May 2021

Mental Workload Detection based on EEG Analysis

José YAURI [a,1], and Aura HERNÁNDEZ-SABATÉ [a] and Paul FOLCH [b] and Débora GIL [a]

[a] *Computer Vision Center, Universitat Autònoma de Barcelona, Spain*
[b] *Aslogic, Barcelona, Spain*

Abstract. The study of mental workload becomes essential for human work efficiency, health conditions and to avoid accidents, since workload compromises both performance and awareness. Although workload has been widely studied using several physiological measures, minimising the sensor network as much as possible remains both a challenge and a requirement.

Electroencephalogram (EEG) signals have shown a high correlation to specific cognitive and mental states like workload. However, there is not enough evidence in the literature to validate how well models generalize in case of new subjects performing tasks of a workload similar to the ones included during model's training.

In this paper we propose a binary neural network to classify EEG features across different mental workloads. Two workloads, low and medium, are induced using two variants of the N-Back Test. The proposed model was validated in a dataset collected from 16 subjects and shown a high level of generalization capability: model reported an average recall of 81.81% in a leave-one-out subject evaluation.

Keywords. Cognitive mental states, Mental workload, EEG analysis

1. Introduction

A cognitive state is the state of the mind, often named cognitive status, and it is related with the human performance and awareness in a specific time. Usually, cognitive states of workload, distraction, and fatigue are among the most studied due to their association to human performance and reliability and their risks for catastrophic effects in, for instance, aviation and automotive accidents [1,2,3].

In particular, if we define mental workload as the cognitive and psychological effort to conclude a task [4] we can observe that when workload is too heavy or too light it can degrade the human performance [5,6], so workloads have high effects on the daily life human performance. The more difficult the task is, the greater the mental workload [7] results. Thus, the study of workload becomes essential to prevent accidents, since it could compromise human task performance [8].

Since workload involves cognitive, neuro-physiologic, and perceptual processes to resolve a task, it is affected by individual capabilities, motivation to perform the task,

[1]Corresponding Author: Campus UAB, Edifici O, s/n, 08193 Cerdanyola del Vallès, B, Barcelona, Spain; E-mail: jyauri@cvc.uab.cat

and its physical and emotional state [9]. This multifaceted nature of workload prevents to study workload directly, but it is feasible to be inferred from different quantifiable variables [10]. There exist many proposals for recognizing workload based on physiological features, such as hearth rate, eye movement and dilation, electroencephalogram (EEG) and electrocardiogram (ECG) [11,12].

Besides, the recent emerging of low cost EEG headsets has driven to new researches further than medical screening of neurological disorders. Mentally control robotic arms, drive wheelchairs, and interact with home devices are closer but also improving teaching-learning educative methods are feasible thanks to EEG signals research. In the particular case of cognitive state, EEG alone is becoming the preferred sensor for addressing its characterization [13,14,15]. However, there is not enough evidence in the literature to validate how well models generalize in case of new subjects performing tasks of a workload similar to the ones included during model's training.

In this paper we propose the use of EEG for characterizing workload by means of a neural network and show its ability to generalize the model across a wider population.

The remainder of this paper is organized as follows: Section 2 presents relevant related works, Section 3 explains the process followed to collect the data. Section 4 presents our proposal for the analysis of EEG generalization capabilities, while Section 5 is devoted to present the results of our experiments. Finally, Section 6 outlines the conclusion and future work.

2. Related work

The most generalized mechanisms to measure workload can be split in two main categories [7,11,1]: subjective measures and physiological ones. On the one hand, subjective measures are still the most used to assess mental workload, being the NASA Task Load Index (TLX) [16] the most prominent to gain insight about the perceived workload levels while a subject works with various human-machine interface systems [4,17]. It measures the mental workload based on a weighted average of six sub-variables: mental demand, physical demand, temporal demand, performance, effort and frustration and it is widely used in aviation to assess mental workload of pilots while interacting with plane controls [18,19]. On the other hand, physiological measures provide a more reliable data of workload by measuring physiological dynamic changes which cannot be controlled consciously, so they are becoming more popular among researchers in recent years [20,21,22]. The most common sensors to record physiological data area: electrocardiogram (ECG) to register heart's electrical activity, electromyograph to read skeletal muscles electrical activity, electroencephalogram (EEG) to detect electrical activity in the brain, photoplethysmography to register volumetric changes in the blood flow, respiration rate sensors, electro-dermal activity (EDA) to read skin surface temperature, oxygen density in the blood in the brain, and eye movement trackers, among others [23]. TLX surveys allow to assess the perceived workload [16], but it is highly subjective. However, physiological data occurs spontaneously, and, together with TLXs, provide a more reliable information [20,11,4].

