Ett realtidsmått på kognitiv belastning

Online ocular measurement of cognitive load

Erica Jostrup

Handledare/Supervisor(s)

Christian Balkenius

KOGM20

2017-06-09

Masteruppsats i kognitionsvetenskap
Avdelningen för kognitionsvetenskap
Filosofiska institutionen
Lunds universitet

Master's thesis (2 years) in Cognitive Science
Lund University Cognitive Science
Department of Philosophy
Lund University

Online ocular measurement of cognitive load

Erica Jostrup

Five different ocular measurements were investigated as a means of predicting cognitive load during a visuospatial memory game. Eye data was collected from the participants during a non-related n-back task, trained on both classification and regression models and investigated in relation to their ability to make predictions regarding how the participants were performing on an n-back task. All the algorithms performed well in their ability to make predictions on the data and a second task was created to examine how stable the predictions were over different tasks. During the second task a number of participants from the first part completed a visuospatial memory game while information about their ocular reactions were collected. The information was continuously sent to the learning algorithms previously investigated, making predictions about the participants' performance online. Based on the predictions, changes were made in the level of difficulty of the game. The results of the study show that it is possible to use a combination of several different ocular measures, collected from several participants during one task, to predict performance of individual agents during a second task. The tasks in this study were not related and the data used for prediction were not individually adapted, in contrast to previous studies.

Keywords: Cognitive/Mental workload, Fixations, Blinks, Pupillometry, N-back

1 Introduction

The pace at which new technology is introduced into our lives today does put a remarkable load on our human minds, in order to keep up. Today, many technological applications are on the verge of being connected to the internet. Internet of Things (IoT) is estimated to include 50 billion items in 2020 (Evans, 2011) and this new technology demands new ways of interacting. An interaction where information about the state of mind of the user is gathered, and the feedback given to them is adapted accordingly, to lessen the cognitive demand of the user.

There have been many approaches in trying to find a way to study and read the human mind. Methods such as EEG, CT and fMRI has been used as a tool to get new insight into the workings of the brain, but none of them are able to give us complete insight and many of them are both sensitive to their surroundings and/or space contingent. This brings forward the need for a new way of getting insight into the workings of our minds, in order to deal with the overwhelming amount of information that will be seeking our attention in the near future.

To create systems that are able to predict cognitive load have been proven difficult so far, and "online" measurements are required to create truly adaptive systems (Van Orden, Limbert, Makeig & Jung, 2001). Online measurements can be described as the kind of situated systems able to adapt to both implicit and explicit information given from the user and its surroundings, in real time. In contrast, today's offline systems are only able to interpret and create changes in their adaptation process after they have received and analyzed information explicitly given from the user, or someone else. The combination of information from both external and internal measurements can, in the future, serve as a complete ground for identifying, evaluating and regulating cognitive demand posed on the agent (Pederson, Janlert & Surie, 2011).

Information regarding the environment can, in many cases, be measured by sensors and information about the agent activity can be assessed through system interactions, while information about the agent state is harder to obtain. To assess an agents' workload, online, there are many aspects to take under consideration. Initially, there are three distinct measures of workload; subjective, performance and physiological measures (O'Donnell & Eggemeier, 1986). Subjective measurements reflect the individuals own assessment of the workload, while performance measurements reveal how well an individual is performing on the given task. Physiological responses may include changes in heart rate and skin resistance, amongst others (Kahneman, Tursky, Shapiro & Crider, 1969). Physiological responses are the most appealing way of measuring workload online, since these measurements can be made without the need of any input from the agent or any assessment of performance. The drawback is that these kinds of measurements often rely on electrodes placed on the body, connected to wires and other equipment that easily get in the way of the activity performed by the agent. Thus, a non-obtrusive and non-distracting technique for assessing workload through physiological measurements, in real-time, is needed in the future (Marquart, Cabrall & de Winter, 2015). This to be able to meet the demands of the technological development.

By using information from the eyes, as a means of input, showing our level of mental processing and fatigue as output, it is possible to collect the above mentioned online information and create a situated and adaptive system without disturbing the agent.

Ocular measures provide information in a high temporal resolution, better than any response time and accuracy measures do, without being intrusive or technologically demanding such as previously mentioned equipment. This makes it possible to measure how people respond to changes in task demand in real time (Eckstein, Guerra-Carrillo, Singley

& Bunge, 2016). The information can be used as a tool to evaluate when a person is in a high cognitive load state and, by that, make the conclusion that adding more stimuli or information will impair their performance at the given point in time.

Previous studies have found several ocular measures that correlate with cognitive demand imposed by a task. These are; blink rate, duration of blinks, frequency and length of fixations and pupil size.

Blinks have been studied, many times in combination with other physiological measurements such as heart rate and respiration, during tasks designed primarily in either air traffic control situations, car driving or flight simulations (Brookings, Wilson & Swain, 1996; Marquart et al., 2015; Veltman & Gaillard, 1998). In one flight simulation experiment, 12 pilots had to navigate through a tunnel while simultaneously conducting a memory task. The difficulty level of the tunnel was matched with the difficulty level of the memory task, creating continuously increasing demand on the pilot. As more visual information had to be processed by the pilot, a decrease in blink frequency were seen, along with decreased blink duration. Although, in relation to the increasing difficulty in memory task, blink frequency increased (Veltman & Gaillard, 1998). This indicates that there are several factors contributing in regulating blink frequency, but in relation to cognitive demand it is generally seen to increase (Marquart et al., 2015).

Fixations, including their frequency of and duration, have been found to increase with increasing demand created by a task. This have been found in experiments including both driving and completing mental tasks simultaneously (Recarte & Nunes, 2000), and during human computer interactions with varying degree of system autonomy levels (Evans & Fendley, 2017).

Pupil dilation is the most studied ocular measurement of the ones used in this study. Changes in dilation can be used as a measure of task engagement, physical arousal, attention, mental effort and allocation of cognitive control (Kahneman & Beatty, 1966; Johnson et al., 2014; Holmqvist et al., 2011; Naber, Alvarez & Nakayama, 2013). Studies investigating the relationship between short-term memory and pupil size, using a digit span task, have seen a steady increase in pupil size in relation to the number of items to be remembered. The dilation keeps on growing as long as the number of items to be remembered increases, until the participant reaches its' maximum load (Kahneman & Beatty, 1966; Johnson et al., 2014). Changes in pupil diameter up to one millimeter can be detected due to mental workload (Beatty & Lucero-Wagoner, 2000) and after reaching their maximum dilation, when the Working Memory (WM) load is at its highest, pupil dilation seizes. It then either continues in the dilated state, as long as the items to be remembered are still contained in memory, or start to shrink again, after the items to be remembered are reported. As long as the agent is engaged in the task and continues to store information in WM, their pupil dilation remains in its dilated state (Granholm, Asarnow, Sarkin & Dykes, 1996; Johnson et al., 2014).

Pupil dilation has been shown to reflect variations in processing load between very different cognitive tasks. This means that the measurement is a basic physiological one, stable over different tasks, and the dilation values can be directly compared between different experiments in different labs, according to Beatty (1982), which makes the measurement very useful in many applications.

Although there has been many different definitions of which cognitive processes are measured during the tasks mentioned above, it is clear that all the ocular measurements do reflect changes in level of attention and intensity directed towards the task. It thereby does reflect different aspects of cognitive processing. Given the broad background in studying cognitive load, in this paper, cognitive load is more generally defined as the amount of mental effort used by an agent, imposed by a task.

All of the above mentioned ocular measurements might be useful tools in our seeking of a way of measuring what is going on inside the human mind in real time. The combination of several eye-measurements does not only exclude some of the risks of only using a single, or a few, source(s) of input to analyze. Marquart and colleagues (2015) also conclude that focus of future research should be in combining different eye-measurements to create a robust and valid assessment method.

Eckstein and colleagues (2016) state that using multiple eye-measures during the same task creates a great opportunity for studying their relation to each other and how they can be used to assess cognitive processes. The aim of this study is to investigate precisely that.

The study examines whether the combination of above mentioned ocular measurements can be used as a method for measuring an individuals' cognitive workload, in real-time. The goal of this study is to examine whether it is possible to create an online ocular information measurement that is able to make predictions regarding the level of demand posed on an agent.

