Auto detect cognitive load using machine learning.

Quang Dang, Faisal Rasheed Khan, Sravya Chirakata

Abstract

Cognitive load significantly influences an individual's performance and memory retention. This study aims to explore the relationship between cognitive load and pupil dilation and examine the reliability of AI in detecting cognitive load using pupil data. We analyze data collected from 58 subjects performing various mental tasks and utilize Convolutional Neural Networks (CNN) to classify the presence or absence of cognitive load. Our research contributes to the development of adaptive systems that deliver tailored learning and workload experiences for individuals, ultimately improving user experience, satisfaction, and performance.

1. Introduction

In this paper, we investigate the potential of artificial intelligence (AI) in detecting cognitive load by analyzing pupil dilation. Recognizing cognitive load is essential for optimizing learning and working environments, enabling the allocation of appropriate workloads for individuals. We explore the reliability of AI in predicting cognitive load by examining performance metrics such as recall rate and precision, as well as identifying the contributing factors to this detection.

1.1. Motivation

Cognitive load refers to the mental effort required to process incoming information, which significantly influences an individual's performance and memory retention [1] (Sweller, J., 1988). Each person has a unique working memory capacity, making it essential to accurately identify their cognitive load. Recognizing cognitive loads allows us to determine when individuals are either overwhelmed or not sufficiently challenged, enabling us to allocate the appropriate amount of workload for each person. This optimization enhances learning and working environments, promoting efficient information processing and long-term knowledge retention Paas, (F., & Van Merriënboer, J., 1994). By understanding cognitive load, we can develop adaptive systems that deliver tailored learning and workload experiences for individuals, improving user experience, satisfaction, and performance (Kirschner, 2006). Such systems also have the potential to contribute to personalized learning experiences and more effective learning outcomes.

1.2. Problem

When an individual experiences cognitive load, their mental effort increases, subsequently activating the sympathetic nervous system. This activation puts the body into a "fight or flight" state, resulting in physiological changes such as increased heart rate, blood pressure, and pupil dilation (Arnsten, A. F. 2009). Our research aims to explore the relationship between cognitive load and pupil dilation, which has been widely investigated in previous studies (Cohen Hoffing RA, 2020)[2].

The primary question we seek to examine is whether artificial intelligence (AI) can accurately detect cognitive load by analyzing pupils. We will investigate the reliability of AI in predicting cognitive load by considering performance metrics such as recall rate and precision. Specifically, we will explore the balance between the AI's ability to correctly identify instances of high cognitive load (recall rate) and its overall accuracy in detecting cognitive load without generating false positives (precision). The secondary question will investigate the reliability of AI in predicting cognitive load and identify the contributing factors to this detection. Specifically, we will explore whether pupil diameter, pupil position, or a

[1] The concept of cognitive load was first introduced by John Sweller in the late 1980s, highlighting the importance of mental effort in processing information and its impact on learning outcomes.

[2] Pupil dilation has been widely used as a physiological indicator of cognitive load, as it reflects the activation of the sympathetic nervous system and the allocation of cognitive resources.

combination of both factors play a significant role in AI's ability to detect cognitive load. Additionally, we will consider potential limitations and confounding variables, such as lighting conditions and individual differences, that may affect AI's performance in this context.

2. Previous Research

There is a considerable body of research examining the relationship between cognitive load and human responses. One such study conducted by Xiaolin Yu (2008) investigated autonomic nervous activity during mental arithmetic tasks (MA). While Yu's study shares similarities with our research in terms of task design, the primary distinction lies in the analysis of brain signals through electroencephalograms (EEG) rather than examining pupil responses as we do.

DiCriscio (2018) also explored pupil responses during mental tasks, which aligns with our research focus. However, their study employed specialized tasks that are not as widely recognized as the common mental tasks utilized in our investigation.

Hoffing (2020) conducted a study most closely related to our research, analyzing the relationship between pupil data and cognitive processes during MA tasks. Our preliminary analysis is consistent with their findings, confirming the connection between pupil responses and cognitive load. However, Hoffing's research stops at establishing this relationship, whereas we aim to further explore the potential of artificial intelligence (AI) in detecting cognitive load through pupil analysis—an area that remains largely uncharted.

3. Solution

Our approach is to assess cognitive load by analyzing changes in the pupil diameter (PD) when performing various tasks, such as the Psychomotor Vigilance Task (PVT), Dot Probe Task (DPT), Mental Arithmetic (MA), and Visual Working Memory (VWM). We measure PD, as well as its X and Y coordinates.