The combination of several physiological sensors to classify workload states gives better results than using a single one. The approach proposed in [24] combines EEG, ECG, and electrooculography (EOG) and results show a highest predictive power for

their combination (80%) rather than the analysis of each one independently (70%). Besides, the study in [12] reports an accuracy average of 85.2 ($\pm$ 4.3%) combining EEG, ECG, respiration rate, and EDA to classify 4 mental states. The work in [25] still shows better results combining EEG, ECG and EDA than using only EEG signal from classifying four mental states, although results from the single sensor are promising (86.66%).

At that point, Deep Learning (DL) approaches are gaining ground over more classical machine learning techniques due to their ability to automatically extract the features [23,26]. For instance, the study in [8] proposes a concatenated structure of deep recurrent and 3D convolutional neural networks to combine both raw and spectral EEG data and assess two degree of mental workloads reporting an average accuracy of 88.9% in a cross-task assessment. However, none of the last previous works were tested on a dataset totally unseen in the training set, being their ability to generalize an unknown.

In this work, we propose to investigate the ability of 1D-CNN models to recognize two types of mental workload from EEG signals and generalize the model to an unseen population in the training set. To induce low and medium workload, we use the N-back test [27], while for workload classification, we train a simple neural network (NN) using only the power spectrum of theta waves. To assess the generalization abilities of the general model we propose a a personalized model for each individual and a generalist one. We perform our models in a new dataset collected from 16 subjects, showing outstanding results in a leave-one-out subject test, so the model highly generalizes to new unseen subjects.

3. Data Set Collection

When performing different tasks along the day, people experiment different levels of mental workload depending on the level of attention required, the difficulty of such task and how many sub-tasks are needed to take care off. In order to induce different levels of workload in a controlled manner, we implemented different variants of the N-Back-test [27].

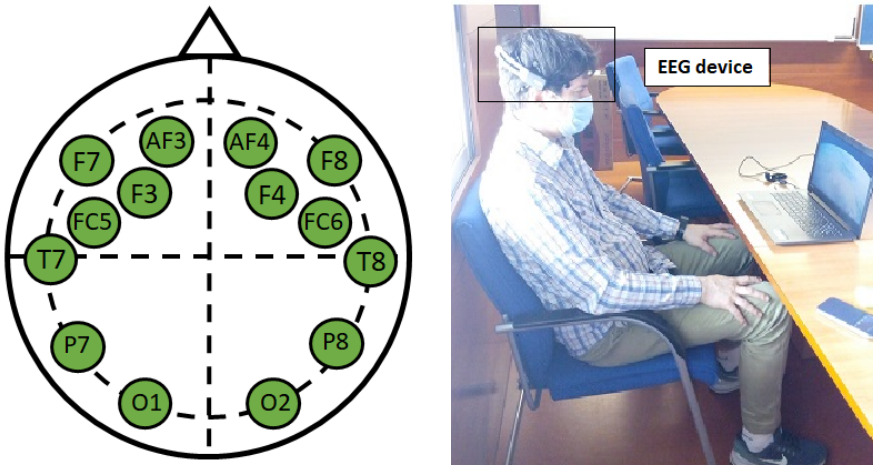
N-Back-tests are memory demanding games requiring the resolution of simple arithmetic operations. A grid with a number in varying positions is shown to the player. The player has to introduce the result of an arithmetic operation using the number currently shown and the number shown to the player N-screens before and (for some variants) recall whether both numbers were or not in the same position. We implemented three variants of the N-Back-test [27] to induce low, medium, and high mental workload:

1. *Position 1-back for low workload*. A square appears every few seconds in one of eight different positions on a regular grid over the screen. Players must press a keyboard key in case the position of the square on the current screen is the same as the square of the previous grid.
2. *Arithmetic 1-back for medium workload*. An integer number between 0 and 9 appears every few seconds on the screen while an audio message says an arithmetic operation (plus, minus, times and divide). Players have to solve this operation using the current number and number that appeared in the previous screen.
3. *Dual arithmetic 2-back for high workload*. This test combines the two previous ones. An integer number between 0 and 9 appears every few seconds in one of eight different positions on a regular grid. At the same time, for each number that

appears on screen, an operator is presented with an audio message. As before, players have to solve this operation using the current number and number that appeared in two screens before. In addition, players have to press a key in case the position of the current number is the same as the position of the number shown two screens before.