Differing from previous studies, data is collected from several participants during one task and all the data is used for prediction in a second task. In relation to these predictions, changes in the demand-creating stimuli can be made depending, to keep the agent in a continuously challenging level of demand. The study thus addresses both the possibility of using more general data for prediction and the stability of this measurement over different tasks.

2 Experiment 1

The experiment consisted of two parts. The first part, described in this section, was used to collect information about the participants' ocular behavior during a task created to maximize their WM load. The information collected was then used as basis for the second part of the experiment and to evaluate

the performance of different learning algorithms in trying to predict participant mental demand in relation to the stimuli.

2.1 Method

Participants

23 participants (12 men and 11 women) having normal (color) or corrected to normal vision participated in the study. All participants that had been informed of the study and undertaken a written consent. They were asked to use glasses in case they had to correct their vision, which was recommended by Eckstein and colleagues (2016) in relation to studying blinks. While reading, and signing the given consent before joining the study, the participants were asked not to participate if they were under the influence of any drugs. They were asked whether they had taken any caffeinated drinks or other types of caffeinated stimulants, in such case this was noted. One of the participants were excluded from the analysis, since the eye tracker did not record properly.

Task description

To assess the relationship between eye responses and cognitive load, measured as the level of demand on WM capacity, an n-back task with visual stimuli was used. The stimuli consisted of 3D block images with different shapes (Fig. 1). The 3D block images in this experiment have been chosen because they are unrelated to any emotion eliciting or arousing stimuli, which could affect the eye responses and thereby the results of the study. A similar experimental stimulus has been used by Lamp and colleagues (2016) in comparing neural mechanisms during fMRI, when either maintaining or maintaining and rotating 3D block images in a mental rotation n-back task. In the current experiment, no mental rotation was included in the task.

Initially, all participants completed a test-round, getting acquainted with the task at hand, consisting of a 0-, 1-, 2- and 3-back task. After completing the test-round, the participants were announced that the real experiment was to begin and the eye tracker to be calibrated before the data collection started. At the beginning of the experiment, a baseline measurement of the participants' eye-responses was made. This was done in a similar way as by Léon-Domíngues and colleagues (2015), using a 0-back task which creates no memory load. The target stimuli, which was chosen as one, out of 20 stimuli images, was shown during the instructions of the 0-back task. The participants were instructed to respond to the chosen image as matching, while all other stimulus pictures should be responded to as mismatching. This was done by pressing 'left key' for yes (matching) and 'right key' for no (mismatching).

A sound was played immediately after each response, indicating whether the participant's answer was correct or incorrect. Since the experiment leader was present in the room during the experiment, the sound was audible for both the participant and the experiment leader. The sounds served as a marking of how well the participant was performing during the task, and since most people do not want to fail in front of someone else, the sound also served as a way of keeping the participants trying their best throughout the task.

The participants completed a series of visual n-back tasks, consisting of the same 20 3D block images as in the 0-back task. Task workload was modified by the number of n-backs needed to be remembered in the n-back task. In the 1-back condition the participants had to respond whether the current stimulus was identical to the one previously shown, while in the 2-back task participants were required to respond whether the stimulus currently shown was identical to the one shown two trials earlier, and so on. In relation to the continuing increase in the number of stimuli needed to be remembered by the participant, as the n-back increased, the workload of the task increased.

The memory load produced by the n-back task can be compared to the load created by, for example, the WISC forward digit span used by Johnson and colleagues (2014), when studying the relation between Task Evoked Pupillary Response (TEPR) and Short Term Memory (STM) capacity. In their task, as in this one, there is a continuing increase of items needed to be held in memory, which should produce similar effects on ocular measures between studies.

The stimuli sequences were presented in blocks, one for each n-back. The level of the n-back series the participant performed was announced at the beginning of each block. Each block consisted of 40 trials, where half of the targets were matching and half mismatching. The stimulus sequence was randomized for each block and each participant and the stimuli presented was randomly chosen in each trial, out of the 20 different images. The stimuli was shown for 750 ms each or until the participant identified it as either matching or mismatching with the image n-backs earlier. After the stimuli was presented, a scene, scrambled to create the same luminance as the stimulus picture, was shown for 500 ms before the next stimulus picture was presented. To create a maximal load on participants WM, the value for n kept increasing until they completed the 5-back version.

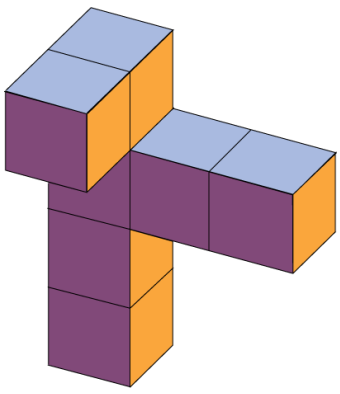


Figure 1. Example of 3D block stimulus. 20 different stimulus-pictures were used during the task.

Both the stimuli (3D-block images) and the scrambled slides shown between the stimulus pictures were comparable in brightness and contrast to avoid any changes in pupil size due to variations in light intensity (Holmqvist et al., 2011). It was also of great importance that the stimuli pictures were placed in the center of the screen, due to the sensitivity of pupil measurement when the gaze is directed to the edges of the screen (Brisson et al., 2013).

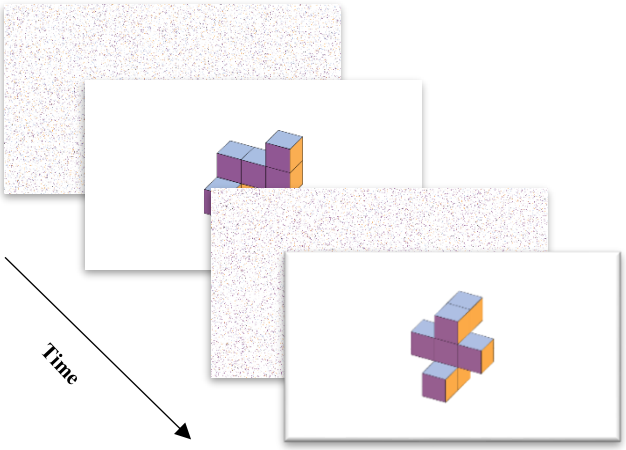


Figure 2. The presentation of the n-back sequence. 40 3D block images were randomly chosen from a list of 20 different pictures and shown for 750 ms, or until a key was pressed by the participant. Between every 3D image a scrambled stimuli-image was shown for 500ms.

The entire session lasted about 20 minutes, including training, setup and calibration of the eye-tracking equipment. The participants were asked to do their best during the experiment, to ensure task engagement was as big as possible, and they were instructed to answer as quickly and accurately as possible.

After the experiment was accomplished, the participants were asked whether they gave up during the task and if so, at what point. All participants were then debriefed, thanked for their participation and asked to return another day, to participate in the second experiment.

Eye tracking

Participants were placed in a head- and chinrest in a windowless room approximately 80 cm in front of a computer screen. The usage of a chin- and headrest was of great importance, because of the sensitivity in pupil-measurements due to movement. If the head is moved during tracking, either closer or further away from the eye tracker, the pupil measurement will be impaired. In the same way, if the gaze is directed to the edges of the screen, the pupil-measurement will differ from what is measured if the gaze is directed to the center of the screen. This underlines the importance of a centered stimuli during the task.

The participants initially completed a trial consisting of a 0, 1, 2 and 3-back task to make sure they understood the span of the task. After completion of the test-round, the participants were calibrated on the Eyelink 1000 Plus eye tracker (The Eyelink 1000 plus, 2017), mounted on the desktop. After the calibration, the 0-back task served as a way to measure the baseline eye-responses of the participants. This measurement was used to evaluate the size of the dilation, changes in fixation duration, fixation rate, blink rate and time, of every participant during the experiment. Because of the sensitivity of the pupil size measurement it was of great importance that all experiments were done in a room with constant luminance, including a similar luminance in stimuli throughout the trials. Eckstein and colleagues (2016) recommended performing the experiments in a moderately lit room, because if the room is too dark the cognitively-evoked pupil dilation is smaller than in moderate light.

