Comparative Analysis of Cognitive Tasks for Modeling Mental Workload with Electroencephalogram

Taeho Hwang, Miyoung Kim, Minsu Hwangbo, and Eunmi Oh, *Senior Member IEEE*

***Abstract* - Previous electroencephalogram (EEG) studies have shown that cognitive workload can be estimated by using several types of cognitive tasks. In this study, we attempted to characterize cognitive tasks that have been used to manipulate workload for generating classification models. We carried out a comparative analysis between two representative types of working memory tasks: the n-back task and the mental arithmetic task. Based on experiments with 7 healthy subjects using Emotiv EPOC, we compared the consistency, robustness, and efficiency of each task in determining cognitive workload in a short training session. The mental arithmetic task seems consistent and robust in manipulating clearly separable high and low levels of cognitive workload with less training. In addition, the mental arithmetic task shows consistency despite repeated usage over time and without notable task adaptation in users. The current study successfully quantifies the quality and efficiency of cognitive workload modeling depending on the type and configuration of training tasks.**

I. Introduction

With the progress in sensors and algorithms for electroencephalogram (EEG) data, it has become possible to perform real-time estimation of human cognitive states using EEG. In particular, mental workload, or the cognitive load that is related to working memory, is one of the most popular research topics. Mental workload is involved in smart device applications in a variety of fields such as education and mental healthcare. Recent researches have shown that the applications might be realized by utilizing simple dry and wireless EEG sensors such as Emotiv EPOC and Neurosky MindWave [1-2]. It is necessary to derive cognitive models to estimate workload level based on real-time EEG data. The cognitive models are generated by applying machine learning algorithms to EEG data recorded from subjects exposed to different workload levels in training sessions. Although there is no standardized method for manipulating a subject's workload level, we can take advantages of well-defined cognitive tasks that are known to be associated with working memory. At the very least, the cognitive task should be able to induce a user's baseline workload level and high-level workload. In addition, the cognitive task used in training sessions would be more effective if it is intuitive to typical users and time-efficient. Furthermore, these requirements should be met even in realistic use-case scenarios where convenient EEG sensors are more vulnerable to unwanted noises with respect to those used in experimental settings.

Taeho Hwang, Miyoung Kim, Minsu Hwangbo and Eunmi Oh are with the DMC Research Center, Samsung Electronics, Suwon-si, Gyeonggi-do, 443-742, South Korea. (phone: +82-31-279-0528, email: taeho.hwang@samsung.com)

The majority of the previous researches on EEG-based workload estimation has used simple and well-controlled working memory tasks that allow us to easily control difficulty level in order to induce at least 2 clearly separable workload levels (i.e., high and low). Mental arithmetic task and n-back task have been widely used to generate classification models to estimate workload level. The n-back task is used to measure subjects' working memory [3]. The n-back task requires a subject to view a sequence of letters or figures and answer whether the currently displayed letter or figure is the same as that which appeared n-steps before. The easiest form of the n-back task, the 0-back, is often used to set the baseline workload level, and subsequent n-back ($n > 0$) tasks are used to induce a higher level of workload by asking subjects to remember and compare a current letter or figure to the previously displayed one [4-6]. The 2-back task is preferred since the 3-back task has been shown to be too difficult for subjects to complete in order to properly induce a high workload [7]. Mental arithmetic tasks have also been frequently used to train model users' workload [8-9]. Typically, mental arithmetic tasks require subjects to solve several problems of addition between two numbers within a predetermined time. Mental arithmetic tasks do not rely on any tools (e.g., computers, pens and paper) and are known to be closely related to working memory [10]. Although many studies on these tasks show that users' workload can be estimated with acceptable accuracy based on EEG data, there has been little investigation of the feasibility of these tasks with regard to practical cases where a user with a simple EEG sensor performs a training session with no or little instruction about the tasks.

Thus, in this paper, we look into practical issues regarding n-back and mental arithmetic tasks in order to obtain accurate classification models of high and low workload that can be used in real applications. Given the dependency of tasks on the quality and efficiency of workload modeling, we compared the 2 tasks with the same participants and the same signal processing algorithms. First, we compared accuracy of classification models generated from each task to evaluate the consistency in inducing high and low workload levels within and between training and post-training sessions Second, we compared the time-efficiency of the tasks based on the between-session classification accuracy of the models achieved with varying lengths of the training session. Third, we conducted a time-course analysis to compare the 2 tasks in terms of consistency of workload induction in a series of post-training sessions over a longer period using intermediate sensor re-installation. Finally, we analyzed the behavioral data recorded during multiple post-training sessions to determine the existence of training effects. In Section III, we discussed the characteristics of tasks that can be considered to

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

generate reliable classification models of workload for real-time application.