The neurophysiological response of a subject against mental demanding tasks depends on its baseline state, which is prone to vary accross time. In order to account for differences in the baseline state of subjects, previous to the N-back-tests participants watched a relaxing video for 10 minutes. For each experiment (low, medium and high workload), we call the video watching, phase 1, and the N-back-test, phase 2. After the game, participants ask a TLX questionaire to collect their subjective perception of game difficulty and workload.

Although, this work presents results on EEG, we also recorded the electrocardiogram (ECG) data during the video watching and the game. For EEG recording, we used the EMOTIV EPOC+ headset [28] which has 14 electrodes placed according to the 10/20 system. This sensor provides both raw data and power spectrum for the main brain rhythms (theta, alpha, beta low, beta high, and), at 128 Hz and 8 Hz, respectively. Figure 1 illustrates the distributions of electrodes of this sensor (a) and a volunteer during a session task (b).



(a) (b)

Figure 1. Data collection with Emotiv Epoc+ headset. (a) Electrodes distribution over the head scalp. (b) A volunteer during a N-Backtest.

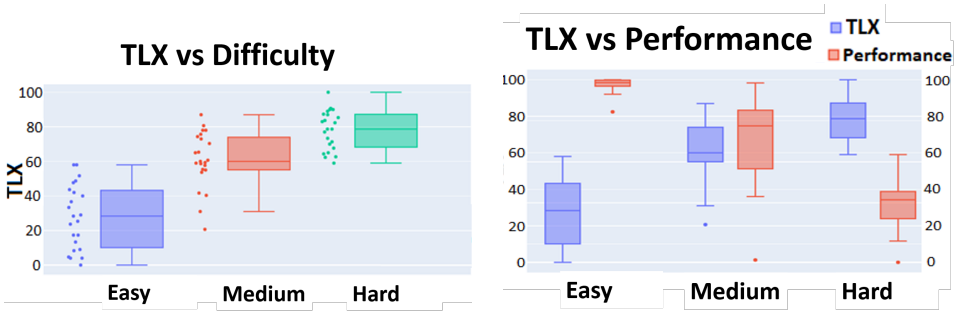
A total of 24 subjects participated in the experiment. Subjects were adults between 20 and 60 years, all of them were healthy without any condition that might have cause an imbalance in the data recorded. The sequence of tasks were randomly assigned to subjects, and recording of each session was in different days and hours.

4. Machine learning approaches

In order to assess to what extend a general model trained over a set of individuals can successfully predict a new unseen individual, we have considered the following approaches for the analysis of EEG generalization capabilities:

- **Personalized model for each individual**. A different model is trained for each subject of the data set to account and compensate for large intra subject variability in EEG signals.
- **Generalist model for the population**. A single model using all subjects is trained to assess whether inter subject variability can be properly modelled.

For both approaches we implemented a binary neural network trained to classify between workload (phase=2) and base lines (phase=1) phases. The experiment used to define the training data of the WL class was the phase 2 of the second experiment (noted as WL2). The phase 2 of the first experiment was discarded as training data because, according to TLX, it did not demand any significant mental effort for most users (Figure 2.a). The phase 2 of the third experiment was also discarded as training data because, according to TLX and users' performance, most users considered the task too difficult and gave up at some point of the experiment (Figure 2.b). Regarding the baseline class, all phases 1 can be considered for training. This data will be noted BLi, i=1,2,3, for i indicating the experiment. In order to discard any dependency of models with respect base line acquisition, we trained 3 different models for each approach using a different BLi for the base line class: WL2 vs BL1, WL2 vs BL2 and WL2 vs BL3. Additionally, for the generalist approach we trained an extra model using all 3 baselines phases in an attempt to account for any variability across them and better model the space of the baseline class. This model will be noted as BLall vs WL2.