The task-evoked pupillary response (TEPR), blink rate, blink duration, fixation frequency and duration was recorded, sampled at 1000 Hz. To be able to get a reliable estimation of the participants' workload, multiple TEPRs had to be recoded and averaged across participants, why the many trials and blocks were necessary.

Data Examination, Reduction & Programming

The n-back task was created in python, using PsychoPy2 (http://www.psychopy.org) to present it.

Blink duration, blink frequency, fixation duration, fixation frequency and average pupil size (area) was extracted from the data collected by the Eyelink 1000 eye tracker. The data was studied in relation to the level of the n-back task and segregated according to trials and stimulus onset (Society for Psychophysiological Research, 2011).

The data was organized and presented in three different ways, to evaluate the best way of data presentation for the machine learning algorithms. A Z-score cutoff, including datapoints with scores within the span of plus/minus 3.5, was evaluated as one of the datasets. The complete raw data was used in another dataset and the percentage of change in the data from the 0-back in the third dataset. In the "percentage of change"-dataset the 0-back was set as the null-limit and the 1-5-back task was calculated as change from the null-limit, presented in percentage. Blinks lasting longer than 1000ms was always counted as artifacts and thereby excluded from all the datasets.

The mean blink and fixation duration was calculated for each level of the n-back task. Blink- and fixation frequency was calculated by dividing the number of instances, during each level of the n-back, with the time (in ms) spent on each level. This resulted in a unit per millisecond, as described by Holmqvist and colleagues (2011). Mean pupil size during each n-back was measured as the area of the pupil. For the dataset presenting percentages of change from the 0-back task, Task Evoked Pupillary Response (TEPR) was calculated in change from the baseline measure of pupil size, which represented 0.

The error percentage of each block was calculated, where 40 out of the 40 blocks were the maximum number of correct responses that could be given.

The blocks were also classified as either easy or hard for the participant, and were classified according to whether the participant had given up or not. The easy/hard classification was done evaluating the error percentage. Errors above or equal to 37,5% were classified as hard since this level of error was both found in n-backs of four and five (where most of the participants did report on given up on the task), but also in level three. For many of the participants the error score did not continue to increase in relation to the level of the n-back, but did instead show levels of errors between 37,5% and 45% in the last two n-back levels, which might equal to chance.

The given up/not given up classification was done evaluating the time spent on each block. If the time spent on a higher block was less than the previous one, it was defined as the participant had given up on the task. This was combined with verbal reports taken from the participants after their participation, which were very useful in cases where no changes in time on task could be found.

The different datasets described were examined in an explorative way, with several different Machine-Learning (ML) algorithms collected from the scikit-learn (http://scikit-learn.org) package for Python. Since all datapoints were labeled, all the algorithms evaluated are included in supervised learning (Mohammed, Khan & Bashier, 2016). Several different algorithms, with different strengths and weaknesses, were chosen to get a variety in the different algorithms evaluated. The different algorithms might find different relations in the different datasets, and thereby be more or less fitting in making predictions in this particular case. The algorithms chosen for exploration included both classification and regression models.

The number of samples included in the dataset created the limits for which algorithms to explore and some of the most common algorithms for data with these limitations were chosen. The evaluated algorithms were: Logistic Regression (LogReg), k-Nearest Neighbors (k-NN), a Fisher Linear Discriminant Analysis (LDA), Linear Regression (LinReg) and a Multi-Layered Perceptron (MLP) for both classification and regression.

Both classification and regression models used the default parameters set by scikit-learn for the models (Scikit-learn Linear Regression Model, n.d.; Scikit-learn Multi-layer Perceptron Regressor, n.d.), except for the selection of the solver for the Multi-Layered Perceptron for regression. A stochastic gradient descent was used for the MLP regressor in this case.

Data from six, randomly chosen, participants was used to make predictions from while the rest of the data was used as training data for the algorithms. When evaluating the algorithms in relation to the different datasets, a scaler was used to normalize the data, when presented either as raw data or when filtered with a Z-score cutoff.

The classification algorithms predicted if the task was easy or hard, or whether the participant had given up on the task, while the regression algorithms predicted the percentage of error on each block for each participant.

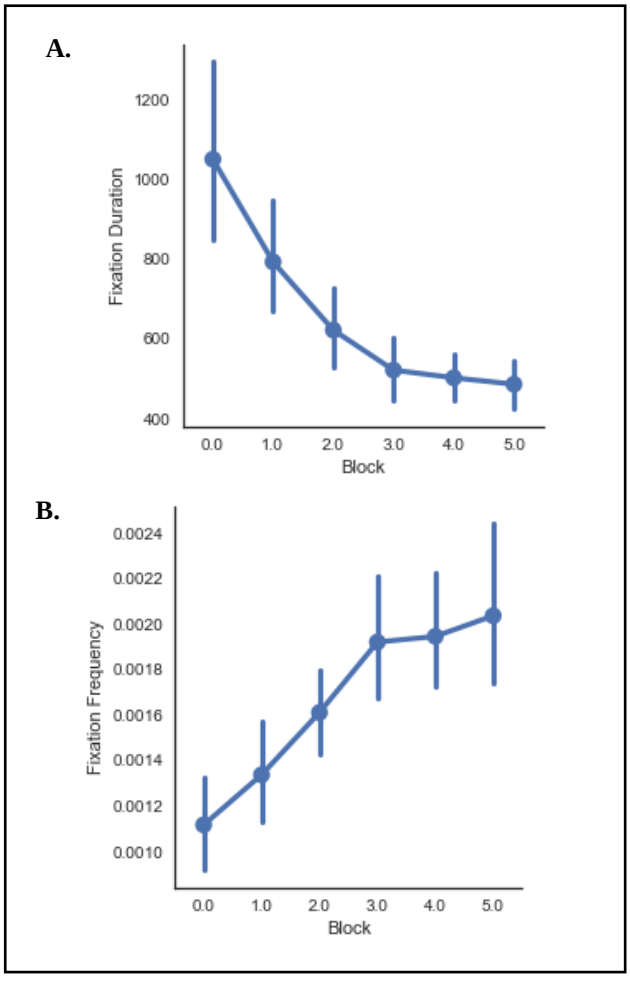
2.2 Result

Ocular measurements

When inspecting the behavior of the five ocular measurements that has been collected during the experiment, one by one, it can be seen that the measurements does differ quite a bit in their relation to the task (Fig 3A-F). Fixation duration, fixation frequency, blink frequency and average pupil size (APS) were all calculated as a mean value across each n-back, for each participant. For blink duration, both mean and median was calculated (Fig. 3C-D), because there were a lot of variance in the max and min blink duration measurement (Aron, Coups & Aron, 2013), both within participants and between. The different calculation methods (mean/median) thereby gave quite different results.

The two ocular measurements that gave the most salient results was the decreasing fixation duration $F(5, 126) = 11.7$, $p < 0.01$ (M = 661.6, SD = 360.5), Fig. 3A, and the increasing fixation frequency $F(5, 126) = 7.7$, $p < 0.01$ (M = 0.0017, SD = 0.00071), Fig. 3B, in relation to the increasing n-back.

No significance was found, either in blink duration mean (M = 122.1, SD = 38.6), blink duration median (M = 112.6, SD = 34.7), blink frequency (M = 0.0003, SD = 0.00019) or average pupil size (APS) (M = 904.6, SD = 247).