In our data, Stimulus Time (ST) will mark the beginning of tasks our subjects. ST will also represent the beginning of cognitive load. Our aim is to identify the location of ST which will determine where the cognitive load begins.

To do this, we convert it into a binary classification problem, where we divide the data into smaller samples of one-second intervals and label them as true "1" if they contain ST and false "0" if they do not.

We start by cleaning the data, either by dropping missing samples or interpolating the data. We use a Convolutional Neural Network (CNN) to reduce the data's dimensionality and extract important features related to the cognitive load of the tasks. CNNs help in analyzing patterns and identifying significant features in the data.

4. Data analysis

We obtained the data from professor Juntin Brooks. There are a total of 58 subjects participating in our study. All subjects performed the tasks at University of California, Santa Barbara. Each subject will perform four distinct mental tasks, with the option to retry each task up to ten times. The chosen tasks represent popular cognitive tests, designed to measure different aspects of mental functioning. These tasks are as follows:

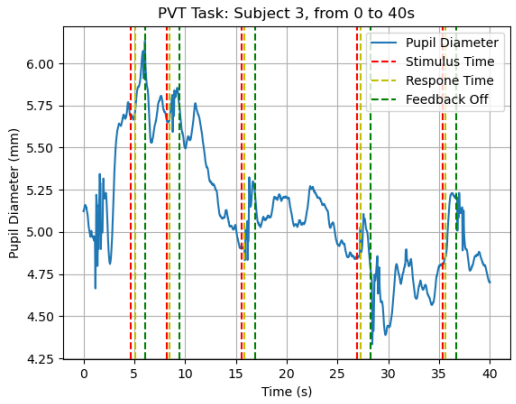
- Psychomotor Vigilance Task (PVT): A sustained-attention test that measures reaction times in response to visual stimuli, typically used to assess fatigue and alertness levels.
- Dot Probe Task (DPT): A test that evaluates attentional bias by presenting participants with two stimuli (one neutral, one emotional) simultaneously and measuring response times to a subsequent probe appearing in the location of one of the stimuli.
- Mental Arithmetic (MA): A task that involves solving arithmetic problems mentally, often used to induce cognitive load and assess numerical processing skills.
- Visual Working Memory (VWM): A task that measures an individual's ability to maintain and manipulate visual information in their working memory, typically through the presentation and recall of visual stimuli.

The data collection process involves recording at a sample rate of 250 Hz, capturing the subject's Pupil Diameter (PD), and the pupil's position on the X and Y axes (Pupil X and Pupil Y). The data also document the time when the test begins (Stimulus Time), the time when the subject responds (Response Time), and whether their response is correct or not.

The stimulus time (ST) represents the cognitive load experienced by the subject, and our goal is to accurately identify the ST using machine learning.

4.1. Data Visualization

To enhance our analysis, we aim to visualize each sample individually. We create plots using PD and time. Here are two example plots:



- The red line represents the stimulus time, which is the target we want to predict.
- The yellow line indicates the user's response time, capturing the moment when the subject reacts to the stimulus.
- The green line signifies when the feedback turns off, marking the end of the subject's interaction with the task.

By examining these plots, we can gain a better understanding of the relationship between PD, stimulus time, and response time, providing valuable insights to decide which feature we might consider extracting from the data.

4.2. Missing Data

We have observed that a significant portion of our data is missing, accounting for 18.46% of the total dataset. To address this issue, we consider using linear interpolation to fill in the missing data. However, we are concerned that the accuracy of our dataset may be compromised if there are large consecutive gaps in the data.

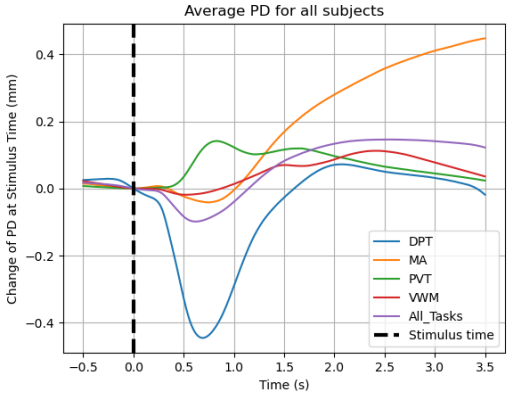
To assess the potential impact of consecutive missing data, we have conducted a statistical analysis with the following results (units are in data point):

- Mean: 728.6144
- Std : 2,694.344
- Min : 1
- 25th percentile: 32
- 50th percentile: 143
- 75th percentile: 659
- Max : 71,997

Given that our sample rate is 250 Hz, half of the missing data spans less than 1 second, and three-fourths of the missing data is less than 3 seconds. Based on these findings, we can reasonably conclude that it is possible to salvage the majority of the dataset using linear interpolation, without significantly compromising the integrity of the data.

