Pupil Dilation as an Index of Learning

Ciara Sibley[1], Joseph Coyne[1], Carryl Baldwin[2]

[1]Naval Research Laboratory, Washington, DC
[2]George Mason University, Fairfax, VA

Physiological assessment of cognitive processes has become a topic of increased interest. The value of understanding and measuring brain function at work has the potential to improve performance. The emphasis of this paper is to discuss how pupil diameter can be applied to learning. The link between pupil diameter and task difficulty, or cognitive load, has been repeatedly demonstrated for the past 40 years. However there has been little work to date on measuring cognitive load during training or looking at how real time metrics of cognitive load could be used to adapt training. According to Cognitive Load Theory, cognitive load should be reduced as an individual learns a task and he/she relies more on long term memory than working memory. Ten participants completed a simulated unmanned aerial vehicle task in which they had to identify targets and report their direction of movement. There were three levels of increased difficulty. As expected, pupil diameter significantly dropped within each block as participants learned the task, and then increased again at the start of the next level of difficulty. The results suggest that pupil diameter may be a useful metric for assessing when an individual has transferred information into long term memory. Implications for how pupil diameter can be used to drive an adaptive training system are discussed.

INTRODUCTION

Neurophysiological measurement affords a unique opportunity to understand the interaction between working memory utilization and skill development. The ability to track workload in real time during training could provide insight into when an individual has acquired a new skill and then be used to advance training. Although a number of sensors have been used for measuring workload (e.g., EEG, fNIR, EKG), this study focuses on pupillometry. The ability of modern eye trackers to monitor pupillometry and eye movements without requiring the operator or learner to be fitted with any surface contact equipment makes it one of the most unobtrusive neurophysiological recording systems available. Numerous researchers have found changes in pupil diameter, divergence, fixation duration, and blink frequency to be predictive of various levels of cognitive demand in a task (de Greef, Lafeber, Oostendorp, & Lindenberg, 2009; Tsai, Viirre, Strychacz, Chase, & Jung, 2007; Van Orden, Limbert, Makeig, & Jung, 2001; Veltman & Gaillard, 1996).

In particular, the link between pupil dilation and task difficulty, with increasing tasks demands being associated with larger pupillary dilations, has been well established (Beatty & Kahneman, 1966; Marshall, 2007; Moresi et al., 2008; Siegle, Steinhauer, Stenger, Konecky, & Carter, 2003; Van Orden et al., 2001). Siegle et al. (2003) showed that pupil dilation increased parametrically with a working memory task (digit sorting) both inside and outside an fMRI machine. Additionally, concurrent activation of the middle frontal gyrus, which has been associated with central executive and high demand functions (Spinks, Zhang, Fox, Gao, & Hai Tan, 2004) was shown during the time course of pupil dilation.

Pupil diameter has also been found to be linked to motivation. Heitz et al. (2008) hypothesized that individual differences in working memory span performance were largely influenced by motivation. Specifically, that individuals with a low working memory span would be less motivated to perform. Heitz provided a performance-based financial incentive during a working memory task of multiple levels of difficulty, and compared high and low span individuals under rewarded and un-rewarded conditions. Results confirmed that pupil diameter significantly increased as the task increased in difficulty, and that significant increases in pupil diameter occurred when a financial incentive was provided.

The present study investigates the use of pupil diameter as a means of assessing cognitive load during a training task. Classifying working memory or cognitive load during training presents unique challenges. By virtue of the fact that people are *learning* a task or skill, working memory demand is expected to change with increased experience and familiarization with a certain difficulty level. That is, cognitive load would not be expected to be static across a given level of task difficulty. But rather, as the learner develops skills and associated improved performance strategies, cognitive load would be expected to decline *within* a given task difficulty level. Further, the presentation of the various difficulty levels can't be randomly assigned because learners can't be expected to perform a more difficult version of the task until they have mastered the easier levels.

Although physiologically driven metrics of cognitive workload have been used on stable tasks (i.e., learned tasks) they have not been heavily investigated during learning. The present experiment examined the potential for pupil diameter to provide a sensitive index of cognitive workload as an individual learned two UAV tasks of increasing difficulty. The ultimate goal of this research effort is to establish the groundwork for automating methods for monitoring cognitive resource utilization during learning.