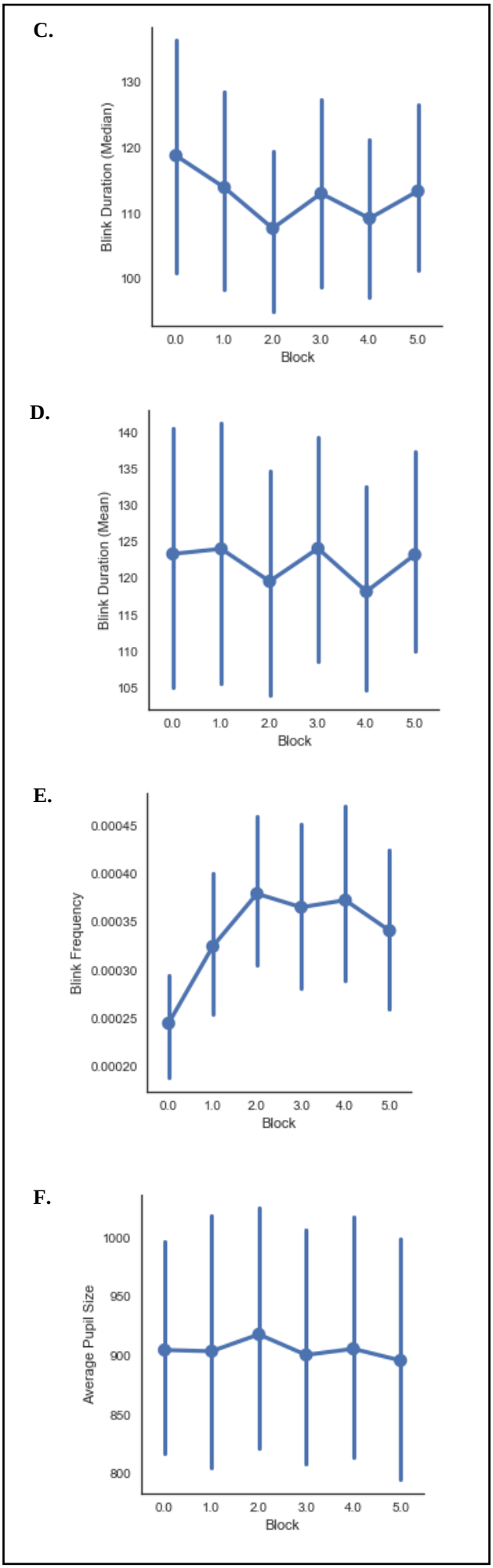


Figure 3. A-F. A. Fixation duration in relation to n-back. B. Fixation frequency in relation to n-back. C. Mean blink duration in relation to n-back. D. Median blink duration in relation to n-back. E. Blink frequency in relation to n-back. F. Pupil area in relation to n-back. Data from all participants has been used in all graphs.

There were some tendencies for an increase in blink frequency for the first three blocks (0, 1 and 2) in the n-back task (Fig. 3E), and if only observing the first three blocks there was a significant increase in blink frequency for all the participants $F(2, 63) = 3.5$, $p = 0.035$. The same tendencies for a decrease in blink duration could be seen in the first three blocks in median blink duration (Fig. 3C), but no significance was reached.

Data examination

The dataset containing raw data, including as unfiltered data as possible (extreme outliers lasting longer than 1000ms excluded), gave the best predictions from the algorithms in general, with the highest correlations in classification (Table 1, 2). Figure 3 is presented using this dataset.

Table 1. Mean correlation coefficients from the six randomly chosen participants, when predicting if the task was easy or hard for the different datasets. The rest of the dataset was used to train the classification models.

	Raw data	**Z-cutoff**	**% Change**
k-NN	0.78	0.72	0.67
LDA	0.89	0.86	0.69
LogReg	0.92	0.89	0.69
MLP	0.78	0.72	0.64

Two exceptions were made from the best predictions seen in Table 1, where raw data created the highest correlations. In Table 2, the dataset presenting the data as percentage of change from the 0-back gave the best predictions for the k-NN and MLP models, in predicting whether the participant had given up or not.

Table 2. Mean correlation coefficients from the six randomly chosen participants, when predicting whether the participant had given up or not, for the different datasets. The rest of the dataset was used to train the classification models.

	Raw data	**Z-cutoff**	**% Change**
k-NN	0.64	0.58	0.75
LDA	0.78	0.78	0.58
LogReg	0.81	0.78	0.67
MLP	0.61	0.64	0.67

For the regression-models there were no major differences between the raw dataset and the one using a Z-cutoff, while the percentage of change dataset did not produce predictions

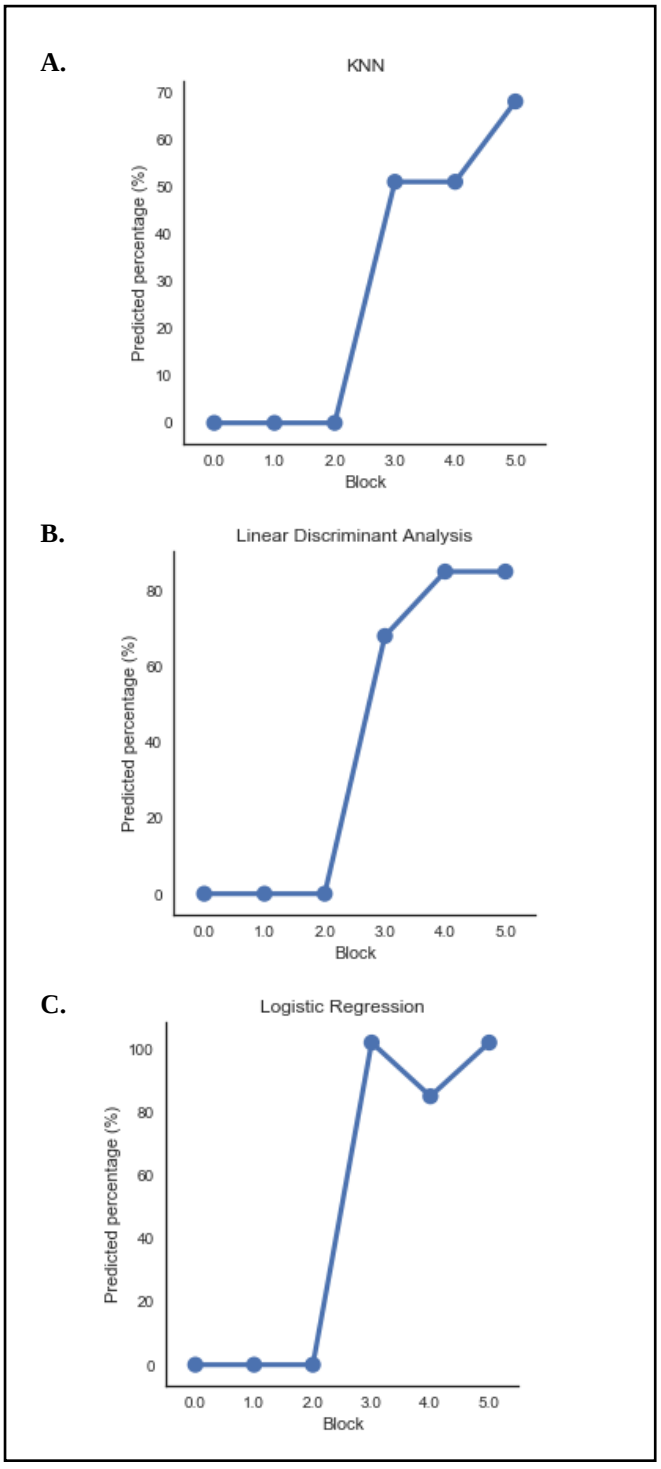
that correlated with the actual performance of the participants in any high degree (Table 3).

Table 3. Correlation coefficients from the six randomly chosen participants used for prediction, as a relation between model type and dataset. The rest of the dataset was used to train the regression models.

	Raw data	Z-cutoff	% Change
LinReg	0.47	0.45	0.21
MLP	0.5	0.5	0.14

Classification algorithms

The classification algorithms predicting whether the participant found the task easy or hard gave the best predictions (Table 1, Fig. 4A-D).