(a) (b)

Figure 2. TLX-based subjective perceptions. (a) Perceived difficulty of tasks. (b) Achieved performance on tasks.

Given that proposed N-back tasks are memory demanding stressing games and base line phases consist in watching a relaxing video, the theta wave [29] is the best candidate for discriminating the different mental loads of our experimental phases. In this work, we use the power spectrum of theta wave (4–8 Hz) sampled at 8 Hz.

Eye blinking and sudden head movements introduce abrupt sharp peaks of large amplitude in the power spectra wave that should be filtered before using them as predictors

of a mental state [20]. In particular, we use an Inter Quartile Range (IQR) [30] filtering strategy to detect outlier values associated to muscular movement wave peaks. Our IQR filtering is based on setting the value of the 99% percentile of the distribution to all points above it.

To ensure a high quality of signals, we further filter data according to the quality of the EEG during recordings provided by the headset itself. For each sensor and recorded sample, Emotiv reports the quality of the recording in a discrete scale with values in 0,1,2,3,4 indicating how good the contact between sensor and head is: 4 for optimum; 3 for good; 2 for medium; 1 for bad; 0 for none. For the sake of data with the highest possible quality while keeping a reasonable sample size signals with a 25% of bad recordings are discarded (<3). Further, since there is no evidence about what are the most discriminative sensors that best correlate to the detection of mental workload, the whole phase is discarded if the signal of two or more of the sensors has a low quality. Finally, a subject is discarded if either all its base line or its workload phases are discarded, since, in this case, there is not enough data to define the binary classification. After this quality filtering, only 16 of the 25 subjects were selected for models training and testing.

In order to feed data to models, θ signals were cut in temporal windows of 5 seconds without overlap [12]. The input variables of the networks are the concatenation of the 5 second windows of IQR filtered theta signals for the 14 EEG sensors. The concatenation of the 14 windows, defines a 14*40=560-dimensional feature space. In order to account for the difference in units and magnitudes, variables were normalized to have 0 mean and sigma=1 using the mean and standard deviation of the training set.

Networks have a hidden layer of 128 neurons with rectified linear (ReLU) as activation function. A dropout (0.1) layer was added before the classification layer to alleviate overfitting. All models have been trained using a weighted cross-entropy loss to compensate for the different length of base line and workload phases which introduces some unbalance in data samples.

5. Results

The performance of the different approaches for detection of mental workload has been assessed using the accuracy (or sensitivity) for each class. Sensitivity measures the capability of the system to detect BL and WL classes. Since we have a binary classification problem with WL the positive class, then the sensitivity for BL corresponds to the specificity of the model.

In order to validate the reproducibility of each model, the following experiments have been conducted:

1. **Model Personalized for each Individual**. Reproducibility of personalized models has been assessed at intra-experiment level. For each model trained with a different base line, WL2 vs BLi, i=1,2,3, 10% of the samples were randomly chosen for testing the capability of discriminating workload at different times of the task. A high accuracy would proof that the variability of the EEG is stable and low while continuously repeating the same task.
2. **Generalist Population Model**. The validation of the capability for modelling a population was tested using a leave-one-out scheme for the 4 models considered:

the 3 trained using a single baseline, BLi vs WL2, i=1,2,3, and the one trained using all 3 baselines phases, BLall vs WL2.

Table 1 summarizes the recall of baselines (BL) and workload (WL2) for the intra-experiment reproducibility. We report the 95% confidence interval for each class computed for all subjects (for each subject the average of BLi vs WL2, i=1,2,3 is computed). The overall recall for both classes is above 90%, which shows that workload and base line signals are different regardless of the time the experiment was conducted. However, the variability is large, which might be attributed to a suboptimal size of the temporal windows and the variability in mental effort across the task.

Table 1. Personalized model. Intra-experiment Reproducibility.

BL	WL2
92.8 ± 7.03	91.17 ± 5.35

Table 2 summarizes the recalls of baselines (BL) and workload (WL2) for the generalist model trained using a single BL and the agregation of the 3. The model trained aggregating the 3 baselines has a higher performance in detecting baseline states. According to a Student t-test of paired data this difference is significant (p-val= 0.0054) with an average improvement range of (-17.2522, -3.5812). Regarding detection of workload phases, both approaches perform similarly (p.val= 0.7159). For both approaches, there are 3 outliers in WL detction rate that, given the small sample size, are highly influential. If we remove them, we have that for the remaining 80% of the subjects, the average detection of idle and work load stages for the model that aggregates all BLs for training is, respectively, 76% and 73%. This suggests that the variability and nonstationarity that psychophysiological data exhibits could be modelled if enough data from subjects was gathered.