Previous work (Sibley et al., 2010) has shown that pupil diameter is sensitive to increasing difficulty levels of a task, when the overall pupil size of the participant is averaged over an entire block of difficulty. This research also showed that pupil size was even greater during the periods of mental calculation, compared to the pupil size during the rest of the trial. For this most recent effort, it was hypothesized that pupil diameter would be sensitive to cognitive effort within a specific task difficulty level (as the learner became more proficient with the task). Specifically, we hypothesized that pupil diameter would be highest during the periods of greatest mental effort, but that it would decrease with experience or practice at a specific difficulty level.

METHOD

Participants

Fifteen participants (4 male) ranging from 18- 41 years of age (M= 22, SD= 6.2) participated in the experiment. Participants were volunteers from the George Mason University who gave informed consent to engage in a UAV training simulation experiment in exchange for course credit. Five additional participants' data were excluded due to inadvertent problems with missing data. Additionally, data for one of the participants was excluded specifically in the ID analysis, due very low performance on each of the blocks. Data from that specific participant was however able to be included in the heading task analysis.

Materials

Virtual Battlespace 2 (VBS2) was used to construct simulated UAV video files that were played back in a separate application created for this experiment. VBS2 is a high-fidelity, three-dimensional virtual training system used for military training exercises. In addition, the Tobii X120 off the head unit was used to collect eye tracking data. The unit sat in front of the participant and just below the surface of the monitor running the simulation. The system recorded both eyes at 60 samples per second. Neuroscan was used to collect EEG data at 500 samples per second. EEG data however is not considered in this paper.

UAV Desktop Simulation

After receiving a brief PowerPoint training about the task, participants engaged in a UAV desktop simulation in which he or she was trained to report information on moving vehicles as seen from a UAV (see Figure 1). Participants were asked to identify and report heading information about the target vehicles crossing the screen. At the beginning of each experimental block, participants were given ten seconds to study a picture of each vehicle at four different views (bird's eye, left side, right side, and front-on). Participants were expected to learn to recognize each vehicle by name (ID task). An example of one of the vehicles is shown in Figure 2. For each experimental trial, participants were given the heading of the UAV and were asked to estimate the heading of the vehicle on the ground (heading task). A graphical depiction of a compass facing due north with 30 degree increments was provided to the participant for reference. After entering the target heading estimation and the identity of the target, participants were then asked to rate their perceived mental effort in calculating the target heading and identifying the target.

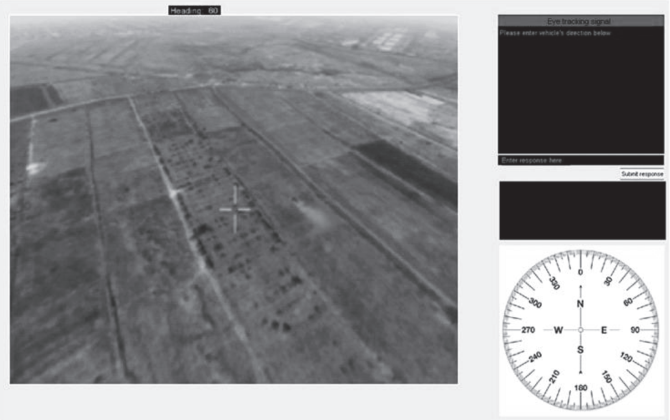
Please enter vehicle's direction below

Figure 1. Interface for the experiment

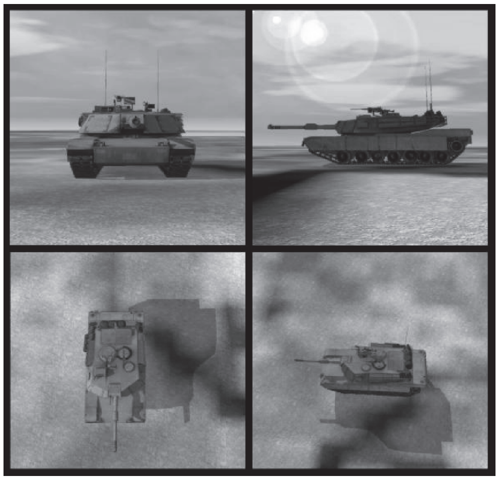

Figure 2. Example picture of an M1A1 that participants were shown and then expected to recognize by name

The difficulty of the task progressed over three blocks of trials. Only one vehicle was shown on the screen at a time and there were a maximum of 20 trials per block. Once the participant had reached an 80% ID accuracy after at least ten trials, he or she would be moved to the next level of difficulty. Participants also could transition to the next heading difficulty level if they reached a certain performance threshold. Participants could progress to the next level of the ID task without moving to the next level of the heading task, and vice versa, but after 20 trials, the participant would be forced to the next difficulty level.