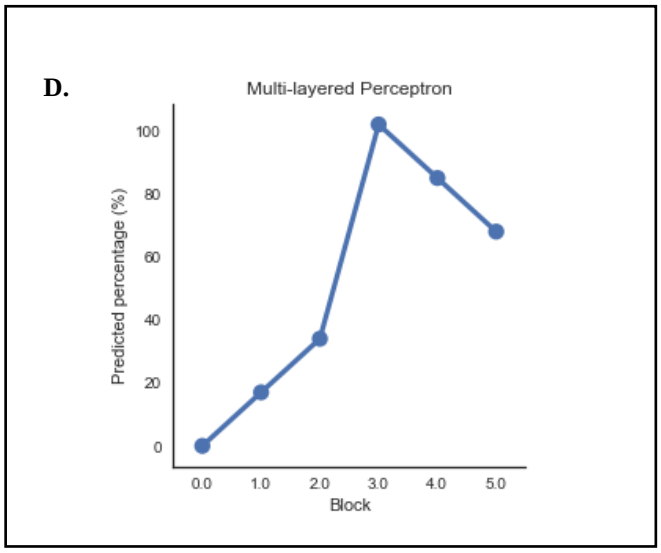


Figure 4. A-D. Percentage of participants (mean) that the classification algorithm has predicted the task to be hard for, in relation to the different blocks of the n-back task. Using raw data from six randomly chosen participants. A. K-Nearest Neighbors. B. Linear Discriminant Analysis. C. Logistic Regression. D. Multi-Layered Perceptron.

Regression algorithms

There were no major differences between the two regression models, which both gave predictions that were correct about half of the time (Table 3). As can be seen from Fig. 5, the regression algorithms error prediction increase in relation to the level of the n-back task, which clearly reflect the error production from the participants during the n-back task.

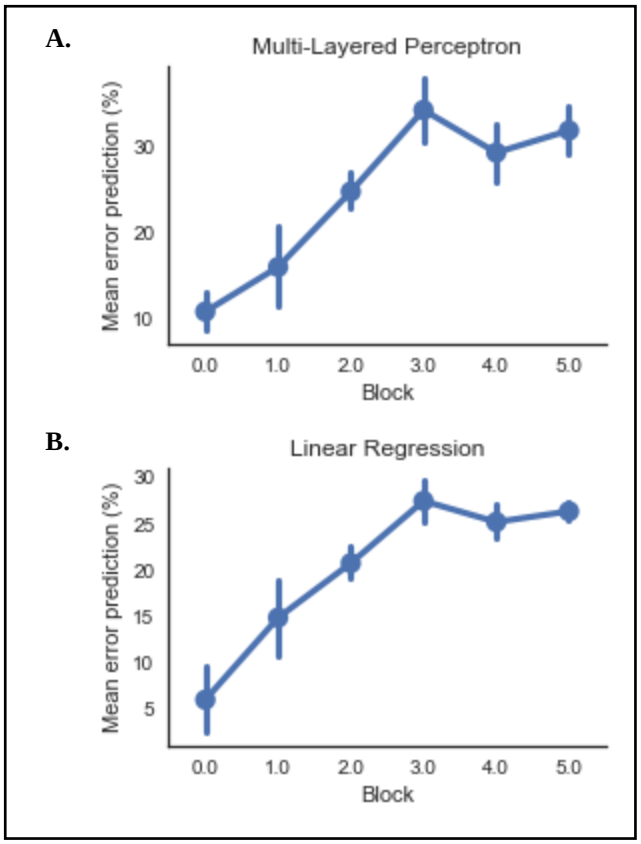


Figure 5. A-B. A. Mean error prediction of the Multi-Layered Perceptron Regression model. B. Mean error prediction of the Linear Regression model. Predictions made from six randomly chosen participants, using raw data.

The mean error prediction, and Standard Deviation (SD), of the regression models show that the models are very similar in their performance for the chosen dataset (Table 4). Although, the Linear Regression model does produce a smaller SD in the higher blocks.

Table 4. Mean (M) and Standard Deviation (SD) of the predicted errors of the Multi-Layered Perceptron (MLP) Regression model and the Linear Regression (LR) model.

Model	Block (n)	M	SD
MLP	0	9.7	4.0
MLP	1	15.6	6.3
MLP	2	24.3	3.4
MLP	3	35.0	5.7
MLP	4	29.2	5.6
MLP	5	32.2	5.1
LR	0	6.3	4.7
LR	1	14.8	5.6
LR	2	20.7	2.3
LR	3	27.4	3.2
LR	4	25.1	2.6
LR	5	26.3	1.3

Studying the contribution of each ocular measurement to the Linear Regression algorithm shows that the measurements with the highest contribution are fixation duration and fixation frequency (Table 5), independent of the usage of either mean blink duration or median blink duration in the model. In column 2 the mean blink duration was used for producing the coefficients and in column 3 the median blink duration was used in producing the coefficients. The two ocular measurements that contributed least to the algorithm was blink duration and blink frequency. The intercept was 20.05 for the model, independent of which blink duration measurement used.

Table 5. Coefficients contributing to the Linear Regression model, using either mean or median calculation of blink duration.

	Mean blink duration	Median blink duration
Blink duration	1.45	0.57
Blink frequency	-0.68	-0.85
Fixation duration	-3.68	-4.04
Fixation frequency	3.72	3.27
APS	-1.90	-2.10

As can be seen, the coefficients do change somewhat in their contribution to the algorithm given the way blink duration is presented. Although, no major difference in the error prediction given the way blink duration was calculated can be seen (Table 3).

2.3 Discussion

The goal of the task was to evaluate whether a combination of five different ocular measurements, collected from several participants during a task created to cover all levels of a participants' cognitive demand (from its lowest to its highest), can be used to predict the performance of another set of participants on the same task. Given the high correlations between the predictions and the results reached by the participants during the task, it is possible to use the studied ocular measurements to create this kind of predictions. The different ocular measurements are discussed in detail below.

There are significant changes in the fixation duration and fixation frequency in relation to the level of the n-back task. This shows that fixations can be used as a measure of cognitive demand created by a task. The fact is further supported as the measurements are the biggest contributors to the algorithms coefficients (Table 5). The steepness of the changes decreases as the demand of the task gets higher, which is coherent with the reports from most of the participants giving up in the last two n-back tasks and the decrease in time spent on the last two n-backs. Even though there are no such changes to be found in the pupillary data in this experiment, Beatty (1982) found that the slope of the pupillary response will vary in relation to both task difficulty and task length. It seems like the same variation can been seen in fixation data in this experiment.

As stated earlier, pupil changes have been shown to reflect changes in mental effort (Johnson et al., 2014) and processing load (Beatty, 1982), amongst others. The measurement has previously been shown to be a very stable one, over different tasks, which should make it comparable between experiments. The fact that there are no clear general findings in this case might be explained with another finding of Beatty (1982), that the slope of the pupil change may vary due to participant and stimuli. Changes in pupil diameter has been shown to be related to the level of intelligence of the participants. Smaller changes in pupil dilation during a number of cognitive tasks, such as digit span task and sentence comprehension, has been seen for individuals with higher scores on an intelligence test, than for participants with lower scores (Beatty, 1982). These findings reflect individual differences in the amount of cognitive effort needed to complete demanding tasks, and since the participants in the current experiment consist of mainly university students, this might have some effect on the results seen here.

When looking at individual cases of pupillary data there are as many instances where the pupil is at its largest at the start of the experiment and then continues to decrease during the following n-back tasks, as the other way around. An explanation could be that some of the participants may have been nervous, even though they had completed a test-round of the task. The nervousness might be due to the calibration and start of the eye-tracker, which were done just before the measurements in the 0-back task were collected. The fact that there are

such individual differences in the changes of the average pupil size (APS) of the participants may result in the evening out of data that is presented in Fig. 3F.

Even though there have not been many studies investigating blink duration, some very differing limitations of what to be counted as a blink has been made in those few cases. Van Orden and colleagues (2000) used a limit of 83.3ms, while Bonifacci and colleagues (2008) set their limit to 96ms and Geng and colleagues (2009) defined a blink to be at least 50ms, in combination with eye movements (saccades) of a certain limit. Initially during the data investigation, the data was sorted with the same limitations as Van Orden and colleagues (2000), but this left very little data remaining. There was a big variance in the blink duration data and excluding data-points with a value less than 83.3ms excluded as much as 95% of the blink-data from one of the participants, which did not seem logical. The fact that the less limitations set on the data, the better prediction, therefore, is very positive, and in the end a similar definition of a blink as Brouwer and colleagues (2005) is used, including all instances where no pupil were detected.

No clear continuity in the changes of the blink data throughout the task has been found. If concentrating on the changes in blink frequency during the first three blocks of the n-back task (0, 1 and 2) a significant change in blink frequency can be seen. This is in line with previous research (Marquart, Cabrall, & de Winter, 2015) stating that blink frequency has been seen to increase in relation to higher workload. The limitation of significance to the first three rounds of the task might depend on blinks being a much more sensitive measurement in relation to the onset of a difficult task.