II. METHODS

A. Experimental Design

Seven healthy subjects (n = 2 females; n = 5 males) aged between 25 and 39 years participated in the experiment. We collected written informed consent from all participants.

We employed the n-back task and the mental arithmetic task. A session of each task was composed of 2 difficulty levels, and each lasted approximately 3 minutes. For the n-back task, 0-back and 2-back formats were selected, as this combination shows the best classification accuracy versus other combinations (i.e., 0-back and 3-back) [4]. In each difficulty level, a series of alphabet letters were sequentially but independently displayed for 500 milliseconds with a 2000 millisecond inter-stimulus interval. For the 0-back task, participants were asked to press a keyboard when the letter "X" appeared. For the 2-back task, participants were asked to respond when the letter was identical to that which appeared two letters earlier. Alphabet sequence presentation was randomized, and the frequency of target letters was set to 30% for both the 0-back and 2-back tasks.

For the mental arithmetic task, we used multiplication problems of two positive integers with two levels of difficulty (i.e., 1-digit multiplication and 2-digit multiplication). To reduce potential fluctuations in difficulty during the 2-digit multiplication, we excluded problems where the product of the last digit did not exceed 10. We devised 2 variants of the mental arithmetic task with different test conditions. In the first condition, training time was fixed to 3 minutes for both the easy and hard tasks. For every trial, a problem was displayed on the screen until the subject entered an answer via a keyboard (hereafter referred to as "time-constrained"). There was no feedback regarding the correctness of the entered answer. In the second condition, the number of problems that participants had to solve was fixed at 40 for the easy level and 7 for the difficult level to balance training time. Instead of fixing the total training time, the training session persisted until the participants entered correct answers for every problem (hereafter referred to as "exhaustive").

Each participant was asked to perform 2 sessions (a training session and a post-training session) of the n-back task, and 2 variants of the mental arithmetic task. During each of these, intermediate breaks were provided; the EEG sensor was kept on participants' heads during these breaks.

For a time-course analysis, a subset of participants (n = 2 male and n = 1 female) were asked to perform 1 training session and 9 post-training sessions of the n-back task and the "exhaustive" mental arithmetic task over 5 days.

B. Data Recording and Processing

We recorded EEG data at a sampling rate of 128 Hz with 14 electrodes using Emotiv EPOC. We also collected response times and the rate of correct answers during task performance. We applied a common reference average for spatial filtering. The EEG data was split into 4-second length data fragments with 3.5-second overlaps. We used the short-time Fourier transform (STFT) to obtain large number of features of higher resolution given the potentially high temporal dynamics of EEG signal [12]. For each data fragment, we applied the STFT to extract the power of 8 equally sized (4Hz) frequency bands from 4 to 32 Hz in each of the 7 data sub-fragments of 1-second length with 0.5-second overlaps. We attempted to fully utilize a total of 784 features (14 channels × 8 frequency bands × 7 time points) in the classification model to explore as diverse features as possible. We used a support vector machine (SVM) with a linear kernel to identify and classify 2 levels of workload. The data recording and signal processing were implemented using Matlab.

III. RESULTS

For each participant, the average time to solve a 2-digit multiplication problem was significantly longer than that of a 1-digit multiplication problem (p = 0.003; paired t-test). These results imply that the 2 difficulty levels were properly devised.

A. Within-Session Consistency

The classification accuracy of the models was computed from the n-back task training sessions and the variants of the mental arithmetic tasks based on 10-fold cross-validation (Fig. 1). The accuracy was 98.6%, 94.2%, and 96.6% for the n-back task and the "time-constrained" and "exhaustive" mental arithmetic tasks, respectively. High overall accuracy (> 94%) indicates that all 3 tasks are capable of generating highly consistent and separable EEG signals during the training session. Six out of 7 participants showed the highest cross-validation accuracy possible in the n-back training session. The cross-validation accuracy of the n-back task was, however, significantly higher than those of 2 mental arithmetic tasks for each user (p = 0.028 and 0.027, respectively; paired t-test).

B. Between-Session Consistency

In addition to within-session consistency, we evaluated performance of the classification models generated above using the EEG data collected from post-training sessions. We assumed that the post-training session data should have similar patterns with the training data, given that the task was identical and the session was conducted immediately following the training session without re-installation of the EEG.