4.3. Pupil Diameter

Our goal is to identify common behaviors among our subjects and learn how their PD changes after STs started. To accomplish this, we divided our analysis into five parts. The first four parts analyze each task individually: DPT, MA, PVT, and VWM. In the final analysis, "All Tasks," we look at the average across all four tasks for each subject. To normalize the PD data, we measure the change in PD by comparing it to when the STs began, not the actual PD itself. Therefore, time zero indicates when the ST started, and the PD at this point is zero for all subjects. We averaged all 58 subjects' PDs for every STs and plotted the results:

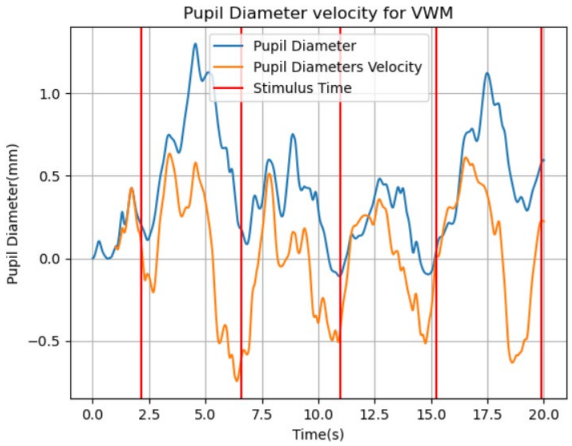


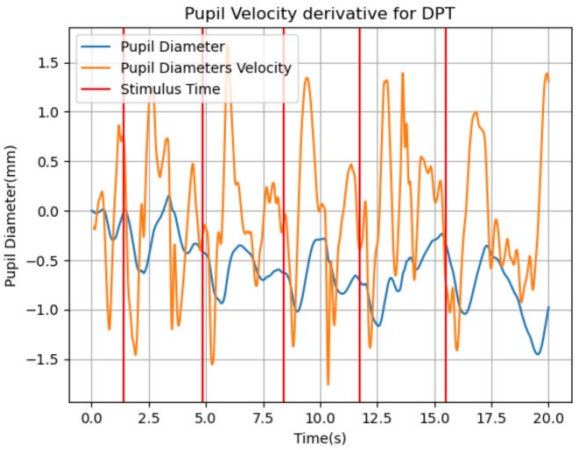
Our result aligned with Hoffing's research, except for the PVT task. In general, the data showed an initial constriction in PD followed by rapid dilation, resulting in a peak PD larger than the initial PD at two seconds after starting the ST. Hoffing found that the peak in PD was higher for tasks with higher cognitive loads, such as the MA task. The DPT task showed a clear constriction, while the MA and VWM tasks had a small constriction before dilation. However, in the PVT task, we did not observe the initial constriction phase. We found a similar pattern of dilation in PD. Our analysis identified a common pattern of PD behavior, which we plan to use machine learning to automatically detect.

4.4. Pupil Diameter's Velocity Analysis

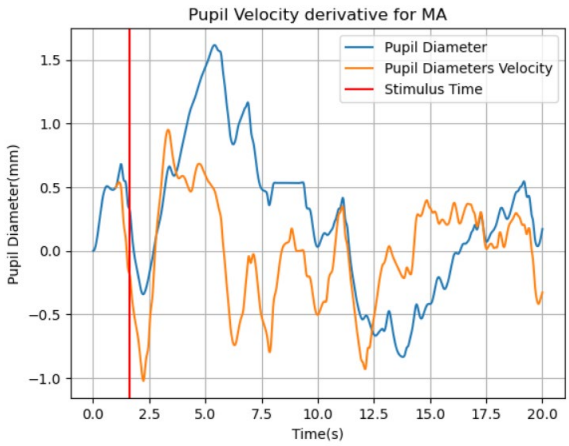
Based on section 4.3, we observed a constriction in PD at the start of ST. Our goal is to amplify this initial constriction in PD. To achieve this, we aim to calculate the velocity of PD by taking its derivative. The velocity will accentuate the constriction in PD.