Since the blink frequency measurement are not as stable as the fixation measurements, which kept on increasing/decreasing throughout the whole trial, the assumption that the blink frequency measurement reflect changes in the level of effort directed towards solving the task can be made. The fact that there is a large leap in the level of difficulty from a 2-back, where the participant should identify if every other image is matching or mismatching, to a 3-back, where you should identify if every third image is matching or mismatching might be reflected in the blink frequency measurement. The 3-back task was in most cases the highest level of the n-back task that the participants reported to have, actually, been trying to solve. But if that was not the case, this might reflect that they gave up earlier than reported. Since links have been seen to correlate with goal directed behavior (Eckstein et al., 2016), the fact that there are changes in only the three first blocks might very well reflect this fact. The ocular changes will be a good reflection of how much cognitive load the task is creating, but only as long as the participants try to maintain good performance throughout the task (Van Orden et al., 2001).

Even though no changes are seen in the blink duration measurements, the learning algorithms do not seem to differ much in their predictions due to the usage of either the mean blink duration or the median blink duration. A conclusion that blink duration does not affect the algorithm in that big extent can be made, which is further supported when studying the coefficients of the Linear Regression model (Table 5).

The fact that the use of raw data gives the best predictions for the algorithms makes the usage of ocular input, in relation to evaluating cognitive load, even more appealing. Not needing to make any big reductions, set limitations or any need of normalizing the data means that eye-data can be collected directly from the tracker, checked for extreme outliers, and then used by the regression algorithm to evaluate the state of the agent. The fewer steps needed between data collection and the regression analysis makes the process faster, with less computational power needed. This makes the possibility of using it in smaller devices, such as wearables and head-mounted displays (HMDs), easier.

The results from the experiment shows that both the classification algorithms and the regression algorithms does perform well in their predictions regarding how the participant is performing in the given task. This creates a stable ground for evaluating the algorithms performance in an unrelated task, to find out how stable these predictions are over different tasks and agents.

3 Experiment 2

Given the result from the first experiment, the investigated ocular measurements are a good source for gathering information and making predictions about a participants' performance during an n-back task. All the investigated algorithms perform well in their task and can be used in further investigations regarding how applicable the data collected during the n-back task is on predicting performance on an unrelated task.

To investigate whether the ocular data collected from the first experiment reflects a general behavior in the participants when their WM is challenged, a game was created for the second experiment. The goal was to investigate whether it was possible to predict a participants' cognitive level of demand in a completely different task than was used when collecting the data. This was done using a memory based game.

3.1 Method

Participants

The second experiment was performed by 5 of the participants included in the first experiment, approximately two and a half months after the first experiment.

Task description

The game consisted of a board with 25 tiles. At the start of every new round, a number of randomly chosen tiles changed color for 2 seconds and then back to the initial color again. The task was to remember which tiles that changed and select them, using the mouse.

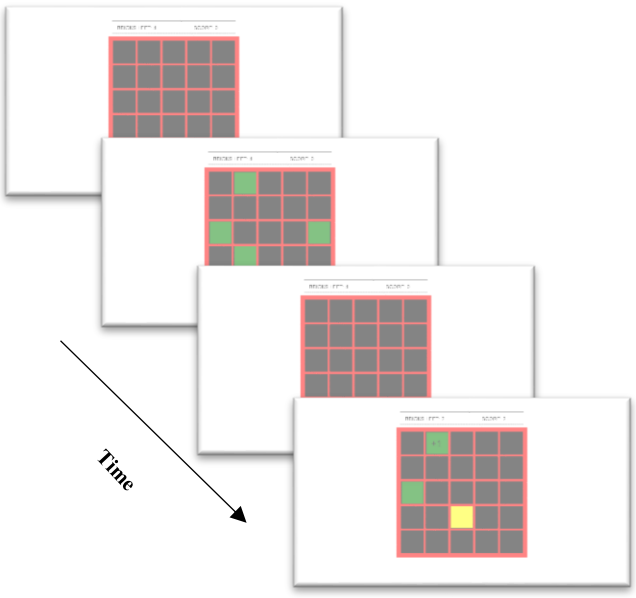


Figure 6. The presentation of the game. Initially a blank board is shown, after 2 seconds a number of tiles changes color for another 2 seconds. The board then changes back to its initial state and awaits participant input. The participant should mark which tiles previously changed color, using the mouse.

Initially the player had to remember 4 tiles. After the first three rounds had been played at that level, the game either changed level or continued at the same level as previously played. This means that the player either had to remember one tile more/less in the coming rounds, or continue to remember the same number of tiles. It was the information collected from the eye tracker, which has been sent to the algorithms for prediction on how the participant is perceiving the task, that was the fundament for the evaluation on what should happen in the game.

If the level of the game increased, the player had to play two rounds of the new level, before the ocular data from the three latest rounds were evaluated. If the level of the game decreased, the player only had to play one round on that level before their ocular data was re-analyzed. This to make sure the player never gave up, which might happen if they reach a level with too high demand, as seen on the n-back task.

The game had a score counter, to help the player keep up with their progress and keep the player motivated throughout the game. The game also had a "bricks left"-counter, to help the player remember how many tiles to press the present round, since the game is contingent on the player pressing all the number of tiles presented in the sequence.

When the participants had played the game for 10 minutes, which equals to approximately 25 rounds, it was terminated and they were asked to assess their experience through a NASA Task Load Index (TLX) assessment (NASA TLX: Task Load Index, 2017).

After completing all parts of the experiment, participants received a gift, thanking them for participating.

Eye tracking

The participants were placed in the same room as during the first experiment, at the same distance from the screen, in a head- and chinrest. The design of the game did consider the same facts about eye tracking as was done during the design of the n-back task. In other words, the stimuli of the game, in this case the board of the game, was centered on the screen and it was presented with constant luminance throughout the game. After completing a brief training-trial, a calibration and validation of the participants' eye-movements were done with the Eyelink 1000 eye tracker. During the game, the number of fixations, fixation duration, number of blinks, blink duration and APS of the participants were measured by the eye tracker. The measurements were collected for 8 seconds during the first round, when the task was to remember 4 tiles, and increased by 1 second every other round. One round of level 4 of the game took approximately 10 seconds to play, setting the limit for how long the data collection could continue. At the end of each round, the collected data was sent to the ensembled learning algorithms and a prediction was made on the data. Since the data that the algorithm had trained on was expressed in error percentage, this was the value predicted by the algorithm.

After the first three rounds of the game, where four bricks were to be remembered, all the predictions made from the algorithms, given the collected ocular measurements, were collected into a mean value. If the mean value of the three rounds was lower than 20 (error percentage) the level of the game increased by one tile. If the mean value of the three predictions was higher than 30 (error percentage), the level of the game decreased by one tile. This continuous evaluation kept on throughout the game, with the player playing the easy levels two times and the hard levels only one time, as described earlier.

Data Examination, Reduction & Programming

The game was created in python, using scikit-learn for the learning algorithms and PsychoPy2 to present it. The data collection online from the eye tracker was also gathered through python code, continuously communicating with the tracker.

During the time the participants played the game, there were several background activities going on. Even though the classification-models did perform very well in predicting whether the task was easy or hard for the participant, the usage of a regression model gives the smoothest way of evaluating the data in an online approach. A regression model can make finer distinctions and its limitations can be set as wished according to the task at hand, which is why the regression models were chosen to make predictions during the game.

Both of the evaluated regression algorithms were used to predict the percentage of error the participants would have created if they were performing the n-back task. The usage of more than one model for evaluation has been shown beneficial in that they can support each other in their weaknesses. By

ensemble learning it is possible to work around their individual problematics and create good generalization abilities (Liu & Yao, 1999).

The ensembled models went through a training phase, where they trained on all the data collected from the first part. During the time the participants played, there was continuous information collection from the eye tracker, as already described. The data collected from the eye tracker was compiled in the same way as the training-data initially used by the learning algorithms, to make sure the algorithms were able to interpret the new data. Before the collected data was sent to the algorithms for prediction, outliers lasting longer than 1000ms were removed, as was done with the data from the first experiment.

