Method for Characterizing and Identifying Task Evoked Pupillary Responses During Varying Workload Levels

Allan Fong[1], Ciara Sibley[2], Joseph Coyne[2], Carryl Baldwin[3]

[1] Naval Sea Systems Command, Washington Navy Yard, DC
[2] Naval Research Laboratory, Washington, DC
[3] George Mason University, Fairfax, VA

Understanding when operators are experiencing high workload is important in the design and implementation of Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance (C4ISR) systems. Fortunately physiological metrics, such as pupillary reflexes, have been shown to correlate with increases in mental workload. This paper proposes an automated method for characterizing and identifying task evoked pupillary responses (TEPR) during various workload levels. This method captures findings and observations from previous TEPR studies in an automated algorithm. This algorithm characterizes the rate of pupil dilation and constriction into a TEPR area metric, which is then used to identify times of increased operator workload. Independent trial analysis shows the benefits of using the TEPR area for distinguishing different workload responses but additional investigation is needed to make the algorithm more robust to individual variability.

INTRODUCTION

As computing power increases and military operational environments become more complicated, warfighters have to constantly push the limits of their physical and mental abilities. Assigning too many tasks to an operator without understanding their effects on cognitive load, or workload, can cause the operator to make poor and even catastrophic decisions. Hence, it is important to measure and understand the effects different tasks and stimuli have on workload, especially when designing human-computer interfaces (Sweller, 2006). Several physiological metrics, including heart rate, electroencephalograph, galvanic skin response, and pupillometry are used to model workload (Ahern and Beatty, 1979; Marshall, 2002; Marshall, 2007; Van Orden, Limbert, Makeig, & Jung, 2001; Wilson, Estepp, & Davis, 2009; Wilson & Russell, 2003; Wilson & Russell, 2007). In this paper we focus on pupillometry metrics because they have been reliably correlated with workload (Iqbal, Zheng, & Bailey, 2004; Klingner, Kumar, & Hanrahan, 2008; Marshall, 2007; Moresi, Adam, Rijcken, & Van Gerven, 2008; Nakayama & Shimizu, 2004; Palinko, Kun, Shyrokov, & Heeman, 2010; Van Orden, Limbert, Makeig, & Jung, 2001). In addition, improvements in eye tracking technologies have made collecting pupillometry metrics less cumbersome and invasive than a number of other methods.

Although an individual's pupil dilates during increased workload, studies have shown that changes in pupil diameter can also be caused by a range of other factors, such as lighting conditions and fatigue (Geacintov & Peavler, 1974; LeDuc, Greig, & Dumond, 2005). Furthermore, pupil sizes and reflexes naturally vary amongst individuals making it challenging to associate pupil diameter averages with workload, especially across individuals and for long complex tasks. In this paper, we propose a method to examine and identify specific task evoked papillary response (TEPR) signatures in a visual unmanned aerial vehicle (UAV) task with varying levels of workload. We assess the utility of this method for identifying TEPR events and classifying workload levels.

Task Evoked Pupillary Response

Task evoked pupil dilations has been shown to correlate with increased mental workload (Ahern and Beatty, 1979; Iqbal, Zheng, & Bailey, 2004; Klingner, Kumar, & Hanrahan, 2008). In addition, an individual's pupil remains dilated longer during more difficult cognitive tasks. A number of methods involving averages, percent changes, and wavelet analysis, have been used to study this pupil reflex (Iqbal, Zheng, & Bailey, 2004; Marshall, 2002; Marshall, 2007). This paper builds on the idea that a pupil reflex can be analyzed in near realtime by proposing a method for detecting unique TEPR characteristics correlated to increased workload. An advantage of this method is that it does not solely rely on pupil diameter block or trial averages.

Our hypothesis is that task evoked pupillary responses can be characterized and can help classify different workload levels. We define the dominant features of the papillary reflex during a mentally challenging task as a rapid increase in pupil diameter followed by a gradual return to normal size, where the constriction rate is inversely related to the workload level experienced (the slower the constriction, the higher the workload). Furthermore, we propose using the pupil diameter area during a TEPR event (the TEPR area) as an indicator for workload: the higher the workload, the longer it takes for the pupil to return to its pre-stimulus size, resulting in a larger TEPR area (see Figure 1).

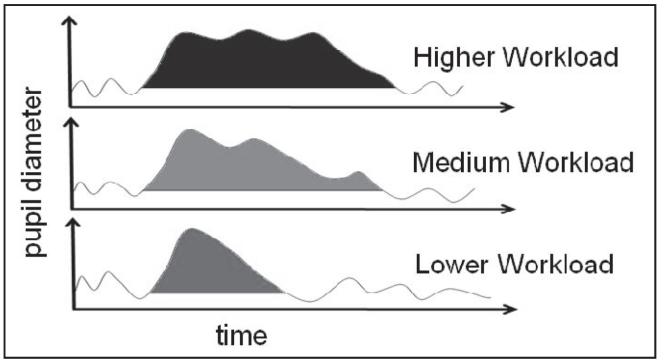

Figure 1. Conceptual illustration of one's pupil reflex under different workloads (shaded area is the TEPR area)

METHOD

Participants

Fifteen students from George Mason University volunteered to participate in our UAV training simulation experiment. All participants had normal or corrected to normal vision. However, data from four students had to be omitted from the analysis due to experimental complications.