Table 2. Generalist Model.

	Model trained with single BL		Model trained with all BLs	
	BL	WL2	BL	WL2
All population	66.57 ± 13.11	65.42 ± 25.59	78.06 ± 10.75	65.00 ± 24.90
80% of population	66.10 ± 13.87	73.95 ± 18.62	76.08 ± 10.87	73.23 ± 19.07

6. Conclusions

The first experiment (Table 1) shows that work load and base lines signals are different regardless of the time the experiment was conducted. However, the large variability in accuracies indicates that the temporal window might be suboptimal and should be adapted to the variant mental effort across a given task. Results of the generalist model show that the variability in baseline cognitive states can be modelled provided that enough training data is available. This is supported by the higher performance of models aggregating all baselines.

The analysis of the results suggests the following improvements.

A delicate issue that has an impact in the performance of methods is the filtering of signals required to remove muscular motion peaks and other artefacts. EEG preprocessing approaches have not been standardized, and even small changes in artefact removal strategy may result in differences with large effects on particular portions of the signal. In this study, we have adopted a filtering approach based on signal probabilistic distribution for outlier removal in the temporal space. We consider that muscular motion could be filtered calibrating muscular signals before test recording to set either the values or the frequency ranges associated to muscular motion. In this context, a classifier based on Fourier features will be further investigated.

Also the size of the temporal window might be a critical issue in order to properly include workload peaks. We have use 5 seconds windows following (Han, 2020), but recent authors suggest to use longer windows to capture EEG non stationary nature. The optimal window size should be further investigated.

In a future work, we have to ensure the availability of more data to achieve convergence without overfitting and to train more complex architectures. Recent architectures like convolutional/LSTM and Lambda Nets including attention modelling will be also studied.

Acknowledgements

This research is partially supported by the H2020 Clean Sky 2 project: "E-PILOT: Evolution of cockPIt operations Levering on cOgnitive compuTing Services" (Grant Agreement No. 831993), Spanish projects RTI2018-095209-B-C21 and Generalitat de Catalunya, 2017-SGR-1597, 2017-SGR-1624 and CERCA-Programme. DGil is a Serra Hunter Fellow.

References

[1] da Silva FP. Mental Workload, Task Demand and Driving Performance: What Relation? Procedia - Social and Behavioral Sciences. 2014 dec;162:310-9.

[2] Wickens CD. Situation awareness and workload in aviation. Current Directions in Psychological Science. 2002 aug;11(4):128-33. Available from: https://journals.sagepub.com/doi/10.1111/1467-8721.00184.

[3] Loft S, Sanderson P, Neal A, Mooij M. Modeling and predicting mental workload in en route air traffic control: Critical review and broader implications. Human factors. 2007;49(3):376-99.

[4] 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. Neuroscience & Biobehavioral Reviews. 2014;44:58-75.

[5] Proctor RW, Van Zandt T. Human factors in simple and complex systems. CRC press; 2018.

[6] Shaw JB, Weekley JA. The effects of objective work-load variations of psychological strain and post-work-load performance. Journal of Management. 1985;11(1):87-98.

[7] Averty P, Collet C, Dittmar A, Athènes S, Vernet-Maury E. Mental workload in air traffic control: an index constructed from field tests. Aviation, space, and environmental medicine. 2004;75(4):333-41.

[8] Zhang P, Wang X, Zhang W, Chen J. Learning Spatial-Spectral-Temporal EEG Features With Recurrent 3D Convolutional Neural Networks for Cross-Task Mental Workload Assessment. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2019 jan;27(1):31-42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30507536.

[9] Li D, Wang X, Menassa CC, Kamat VR. Understanding the impact of building thermal environments on occupants' comfort and mental workload demand through human physiological sensing. In: Start-Up Creation. Elsevier; 2020. p. 291-341.