To conduct this experiment, we will set up a similar procedure as described in section 4.3. We will plot a graph with PD and PD's velocity on the Y-axis, while time will be represented on the X-axis. In order to normalize the PD values, we will subtract the initial PD value at time 0 from all subsequent PD values. To calculate the velocity, we will employ the forward difference approximation method. In the following plots, we will use a time interval of 1 second for differentiation. We have also experimented with different time intervals. When the time intervals are too small, the resulting velocity becomes very noisy and difficult to interpret. On the other hand, if the time intervals are too large, the velocities are delayed and do not align properly with PD. We will analyze each task individually.

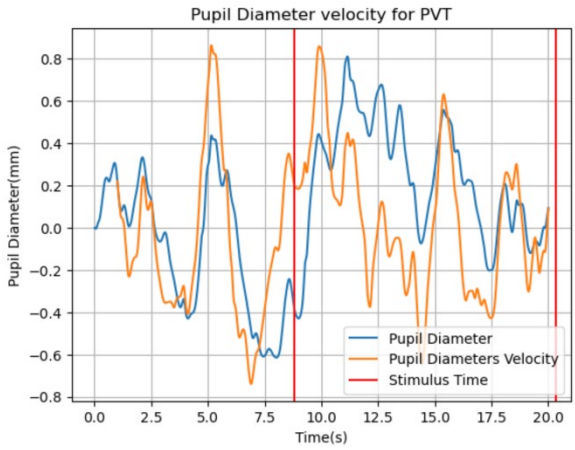




For the VWM task and DPT task, we observed a constriction in PD after the start of ST. The velocity amplifies this decrease. When the derivative experiences a significant drop, there is a high probability that ST occurred near the drop.



Compared to the other three tasks, the MA task is the longest. As a result, ST occurs less frequently. Although we do observe a dramatic drop in velocity, due to the long-time intervals between each ST, we found numerous instances of false positive signals in the velocity data. Therefore, not every dramatic drop in velocity indicates the presence of ST, but ST typically follows a significant drop in velocity.



In the PVT task, we noticed that some samples exhibit a drop in velocity around ST, but the majority of samples do not show any drop around ST. However, this can be understood when considering the analysis from section 4.3. In comparison to the other three tasks, we did not observe an initial constriction phase in PVT.

5. Machine Learning Experiment.

5.1. Method

5.1.1. Sampling

Our tasks will be identified as binary classification tasks, with three samples taken after each ST, each containing one second of data. The first sample, from 0 to 1 second after the ST, will be labeled as "1", while the second and third samples, from 1 to 2 seconds and 2 to 3 seconds after the ST, respectively, will be labeled as "0". This approach should result in an unbalanced data set with a 1:2 ratio. We will drop missing data samples without interpolation. More data was lost in the "0" samples compared to the "1" samples, resulting in a higher number of "0" samples dropped than "1" samples. This caused the ratio to become more balanced, with approximately 43% of the remaining samples labeled as "1" and 57% labeled as "0".

We will perform 5 runs similar to section 4.3, with the first four runs trained on each task individually (DPT, MA, PVT, and VWM) and the last run ("All tasks") trained on data from all 4 tasks. The number of samples for each task is detailed in the table below:

Task Type	Number of Sample		
	1	0	Total
DPT	28,919	65,671	94,590
MA	7,260	19,515	26,775
PVT	14,368	31,269	45,637
VWM	27,792	67,238	95,030
All Tasks	78,339	183,693	262,032

Our model will use 3 features: pupil diameter (PD) and the two dimensions of pupil position (Pupil X and Pupil Y). With a sample rate of 250 Hz and each sample containing 1 second of data, we will have 250 data points for each sample. Therefore, the tensor shape for one sample will be (250, 3).