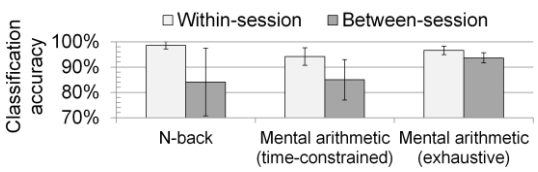

Figure 1. Within-/Between-session accuracy of the classification models. Lines on the bar indicate ranges of 1 standard deviation.

Nevertheless, classification accuracy decreased compared to those estimated by cross-validation (Fig. 1). This result emphasizes the importance of evaluating between-session classification accuracy for unbiased assessment of model quality in an online test.

Notably, the "exhaustive" mental arithmetic task showed the smallest decrease in classification accuracy compared to the n-back and the "time-constrained" mental arithmetic tasks. In particular, the decrease observed for the "exhaustive" mental arithmetic task was significantly smaller than that for the n-back task for each participant ($p = 0.05$; paired t-test).

This indicates that the "exhaustive" mental arithmetic task has more between-session consistency in inducing 2 workload levels than the other tasks do. Moreover, the between-session classification accuracy showed marginal difference between the 2 mental arithmetic tasks for each participant (mean difference=8.7%, $p = 0.05$; paired t-test), even though the task itself was identical. We could not find any significant behavioral difference between the training session and the post-training session in terms of mean reaction time ($p = 0.07$; paired t-test) or number of correctly answered problems ($p = 0.485$; paired t-test) per minute for both types of mental arithmetic task. Although the reason for better consistency in the "exhaustive" mental arithmetic task remains unclear, we postulate that a different paradigm or strategy for the same working memory tasks might have a crucial impact on the quality of the classification model by influencing the user's engagement or motivation during the training session. The "exhaustive" type of task design is thought to motivate the participants to engage in the task so that consistent EEG features related to workload are strongly detected in different sessions. On the other hand, workload modeling based on the other 2 tasks seems sensitive to session-specific signals rather than workload-related signals.

C. Time Efficiency of Training

We conducted an in-depth comparison of the training time required to produce high between-session classification accuracy between the n-back and mental arithmetic tasks. In this analysis, the "time-constrained" mental arithmetic task was included, as the training time was designed to be the same as that of the n-back task. We evaluated the between-session accuracy of the classification models generated from varying lengths of training EEG data (30, 60, 90, 120, 150, and 180 seconds) on the post-training session data (3 minutes) for each participant.

As shown in Fig. 2, the n-back task showed slightly higher classification accuracy than did the mental arithmetic task when using the 30-second training data.

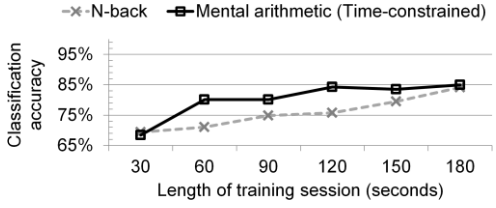

Figure 2. Between-session accuracy of the classification models with varying lengths of training session.

On the other hand, as the training time increased, the mental arithmetic task obtained relatively higher classification accuracy compared to the n-back task. For example, the average between-session classification accuracy of 80% was achieved with only the 60-second training session for each difficulty level for the mental arithmetic task, while the n-back task required more than double the length in training to produce the same level of accuracy. The classification accuracy was, however, saturated at a training time of 120 seconds or above for the mental arithmetic task. This result indicates that each of the working memory tasks might have different time-efficiencies and optimal lengths of training for generating an accurate classification model.

D. Consistency in Multiple Sessions over Longer Periods of Time

We investigated between-session accuracy of the classification models generated in the training session on multiple subsequent sessions over a longer period. For this analysis, 3 of the participants conducted 9 post-training sessions for each of the n-back and "exhaustive" mental arithmetic tasks over 5 days. The average classification accuracy of the 3 participants for the mental arithmetic task was maintained generally higher than that for the n-back task in the 9 post-training session over 5 days (mean difference=9.4%, $p = 0.049$; paired t-test). During the whole 9 post-training session, there was no significant difference between the variance of classification accuracies between the two tasks ($p > 0.05$; F-test).

The between-session accuracy of the classification model based on the n-back task, however, showed an overall decrease during the post-training sessions for all participants (Fig. 3a). Particularly, the average classification accuracy during the first two and the last two sessions showed notable difference (17.5%) in the n-back task. On the other hand, no such trend was observed for the mental arithmetic task (Fig. 3c). Similar to the previous result of between-session accuracy, this result implies that the classification models generated from the mental arithmetic task works more consistently also in the repeated use over long-period of time.