After every round of the game, the algorithms each made a prediction regarding how much error, in percentage, the player would have experienced if the task at hand was the n-back task. The mean of the predictions was calculated and sent to the game which assessed it in relation to the defined restrictions. The restrictions set to 20 error percent, as the lower limit, and 30 error percent, as the upper limit, were thus set to make sure the player was maintained within a limit that kept them experiencing the game as challenging, but not too hard, and at the same time keep them motivated not to give up. The 20% error prediction as the lower limit and 30% error prediction as the upper limit equals to a 2-back task and 3-back task in the n-back task. The 4- and 5-back were for many people in the n-back task too hard and they gave up on the task at hand, why these levels of demand were sought to be avoided.

3.2 Result

The game predicting the error percentage of the participants, given their ocular input collected online by the eye tracker, did not produce any significant results in their change between game levels. Four out of five participants reached a level where 15 tiles should be remembered, while the last one reached a level of 14 tiles to be remembered.

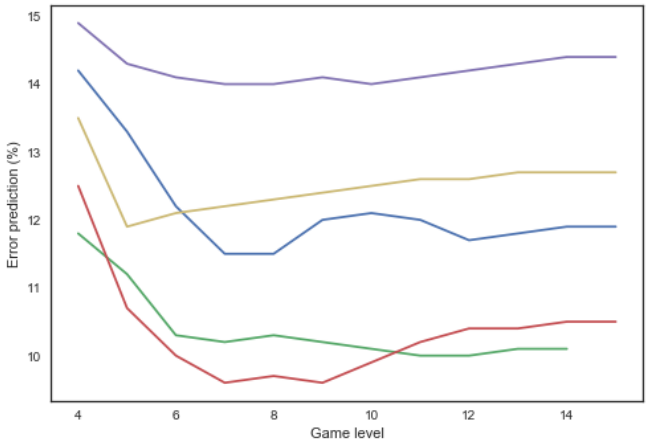


Figure 7. Error prediction in percentage for the five participants during the game. The error prediction is based on the percentage of error seen when showing similar ocular behaviors during the n-back task performed in experiment 1.

From figure 7, it can be seen that the error prediction was at its largest at the beginning of the game, declining somewhat in the following rounds and then making a minimal climb at the end of the game. Although, this initial decline in error prediction (between level 4 and 7) did not reach significance for more than two out of the five participants. As is also seen in figure 7, the error prediction did differ somewhat between participants, but all kept within the limits of 10-15 error percentage (M = 11.96, SD = 1.58).

The NASA TLX assessment did also indicate that the participants did not perceive their mental load during the game as particularly high. The mean value of the participants perceived mental load during the task was 10.4 out of 20 (SD = 4.2), which means that they neither found it very demanding or extremely easy. Their mean rated effort was 15.2 (SD = 2.4), which means that they all worked quite hard to accomplish their level of performance, which they rated as a mean value of 5.4 (SD = 4.5), out of 20. The performance rating means that all the participants felt that they were successful in accomplishing the task.

The general feedback given from the participants after accomplishing the task was that it was much more fun than the previous one (n-back) and was less demanding.

3.3 Discussion

The goal of the second experiment was to evaluate the ability to use the ocular data collected during the first experiment, to predict the level of cognitive load of the participants during a ten-minute game. The goal was reached in such as the algorithm produced predictions that did not significantly differ between the participants, and thereby shows stability over different agents performing the same task.

The limits of the game were set in relation to the errors found in the n-back task and the game was programmed as to change its level of demand according to how the predictions fitted with the limits. Since the game never produced error predictions higher than 15%, the lowest limit of 20% made the game level keep increasing throughout the whole game. This indicates that the cognitive demand created by the game did not affect the participants in any palpable way. One reason for this may be the game design. Since all the bricks were shown simultaneously to the player, they could use the afterimage, produced by the recovering photoreceptors after seeing the pattern made by the colored bricks, to help them remember all the bricks. Another method reported used by a participant was to use chunking as a means of remembering the location of the bricks.

As the NASA TLX assessment also concluded, the participants did not perceive the task especially demanding and a game modification would be needed to produce a higher load on the participants. For example, instead of showing all the

bricks to be remembered at the same time and then only asking the participants to repeat which bricks that changed color, in any order, the bricks could have been presented one by one and the participants could have been asked to repeat the pattern in the same order it was shown. This would avoid participants from using different memory techniques, as mentioned above, to help them remember the placement of the bricks, and put a higher demand on their working memory.

It should also be considered whether the limits were set too high. It might be that the participants did perceive a 3-back task as too demanding, as discussed previously in relation to blink frequency, and that the limits should have been set lower.

Even though the participants did not reach either the 20 or 30 error-percentage limits set of the game, this does not reflect any disabilities in the algorithm or the data collected from the first part of the experiment. Since the algorithm did produce continuous predictions during the game, and the level of the game kept increasing in relation to the predictions, it is only the game and not the predictions that lack in its design. Since the predictions all were in the same limits, with a mean of 12 error-percentage and a low standard deviation, it rather puts validity to the algorithm's ability to produce predictions that are stable between participants performing the same task.

The fact that there is a certain slope in the beginning of the predictions (Fig. 7) for all the participants, although not significant, gives further confirmation of the general behavior seen in ocular measures in different participants performing the same task. It also does add some leverage to the previously stated assumption that participants may perceive some nervousness right after the initiation of the eye tracking. The predictions did reach a more stable state after some time spent on the task, which indicates that the slope was not produced by any heightened cognitive demand at the beginning of the task and should be considered excluded completely from the predictions in upcoming studies.

In general, the results of the second experiment does confirm the stability over participants when it comes to predicting the level of cognitive demand created by any task.

4 General discussion

The present study shows that combining several different ocular measurements is a suitable way of measuring cognitive load online, independently of agent or task. The results are compelling for their applicability in several areas in the future, including HMDs and IoT.

The investigated method reveals that the use of a regression model makes it is possible to find general behaviors between participants. At the same time, the method makes it possible to set individual levels of a comfortable mental workload and sync it with information gathered from the agents' surroundings. The general behavior found between task and participants does broaden the area of applicability, since the algorithms do not have to be individually trained and evaluated before an agent can start using the equipment. The data gathering from the user can be made during the time the equipment is used, and optimize it as the agent uses it.

If needed, the optimization and adaption to the agent can eventually be based on the individual completely, and the coefficients, and intercept, contributing to the algorithm would change in relation. Studies investigating predictions made on single individuals, collected from individualized sessions have been proven to look a bit different in their contributions to the predicting algorithms. Van Orden et al (2000; 2001) used personalized eye data from one or several sessions to train their learning algorithms, in their studies. They have investigated the usage of several different eye measurements to assess both fatigue during a visual compensatory tracking task, and task workload through a target identification memory task. They found that blink frequency, fixation frequency and pupil diameter were the best predictors in relation to how well a participant succeeded in the target identification memory task, while fixation duration and fixation frequency were the best predictors for tracking error in the visual compensatory tracking task.

The fact that there were very few common findings between the studies done by Van Orden and colleagues (2000; 2001) may be due to the fact that they were seeking to investigate different mental processes. Although, in relation to the present study this does pose some questions.

Comparing the studies by Van Orden and colleagues (2000; 2001) with the present study does draw an eye to the similarities in their investigation in making predictions on a participant's performance during a target identification memory task, which should induce more of the same processes as done in the n-back task, than the visual compensatory task. But, when studying the results of the three experiments, the current one is more similar to the visual compensatory task than the target identification memory task. The fact that the same findings as in the fatigue assessing task can be seen in this experiment, produces some questions of what kind of load the n-back task puts on the participant.

The similarities between the results from the fatigue inducing task by Van Orden et al (2000) and the n-back task in this study may be explained by the fact that a n-back task is extremely demanding and does put a lot of demand on the agent's cognitive abilities. It may be that the fixation measurements primarily do reflect the level of fatigue during a task, which also may explain why this is the only measurement that keeps increasing, even after the participant did report on giving up on the task at hand.