5.1.2. Cross Validation and Metric Evaluation

To ensure the reliability and consistency of our result, we will apply the cross-validation method. We will split our group of 58 subjects into 5 sets as follows:

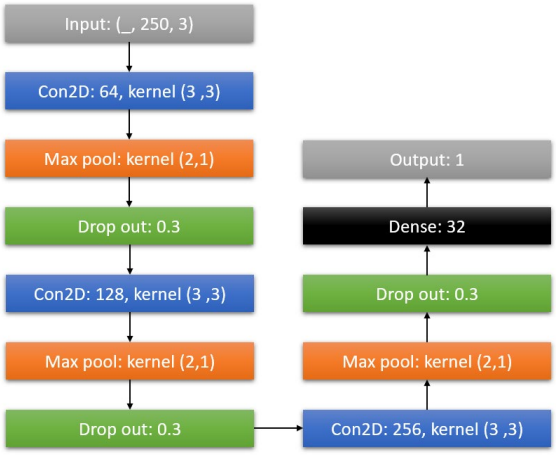
- Set 1: subject 0 to 11.
- Set 2: subject 12 to 23.
- Set 3: subject 24 to 35.
- Set 4: subject 36 to 46.
- Set 5: subject 47 to 57.

We will perform 5 runs for each task, with 4 sets used for training and 1 set used for testing each time. The final results will be the average of the 5 runs.

To evaluate our results, we will use Matthew's correlation coefficient (MCC) as our primary metric, and three secondary metrics: AUC score, F1 score, and accuracy.

5.1.3. Model

We will employ a straightforward Convolutional Neural Network (CNN) model consisting of 3 CNN layers with a 3x3 kernel filter. Following each CNN layer, we will append a Max Pool layer with a 2x1 kernel filter and apply a dropout technique with a 0.3 ratio to regularize the model. At the end of the model, we will attach a Dense layer of 32 and an output. A graphical representation of our model is shown below:



The model will use the sigmoid activation function to output the probability between 1 and 0.

5.2. Result

We have summarized our results in the table below:

Task Type	Metrics			
	MCC	AUC	F1	Accuracy
DPT	0.78	0.94	0.87	0.90
MA	0.58	0.85	0.70	0.81
PVT	0.48	0.80	0.66	0.76
VWM	0.09	0.57	0.38	0.59
All Tasks	0.38	0.74	0.54	0.72

The DPT task had a clear constriction of PD between 0 and 1 second interval, which is why the model performed best on DPT tasks. The MA tasks also had a small constriction, but it was not as clear as in DPT tasks, resulting in the model performing second best in MA tasks.

In the case of PVT tasks, the signals were mixed. Although it did not show in the average plot, we analyzed PVT for each individual subject and found that in some samples, it had an initial constriction followed by dilation, while in some other samples, it only had dilation in PD. It was a mix signal, and therefore the model performed average on PVT tasks.

However, the model performed poorly on VWM tasks, with performance being very close to a random guess. It is understandable if we look at the average plot in section 4.3, where the model can only see from 0 to 1 second because each sample is 1 second. There is not sufficient information within one second interval. As VWM has a very slow dilation and does not peak until around 2.5 seconds, the model could not find any pattern, resulting in poor performance.

Our ultimate goal is to auto detect cognitive load. Therefore, it is important to detect ST generally for all tasks, not just individual tasks. This is why we trained all the data together. Unfortunately, the model did not perform well generally compared to DPT, MA, and PVT tasks. However, the good news is that the model did not make a random guess like in VWM tasks. With an MCC score of 0.38, the model's performance is acceptable for a start. As we continue to improve the model, we can expect it to perform better in the future.

6. Limitation

The current version of our model is not yet suitable for real-world applications as it can only detect the ST at the beginning of the sample. To improve its usability, we need to train the model to detect ST at any point within the sample. Furthermore, we only sampled the first 3 seconds after the stimulus, which means the model did not analyze the remaining data. This could potentially limit the model's ability to accurately detect cognitive load in the real world.

Due to time constraints, we were unable to utilize some potentially useful data, such as subjects' response times and feed times. We were unable to apply the PDs' velocity in section 4.4 into the model. Incorporating this data into our model could provide valuable information for learning.

Additionally, we dropped samples that had missing data, even if they only lacked one or two data points. Interpolating these missing points could save some of these samples.

We also discovered some noisy and outlier data points, which could be removed to make the model more robust.

7. Next Step

The next crucial step is to eliminate noise and outliers from the data. We intend to use the Local Outlier Factor algorithm to detect the outliers and calculate the moving average to reduce noise in the data.

Moreover, we plan to extract valuable information that can be used as features. We intend to use the velocity of PD as an input feature. Additionally, the peaks and latency of PD after ST can also be utilized as useful information and features.