Difficulty of the heading task was manipulated by varying the UAV heading as well as the possible target heading. Additionally, the ID task difficulty was manipulated by increasing the number of vehicles the participant learned for each block from two to four to six. For example, the easiest level (block one) showed the UAV heading at only 0 degrees and the target heading's were randomized in 30 degree increments. During the easiest level of the ID task participants were only expected to learn two vehicles.

The most difficult level (block three) showed the UAV heading at any 30 degree increment and changed after each target, and the target heading could also be any 30 degree increment. During the most difficult level of the ID task, participants were presented with six vehicles and were expected to recognize all six of them by name.

The Experiment

Written informed consent was obtained from all participants and the participants were then introduced to the experimental tasks. This experiment took place over one day with a duration of approximately two to three hours, including EEG preparation, eye tracking calibration, training and experimental trials. After being prepped for EEG, participants reviewed a PowerPoint training on the task and then began the experiment. All participants completed blocks one, two and three in the same order from easiest to hardest.

RESULTS

The researchers considered the most cognitively demanding parts of the simulation to be during the target heading calculation and the ID task. Therefore, this analysis primarily focused on the pupillometry data during these tasks. The data was broken up into sections/tasks in order to compare pupil dilation during the high demand parts of the task to other less demanding parts. Additionally, in order to investigate the affect of learning on pupil dilation, the first and last three trials of each difficulty block were analyzed and compared to each other. We hypothesized that as the participant began to learn the material, it would become less challenging throughout that difficulty block, and that his/her pupil size would get closer to baseline levels (decrease) towards the end of the block, i.e. the last three trials of each block. We also hypothesized that the performance data would show the same pattern, with performance increasing towards the end of each block, as the participant learned and became more proficient at each level of difficulty.

Maximum Pupil Size and Difference Scores

To best capture the exact period of mental effort during the heading calculation task, we took the maximum pupil size (an average of the top five pupil sizes, in order to reduce the effect of outliers) during each trial of the ID and heading tasks. Difference scores were then calculated for pupil dilation by averaging each individual's pupil size during a baseline period of each trial when no mental effort should have been expended. The average of this pupil size was then subtracted from the maximum pupil size during each individual trial. This was done in order to show change from the baseline pupil size and to reduce factors caused by individual variability (i.e. an effect being driven by one individual with large pupils). Positive scores mean that the pupil size was greater than baseline, while negative scores would indicate pupil size was smaller than baseline.

Figure 3 shows the average pupil size difference scores during the heading task for each block, during the average of the first and last three trials. One may observe not only that pupil size is far above baseline (baseline being zero), but also that that pupil size attenuates towards the end of each block and then jumps back up at the beginning of the next difficulty block. This pattern similarly corresponds with performance data (figure 4) where the participants' performance becomes significantly better towards the end of each block.

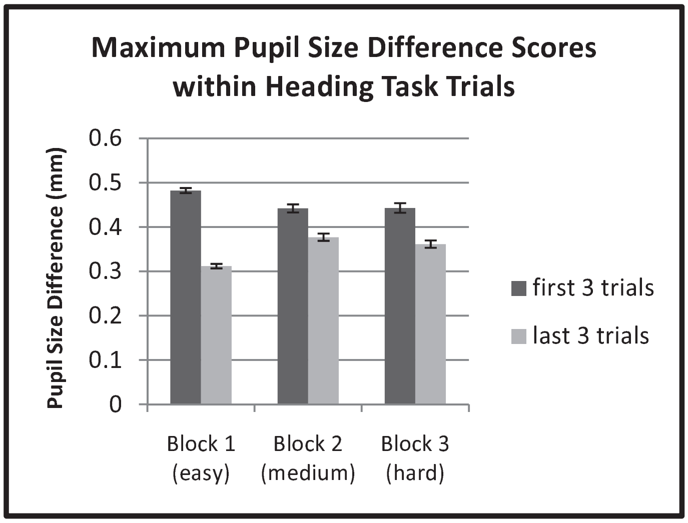

Figure 3. Maximum pupil difference size is highest at the beginning of the new block of difficulty and then attenuates with learning