In that case, it would also add more leverage to the assumption that the reason to why there only were a significant finding in blink frequency in the first three rounds of the n-back task is because of the sensitivity of the onset of an extremely challenging task or that the participants in fact gave up earlier during the task than stated.

The fact that no changes in pupil size could be found in the general dataset points towards the assumption that pupil dilation in relation to task is a very individual measurement. Since others have stated the measure as stable over different tasks (Beatty, 1982), and the measure has been studied for a long time, it is not likely that the stated behavior is missing in this certain study, but more likely that it is a very individual measurement. Pupil dilation might be such a measurement more applicable in evaluating individual changes in cognitive demand, from data collected from the same individual. It thereby has a great ability in contributing in the more individualized prediction algorithms. The relation between an individual's pupil dilation and their performance on the n-back task in this study is of great interest for future studies, giving better insight into its contribution to the predictions. Thus, there is no doubt that it does contribute, since the coefficient for pupil size in the learning algorithms does rank as the third highest, after fixation duration and fixation frequency.

A conclusion that there are many factors contributing to how our ocular behavior relates to the task at hand can be made. Some depending on what mental processes the task demands of the agent, but also a lot of individual behaviors. Although, the results of this experiment show that even more general data can be used to predict task performance, even in different tasks.

Given the unclear distinction between how much the task and agent does affect the preciseness of the predictions of performance, future studies may want to examine how much difference there is between predictions made from a training set consisting of several participants and a training set consisting of data from the same participant that predictions are to be made on. Intuitively the predictions may vary quite a bit, which make them applicable in different areas in the future.

Future studies of interest would also include investigating whether data collected from a different set of participants can be used to make predictions on a new set, evaluating how general the eye responses are across participants. A more thorough investigation of the relation between the different levels of cognitive demand and their subjective evaluation of different agents is of great interest. This to be able to evaluate different levels of cognitive load and their relation to ocular behavior.

Ocular measurements do outperform many of the more invasive or bulky online measuring systems in such that they do not obtrude with the agent, are easy to use and are able to evaluate data in real time, as has been shown in this study. The results show that a combination of several ocular measurements do reflect changes in cognitive demand posed by task. Collecting data from the eyes is a useful way of getting online information about the agent's cognitive state, and by the contribution of regression algorithms the measurements creates a highly useful tool of evaluating the information collected.

But, the ability to gather, analyze and make adaptions to the information given to the agent in real time does not only need ocular data as input, to be able to function. It also puts some demands on the upcoming technology collecting ocular data, in combination with tools able to gather online information from the agent's surroundings. For example, sensoring systems are needed to provide information about the environment, for measuring level of brightness and sound, which have been proven to effect, at least, pupil size (Holmqvist et al., 2011). Fortunately, the separation of luminance from cognitive induced changes in pupil size has been done successfully (Palinko &. Kun, 2011) and the coming years in technological development will have to reveal how to implement all these exciting results found.

The findings in this study shed promising light on our ability to achieve the goal of creating truly adaptive systems in the future, where ocular measurements do have a great chance of contributing to the long sought for online cognitive measuring system.

Acknowledgements

The author gratefully acknowledges Lund University Humanities Lab, and their ever supportive and helping staff, that made this study possible.

References

Aron, A., Coups, E. J., & Aron, E. A. (2013). *Statistics for psychology*. Pearson.

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

Beatty, J., and Lucero-Wagoner, B. (2000). The pupillary system. In Handbook of Psychophysiology, J.T. Cacioppo, L.G. Tassinary, and G. Berntson, eds. (Cambridge, MA: Cambridge University Press), pp. 142–162.

Bonifacci, P., Ricciardelli, P., Lugli, L., & Pellicano, A. (2008). Emotional attention: effects of emotion and gaze direction on overt orienting of visual attention. *Cognitive Processing*, *9*(2), 127-135.

Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eye-trackers. *Behavior Research Methods*, *45*(4), 1322-1331.

Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, *42*(3), 361-377.

Brouwer, G. J., Van Ee, R., & Schwarzbach, J. (2005). Activation in visual cortex correlates with the awareness of stereoscopic depth. *Journal of Neuroscience*, *25*(45), 10403-10413.

Eckstein, M. K., Guerra-Carrillo, B., Singley, A. T. M., & Bunge, S. A. (2016). Beyond eye gaze: What else can eye-tracking reveal about cognition and cognitive development?. *Developmental Cognitive Neuroscience*.

Evans, D. (2011). The Internet of Things. How the Next Evolution of the Internet Is Changing Everything. Cisco Inter-

net Business Solutions Group (IBSG). https://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf

Evans, D. C., & Fendley, M. (2017). A multi-measure approach for connecting cognitive workload and automation. *International Journal of Human-Computer Studies*, *97*, 182-189.

Geng, J. J., Ruff, C. C., & Driver, J. (2009). Saccades to a remembered location elicit spatially specific activation in human retinotopic visual cortex. *Journal of Cognitive Neuroscience*, *21*(2), 230-245.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, *33*(4), 457-461.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Johnson, E. L., Miller Singley, A. T., Peckham, A. D., Johnson, S. L., & Bunge, S. A. (2014). Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Frontiers in Psychology*, *5*:218, 1-8.

Kahneman, D., Tursky, B., Shapiro, D., & Crider, A. (1969). Pupillary, heart rate, and skin resistance changes during a mental task. *Journal of Experimental Psychology*, *79*(1p1), 164-167.

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

Lamp, G., Alexander, B., Laycock, R., Crewther, D. P., & Crewther, S. G. (2016). Mapping of the underlying neural mechanisms of maintenance and manipulation in visuospatial working memory using an n-back mental rotation task: A functional magnetic resonance imaging study. *Frontiers in Behavioral Neuroscience*, *10:*87, 1-10.

León-Domínguez, U., Martín-Rodríguez, J. F., & León-Carrión, J. (2015). Executive n-back tasks for the neuropsychological assessment of working memory. *Behavioural Brain Research*, *292*, 167-173.

Liu, Y., & Yao, X. (1999). Ensemble learning via negative correlation. *Neural Networks*, *12*(10), 1399-1404.

Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, *3*, 2854-2861.

Mohammed, M., Khan, M. B., & Bashier, E. B. M. (2016). *Machine Learning: Algorithms and Applications*. Crc Press.

Naber, M., Alvarez, G. A., & Nakayama, K. (2013). Tracking the allocation of attention using human pupillary oscillations. *Frontiers in Psychology*, *4*, 919.

NASA TLX: Task Load Index. (2017). Retrieved 24th may 2017 from National Aeronautics and Space Administration: https://humansystems.arc.nasa.gov/groups/tlx/

O'Donnell, R. D., & Eggemeier, T. D. (1986). Workload assessment methodology. InK. R. Boff, L. Kaufman and J. Thomas, (Eds.). Handbook of perception and human performance, Vol 2: Cognitive processes and performance (pp. 42.1-42.49).

Palinko, O., & Kun, A. L. (2011). Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. In *Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design*, (pp. 329-336). Iowa City: Public Policy Center, University of Iowa.

Pederson, T., Janlert, L. E., & Surie, D. (2011). A Situative Space model for mobile mixed-reality Computing. *IEEE Pervasive Computing*, *10*(4), 73-83.

Psychology Software in Python. (2017). Retrieved 24th May 2017 from Psychopy: http://www.psychopy.org

Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Esychology: Applied*, *6*(1), 31.

Scikit-learn Linear Regression Model. (n.d.). Retrieved 10th may 2017 from Scikit-learn: http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Scikit-learn Multi-layer Perceptron Regressor. (n.d.). Retrieved 10th may 2017 from Scikit-learn: http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html

Society for Psychophysiological Research Workshop. (2011) Analysis of Pupillary Data. Boston.

The Eyelink 1000 plus. (2017). Retrieved 7th February 2017 from SR-Research: http://www.sr-research.com/eyelink1000plus.html

Van Orden, K. F., Jung, T. P., & Makeig, S. (2000). Combined eye activity measures accurately estimate changes in sustained visual task performance. *Biological Psychology*, *52*(3), 221-240.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(1), 111-121.

Veltman, J. A., & Gaillard, A. W. K. (1998). Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, *41*(5), 656-669.