To enhance our model, we can explore newer robust architectures such as various variations of Transformer architecture. We also contemplate using pre-trained models like GPT or BERT.

Once the model delivers good results in classification, we aim to generate samples in every 0.1 seconds to simulate real-world application.

8. Conclusion

The study found that the model performed best on DPT tasks, followed by MA tasks, and performed average on PVT tasks. However, the model did poorly on VWM tasks, which require slow dilation, resulting in poor performance. When trained on data from all four tasks, the model did not perform well overall compared to individual tasks, but it did not make random guesses like in VWM tasks.

During the sampling process, we did not remove noise, outliers, or extract features from the data. If these steps were taken, the model's performance could be further improved. The model has the potential to be developed further and improve its ability to detect cognitive load.

9. System Requirements and Code.

Hardware Requirements:

The machine learning (ML) code requires 40-50 GB of memory and one or a few powerful GPUs. Our testing was conducted on UMBC's Ada cluster, which is equipped with 5 Quadro RTX 6000 GPUs and 200 GB of RAM. When working with other visualization codes, it is necessary to have approximately 10 GB of memory to load the entire dataset.

Software Requirements:

The ML code requires TensorFlow version 2.4 or later and has been successfully executed on the Anaconda virtual environment. All additional package dependencies are included in the 'requirement.txt' file.

Data Details:

The 'eye_data.csv' file contains information such as 'Pupil Diameter', 'Pupil X', and 'Pupil Y' for all 58 subjects. On the other hand, the 'behavior_data.csv' file records various events that occurred during the experiment, including 'Stimulus Time', 'Response Time', and 'isCorrectResponse'. The columns 'Task Type', 'SID', and 'Trial' will correspond between the two files. Both the 'Time' column in 'eye_data.csv' and the 'Stimulus Time' column in 'behavior_data.csv' refer to the same timeline. The 'Respond Time' column indicates the time elapsed after the corresponding 'Stimulus Time'. The 'Time' and 'Stimulus Time' values are measured in seconds, while the 'Respond Time' values are measured in milliseconds. The 'isCorrectResponse' column indicates whether the subjects responded correctly or not.

Code Details:

The 'model_build.ipynb' file contains all the code related to the machine learning experiments, including data preprocessing, CNN model building, training, and result evaluation.

The 'PD_plot_create.ipynb' file calculates the average Pupil Diameter across all 58 subjects and generates corresponding plots.

The ‘PD_plot_derivative.ipynb’ file will calculate and plot the Pupil Diameter’s velocity for all 4 tasks.

Acknowledgement

We extend our gratitude to Phil Beach, Mario Mendoza, Hannah Erro, and Zoe Rathbun for their invaluable contribution in generating data and coordinating the study. Additionally, we would like to express our appreciation to all 58 volunteer subjects at the University of California, Santa Barbara, for their participation. We have a special thanks to Professor Justin Brooks for providing us with this dataset. We also acknowledge the contributions of Murat Kucukosmanoglu and Steven Thurman for their meticulous review, analysis, and guidance.

References

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. Cognitive Science, 12(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4.

Paas, F., & Van Merriënboer, J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. Educational Psychology Review, 6(4), 351-371. https://doi.org/10.1007/BF02213420.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching.

Educational Psychologist, 41(2), 75-86. https://doi.org/10.1207/s15326985ep4102_1.

Arnsten, A. F. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. Nature Reviews Neuroscience, 10(6), 410-422. https://doi.org/10.1038/nrn2648.

Cohen Hoffing RA, Lauharatanahirun N, Forster DE, Garcia JO, Vettel JM, Thurman SM (2020) Dissociable mappings of tonic and phasic pupillary features onto cognitive processes involved in mental arithmetic. PLoS ONE 15(3): e0230517. https://doi.org/10.1371/journal.pone.0230517

Yu, X., Zhang, J., Xie, D., Wang, J., & Zhang, C. (2009). Relationship between scalp potential and autonomic nervous activity during a mental arithmetic task. Autonomic Neuroscience, 146(1-2), 81-86. https://doi.org/10.1016/j.autneu.2008.12.005

Sabatino DiCriscio, A., Hu, Y., & Troiani, V. (2018). Task-induced pupil response and visual perception in adults. PloS one, 13(12), e0209556. https://doi.org/10.1371/journal.pone.0209556