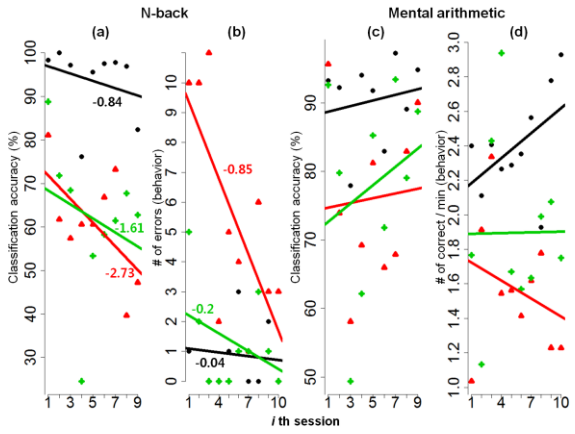

Figure 3. Trends of between-session accuracy and behavior over repeated post training sessions. The colors correspond to each participant. The numbers in (a) and (b) indicate the slope of each regression line.

We further analyzed the data from 3 participants in order to assess whether either task becomes easier through repeated training. We analyzed error trends and reaction time for the n-back task, and the number of correctly answered problems per minute for the "exhaustive" mental arithmetic task, during a total of 10 sessions for each task. As shown in Fig. 3b, the number of errors was drastically reduced as the post-training session was repeated (Fig. 3b). On the other hand, such trend was not observed for the mental arithmetic task (Fig. 3d). In case of the n-back task, the decreasing trend of between-session classification accuracy and behavioral indicators (error rate) seems to have a notable relationship (Fig. 3a and Fig. 3b); there was a positive correlation between the slopes of regression lines for the classification accuracy and for behavioral error ($r = 0.98$) among the 3 participants. Decrease of the between-session classification accuracy was more clearly observed for the subjects who showed more behavioral improvement during the repeated sessions. The result implies that the mental arithmetic task is a relatively consistent workload task despite repeated use.

IV. CONCLUSION AND DISCUSSION

Previous studies have supported the possibility of EEG-based workload classification using simple sensors with advanced algorithms. One factor to consider is the temporally efficient generation of accurate workload classification models in practical scenarios. Because it is virtually infeasible to make users perform a long period of training with arbitrary tasks, proper use of simple but well-defined cognitive tasks seems very important. This concern becomes more significant if our ultimate goal is to bring the technology into the real-world applications using EEG sensors with convenience but limited stability. Nevertheless, it is unclear what characteristics of the cognitive tasks should be considered for robust and efficient workload modeling that preserves high online classification accuracy. With these criteria, we compared 2 representative working memory tasks: the n-back and the mental arithmetic tasks. We used the Emotiv EPOC, a widely used wireless EEG sensor, to show how much modeling can be affected by the choice and settings of the training session.

In our experiments, we analyzed between-session classification accuracy for estimating online performance, between-session classification accuracy with varying lengths of training session, and the effect of repeated training over a long period. The features included in our classification model showed similar spectral change to those previously reported [13-15] for all 3 working memory tasks. We could find that the average theta power (4-8Hz) was increased in frontal channels (AF3 and AF4) under high workload. On the other hand, alpha power (8-12Hz) was decreased in several frontal regions covered by AF3, AF4, F3, F4, F7, F8, FC5, and FC6. Moreover, increase of gamma (near 32Hz) was observed in parietal (P8) and temporal (T8) regions. These observations support that our experiments and signal processing were properly devised to measure cognitive workload.

From our experimental results, we observed that the workload model based on the "exhaustive" mental arithmetic task that enforces participants to solve fixed number of problems without time constraint, consistently accounts for the post-training session data. The relatively poor between-session accuracy observed for the n-back task might have been caused by the following: i) the features associated with workload were not strong enough compared to other session-specific signals, which possibly originated from artifacts from repositioning of the sensor, body movement and fatigue during the task; and/or (ii) the participants adapted to task difficulty through repeated training so that workload induction became inconsistent compared to the first training session. It is yet unclear to figure out the direct cause but we could observe that repeated training on the n-back task resulted in relatively more user adaptation to the level of difficulty. Lastly, we found that the mental arithmetic task is efficient since it can achieve between-session classification accuracy within a shorter period of training session time. The classification accuracy of the n-back task showed steady and slow increase with longer length of training time as previously shown in [16].