[10] Hendy KC, Liao J, Milgram P. Combining time and intensity effects in assessing operator information-processing load. Human Factors. 1997;39(1):30-47.

[11] Heine T, Lenis G, Reichensperger P, Beran T, Doessel O, Deml B. Electrocardiographic features for the measurement of drivers' mental workload. Applied ergonomics. 2017;61:31-43.

[12] Han SY, Kwak NS, Oh T, Lee SW. Classification of pilots' mental states using a multimodal deep learning network. Biocybernetics and Biomedical Engineering. 2020;40(1):324-36.

[13] Zhang P, Wang X, Chen J, You W, Zhang W. Spectral and Temporal Feature Learning with Two-Stream Neural Networks for Mental Workload Assessment. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2019 jun;27(6):1149-59. Available from: https://pubmed.ncbi.nlm.nih.gov/31034417/.

[14] Lee DH, Jeong JH, Kim K, Yu BW, Lee SW. Continuous EEG Decoding of Pilots' Mental States Using Multiple Feature Block-Based Convolutional Neural Network. IEEE Access. 2020;8:121929-41.

[15] Wu EQ, Peng X, Zhang CZ, Lin J, Sheng RS. Pilots' fatigue status recognition using deep contractive autoencoder network. IEEE Transactions on Instrumentation and Measurement. 2019;68(10):3907-19.

[16] Hart SG. NASA-task load index (NASA-TLX); 20 years later. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 50. Sage publications Sage CA: Los Angeles, CA; 2006. p. 904-8.

[17] Index L. Results of empirical and theoretical research. Advances in. 1990.

[18] Wickens CD. Situation awareness and workload in aviation. Current directions in psychological science. 2002;11(4):128-33.

[19] Parasuraman R, Sheridan TB, Wickens CD. Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. Journal of cognitive engineering and decision making. 2008;2(2):140-60.

[20] Wang Z, Yang L, Ding J. Application of heart rate variability in evaluation of mental workload. Chinese journal of industrial hygiene and occupational diseases. 2005;23(3):182-4.

[21] Stanton N, Salmon PM, Rafferty LA. Human factors methods: a practical guide for engineering and design. Ashgate Publishing, Ltd.; 2013.

[22] Jang EH, Park BJ, Kim SH, Chung MA, Park MS, Sohn JH. Classification of human emotions from physiological signals using machine learning algorithms. In: Proc. Sixth Int'l Conf. Advances Computer-Human Interactions (ACHI 2013), Nice, France. Citeseer; 2013. p. 395-400.

[23] Deep Learning in Physiological Signal Data: A Survey. Sensors. 2020 feb;20(4):969. Available from: https://www.mdpi.com/1424-8220/20/4/969.

[24] Ziegler MD, Russell BA, Kraft AE, Krein M, Russo J, Casebeer WD. Computational models for near-real-time performance predictions based on physiological measures of workload. In: Neuroergonomics. Elsevier; 2019. p. 117-20.

[25] Secerbegovic A, Ibric S, Nisic J, Suljanovic N, Mujcic A. Mental workload vs. stress differentiation using single-channel EEG. In: CMBEBIH 2017. Springer; 2017. p. 511-5.

[26] Ismail Fawaz H, Forestier G, Weber J, Idoumghar L, Muller PA. Deep Learning for Time Series Classification: A Review. Data Mining and Knowledge Discovery. 2019 jul;33(4):917-63.

[27] Jaeggi SM, Buschkuehl M, Jonides J, Perrig WJ. Improving fluid intelligence with training on working memory. Proceedings of the National Academy of Sciences of the United States of America. 2008 may;105(19):6829-33. Available from: www.pnas.orgcgidoi10.1073pnas.0801268105.

[28] Emotiv. EMOTIV EPOC+ 14-Channel Wireless EEG Headset;. Available from: https://www.emotiv.com/emotivpro/.

[29] Addante RJ, Watrous AJ, Yonelinas AP, Ekstrom AD, Ranganath C. Prestimulus Theta Activity Predicts Correct Source Memory Retrieval. Proceedings of the National Academy of Sciences of the United States of America. 2011 jun;108(26):10702-7. Available from: https://www.pnas.org/content/108/26/10702https://www.pnas.org/content/108/26/10702.abstract.

[30] Wasserman L. All of statistics : a concise course in statistical inference. Springer; 2010.