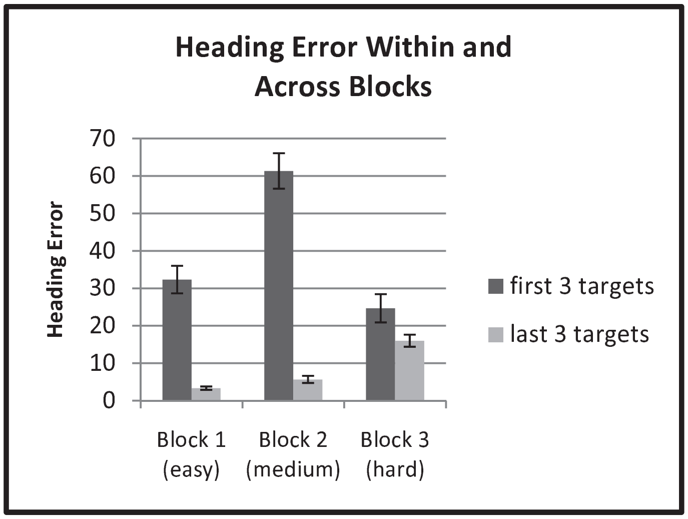

Figure 4. Heading performance data follows the same trend as pupil data, showing participants' error reducing by the end of each block

A two-way (trial placement x block) repeated measures ANOVA was conducted to compare the pupil size during the heading task for the first three trials to the last three trials across each block. There was a significant difference in pupil size between the first three trials of each block (M=0.44, SD= 0.16) compared to the last three trials of each block (M=0.36, SD= 0.14), F(1, 9) = 23.37, *p*= 0.00. The same difference was found in the heading performance data between the first three trials of each block (M=39.44, SD=42.60) and the last three trials of each block (M=8.22, SD=12.09), F(1,9)= 19.26, *p*= 0.00.

A two-way (trial placement x block) repeated measures ANOVA was also conducted for the ID task and showed a significant difference in pupil size between the first three trials of each block (M= 0.35, SD= 0.17) and the last three trials of each block (M=0.28, SD= 0.20), F(1, 8)= 9.66, *p*=0.01. A significant difference was also found between the ID accuracy of the first three trials of each block (M=0.54, SD=0.31) and the last three trials of each block (M=0.70, SD=0.40), F(1,8)= 10.56, *p*= 0.01.

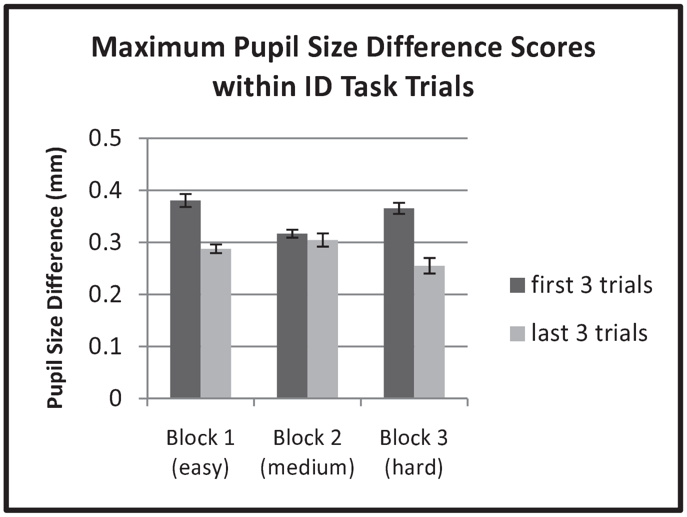

Figure 5. Maximum pupil difference size is highest at the beginning of the new block of difficulty, and then attenuates with learning