Since our study contained a small number of participants and had rather specific formats for task designs, it may be hard to generalize whether the mental arithmetic task is always the better choice for workload modeling. Nevertheless, we believe that the results presented in our study may be helpful in understanding the impact of task choice and design for user training on the quality and efficiency of the resultant workload model, especially in practical situations with simple EEG sensors and limited instruction about the cognitive tasks. In order to develop robust models of workload, it is important to build a large EEG database from a large population of subjects performing diverse cognitive tasks. These subject-independent models should be further calibrated for each user to support more accurate real-time workload estimation. In any case, it is difficult to choose efficient and robust working memory tasks that can be applied to smart devices and associated applications for a real-time workload monitoring. Thus, we will further explore pragmatic configuration of cognitive tasks with larger number of populations by using the comprehensive comparative scheme.

REFERENCES

[1] J. N. Mak, R. H. M. Chan, and S. W. H. Wong, "Evaluation of Mental Workload in Visual-Motor Task: Spectral Analysis of Single-Channel Frontal EEG," *in Proc. 39th Annual Conference of the IEEE Industrial Electronics Society (IECON)*, Nov. 2013, pp.8426-8430.

[2] T. K. Calibo, J. A. Blanco, and S. L. Firebaugh, "Cognitive stress recognition," *in Proc. 2013 IEEE International Instrumentation and Measurement Technology Conference*, May 2013, pp.1471-1475.

[3] M. J. Kane, A. R. Conway, T. K. Miura, and G. J. H. Colflesh, "Working memory, attention control and the n-back task: A question of construct validity," *J. Exp. Psychol. Learn.*, vol.33, issue.3, May 2007, pp.615-622.

[4] A. M. Brouwer, M. A. Hogervorst, J. B. van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld, "Estimating workload using EEG spectral power and ERPs in the n-back task," *J. Neural Eng.* vol.9, no. 4, Aug. 2012.

[5] C. Walter, S. Schmidt, W. Rosenstiel, P. Gerjets, and M. Bogdan, "Using Cross-Task Classification for Classifying Workload Levels in Complex Learning Tasks," *in Proc. Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, Sept. 2013, 876-881.

[6] C. L. Baldwin, and B. N. Penaranda, "Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification," *Neuroimage*, vol.59, issue.1, Jan. 2012, pp. 48-56.

[7] M. Izzetoglu, S. C. Bunce, K. Izzetoglu, B. Onaral, and A. K. Pourrezaei, "Functional brain imaging using near-infrared technology," *IEEE Eng. Med. Biol. Mag.*, vol.26, July 2007, pp.38-46.

[8] S. I. Dimitriadis, Y. Sun, K. Kwok, N. A. Laskaris, and A. Bezerianos, "A tensorial approach to access cognitive workload related to mental arithmetic from EEG functional connectivity estimates," *in Proc. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* July 2013, pp.2940-2943.

[9] P. Zarjam, J. Epps, N. H. Lovell, F. Chen, "Characterization of memory load in an arithmetic task using non-linear analysis of EEG signals," *in Proc. 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* Sept. 2012, pp.3519-3522.

[10] Q. Wang, and O. Sourina, "Real-time mental arithmetic task recognition from EEG signals," *IEEE Trans. Neural. Syst. Rehabil. Eng.*, vol.21, issue.2, Mar. 2013, pp.225-232.

[11] R. H. Logie, K. J. Gilhooly, and V. Wynn, "Counting on working memory in arithmetic problem solving," *Memory & Cognition*, vol.22, Issue.4, July 1994, pp.395-410.

[12] A. F. Rabbi, K. Ivanca, A. V. Putnam, A. Musa, C. B. Thaden, and Reza Fazel-Rezai, "Human Performance Evaluation based on EEG Signal Analysis: A Prospective Review," *in Proc. 2009 31st Annual International Conference of the IEEE Engineers in Medicine and Biology Society (EMBS)*, Sept. 2009, pp.1879-1882.

[13] W. Kilmesch, "EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis," *Brain Research Reviews*, 1999, vol.29, pp.169-195.

[14] A. Mecklinger, A. F. Kramer, and D. S. Strayer, "Event Related Potentials and EEG Components in a Semantic Memory Search Task," *Psychophysiology*, 1992, vol.29, no.1, pp.104-119.

[15] M. W. Howard, D. S. Rizzuto, J. B. Caplan, J. R. Madsen, J. Lisman, R. Aschenbrenner-Scheibe, A. Schulze-Bonhage, and M. J. Kahana, "Gamma oscillations correlate with working memory load in humans," *Cerebral Cortex*, 2003, vol.13, pp.1369-1374.

[16] D. Grimes, D. S. Tan, S. E. Hudson, P. Shenoy, and R. P. N. Rao, "Feasibility and pragmatics of classifying working memory load with an electroencephalograph," *in Proc. SIGCHI Conference on Human Factors in Computing Systems*, April 2008, pp.835-844.