … Stibler, K. (2005). Evaluation of an EEG workload model in an Aegis simulation environment. In *Defense and Security* (pp. 90–99). International Society for Optics and Photonics.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., & Babiloni, F. (2012). Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*.

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

Cegarra, J., & 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., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Johnson, R. R., Popovic, D. P., Olmstead, R. E., Stikic, M., Levendowski, D. J., & Berka, C. (2011). Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biological Psychology*, *87*(2), 241–250.

Tsang, P. S., & Vidulich, M. A. (2006). Mental workload and situation awareness. *Handbook of Human Factors and Ergonomics, Third Edition*, 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), 3–18.