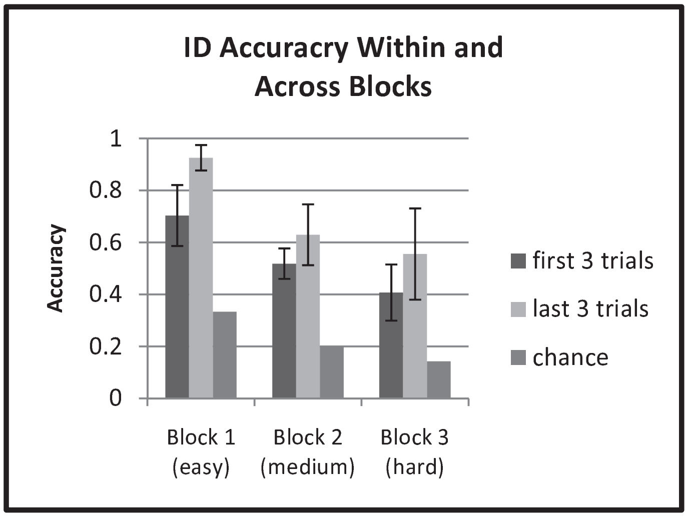

Figure 6. ID accuracy performance data follows the same trend as pupil data, showing participants' accuracy increasing by the end of each block

DISCUSSION

Analysis of pupil dilation data shows systematic changes in pupil size with practice on a task. Findings confirmed our hypothesis that pupil size would peak at the beginning of each difficulty block and then decrease as the individual learned the task. This is also supported with performance data showing participants improving towards the end of each difficulty block. These findings are consistent with Cognitive Load Theory suggesting that when an individual has learned a task the amount of effort that he/she needs to expend reduces as information is transferred from working memory into long-term memory. We speculate that the attenuation in pupil size is related with the transfer of information into long term memory.

This systematic change in pupillometry could potentially provide the ability to assess whether an individual has learned a task and does not need to expend as much effort in order to complete that task effectively. The applications for this finding are numerous. For example, one could identify individuals who are able to successfully complete a task with little effort and assign them to certain tasks in which he/she will excel. This could also be used to ascertain whether an individual is struggling to complete a task or could be given an additional task to perform. Furthermore, pupillometry metrics could be used in situations where performance measures are not available. This could also be very beneficial information to have for setting the pace to train an individual on a novel task.

More research needs to be conducted to investigate the feasibility of using pupillometry to assess learning and cognitive load. This sample size was rather small (N= 10) and thus more research needs to be done to replicate these findings and to help tune algorithms for use in the real world and closed-loop systems. We believe that by analyzing a combination of tonic and phasic changes in pupillometry data, as well as performance data, we should be able to improve conventional training systems that either do not account for learning, or base learning on performance measures alone.

REFERENCES

Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science, 5*(10), 371-372.

de Greef, T., Lafeber, H., Oostendorp, H., & Lindenberg, J. (2009). *Eye Movement as Indicators of Mental Workload to Trigger Adaptive Automation*. Paper presented at the Proceedings of the 5th International Conference on Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience: Held as Part of HCI International 2009.

Marshall, S. P. (2007). Identifying cognitive state from eye metrics. *Aviation Space and Environmental Medicine, 78*(5), B165-B175.

Moresi, S., Adam, J. J., Rijcken, J., Van Gerven, P. W. M., Kuipers, H., & Jolles, J. (2008). Pupil dilation in response preparation. *International Journal of Psychophysiology, 67*(2), 124-130.

Siegle, G. J., Steinhauer, S. R., Stenger, V. A., Konecky, R., & Carter, C. S. (2003). Use of concurrent pupil dilation assessment to inform interpretation and analysis of fMRI data. *Neuroimage, 20*(1), 114-124.

Sibley, C., Coyne, J. T., Cole, A., Gibson, G., Baldwin, C. L., Roberts, D., Barrow, J. (2010). Adaptive training in an Unmanned Aerial Vehicle: Examination of several candidate real-time metrics. In W, Karwowski and G. Salvendy, (Eds.), *Applied Human Factors and Ergonomics*, Boca Raton, FL: Taylor & Francis.

Spinks, J. A., Zhang, J. X., Fox, P. T., Gao, J.-H., & Hai Tan, L. (2004). More workload on the central executive of working memory, less attention capture by novel visual distractors: evidence from an fMRI study. *NeuroImage, 23*(2), 517-524.

Tsai, Y. F., Viirre, E., Strychacz, C., Chase, B., & Jung, T. P. (2007). Task performance and eye activity: Predicting behavior relating to cognitive workload. *Aviation Space and Environmental Medicine, 78*(5 II).

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T.-P. (2001). Eye Activity Correlates of Workload during a Visuospatial Memory Task. *Human Factors, 43*(1), 111.

Veltman, J. A., & Gaillard, A. W. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology, 42*(3), 323-342.