Materials

Virtual Battlespace 2 (a high-fidelity virtual training system) was used to construct UAV simulation scenarios for this experiment. A Tobii X120 desktop unit was used to collect pupillometry data at 60 hertz. The unit was placed below the desktop monitor and in front of the participant. The system was calibrated to the subject before each experiment.

UAV Desktop Simulation

Participants engaged in a desktop simulation in which they were trained to report information on enemy target vehicles as seen from a UAV. Participants were given the heading of the UAV and had to estimate the heading of the vehicle on the ground as it traveled across the screen in various directions. In addition, a graphical depiction of a compass facing north was provided to the participant for reference (see Figure 2). After entering the target vehicle's heading, participants were asked to rate their mental effort in calculating the heading.

Figure 2. User interface for the experiment

For each trial, the vehicle appeared on the screen after a random amount of time ranging between one and five seconds from the start of the video. Once the subject saw the vehicle, he or she had to click on the screen with the mouse. The participant then had to calculate and submit the heading of the target vehicle. The time between acknowledging the vehicle's presence and submitting the heading response was when the mental calculation occurred.

Difficulty Levels

Difficulty The UAV experiment consisted of 60 trials divided into three levels of difficulties: low, medium, and high (see Table 1). In the low workload trials, the UAV heading was set to 0° (North) and target vehicle headings were randomized in 30° increments. In the medium workload trials, the UAV headings varied randomly between 90°, 180° and 270° and the target vehicle headings were randomized in 30° increments. In the high workload trials, both the UAV and target vehicle headings were randomized in 30° increments.

Algorithm Development

Building on findings and observations from previous TEPR studies, we developed an algorithm to detect different workload levels within a task (Ahern and Beatty, 1979; Iqbal, Zheng, & Bailey, 2004; Klingner, Kumar, & Hanrahan, 2008). This algorithm was scripted in Matlab and can be provided upon request.

Data Preprocessing

The raw pupil diameter data was first filtered using an one second averaging window moving every 0.1 seconds. These values were chosen to reduce the noise while providing enough signal granularity. We next calculated the rate of the pupil dilations and constrictions, which helped identify when rapid pupil dilations occurred. Pupil diameter slope was calculated over a two second window every 0.1 second. These initial values were chosen based on observations from previous studies (Ahern and Beatty, 1979; Klingner, Kumar, & Hanrahan, 2008). Identifying and optimizing these parameters for individuals and specific tasks are areas of continued research.

Workload Level	Low	Medium	High
Possible UAV heading (degrees)	0	90, 180, 270	0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330
Possible target vehicle heading (degrees)	0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330	0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330	0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330

Table 1. There were three levels of difficulty in the experiment

Characteristics of TEPR

Typically, a TEPR signal during increased mental workload is characterized by a rapid dilation of the pupil followed by a constriction period as the pupil returns to normal size. Pupil dilation rates are fairly similar across workload levels but the constriction rates vary with workload: slower constriction rates are associated with higher workload. To capture this effect, we calculate the area of the pupil diameter curve during a TEPR event (the TEPR area): the higher the workload, the more area under the pupil diameter curve, the larger the TEPR area (see Figure 1).

Workload Measures

We hypothesized that the TEPR area can be used as workload indicators; the larger the TEPR area, the higher the workload experienced by the individual.

TEPR Algorithm

We developed an algorithm that identifies the times when subjects are experiencing increased workload according to our TEPR model. This algorithm requires pupil diameter, pupil diameter slope, and a validity metric of the eye data as input variables. The algorithm consists of five steps and runs independently for each subject. Step 1 is a batch process while Steps 2-5 incrementally steps through the dataset from the start of the experiment (see Figure 3).

Step 1: Find a pupil dilation criteria

The first step is to find a pupil dilation criteria that distinguishes between rapid pupil dilations from normal pupil oscillations. Because pupils typically dilate faster during a mental stimulus, we set the dilation criteria to include the upper twenty percent of slope values. The upper twenty percent was chosen for simplicity, while providing a range of slope values with reasonable stratification. This value was subjectively assigned and additional research is needed to investigate optimal criterions that can better account for individual variability.

Step 2: Identify times of rapid pupil dilation

Next, the algorithm identifies and marks the times when pupil diameter slope exceeds the dilation criteria determined from Step 1. This marker indicates the beginning of a TEPR event. Furthermore, the pupil diameter at the start of the TEPR event is referred to as the pupil diameter baseline.

Step 3: Integrate the pupil diameter during the TEPR event

Once a TEPR event is detected, the algorithm begins summing the area between the pupil diameter and the pupil diameter baseline. This cumulative sum is referred to as the TEPR area.

Step 4: Check for break conditions

The algorithm continues to integrate the pupil diameter area until either the pupil diameter constricts back to its pre-TEPR/baseline size or the eye data becomes invalid, i.e. the eye tracker loses track of the eyes. Either one of these two conditions can end the TEPR event.

Step 5: Repeat Steps 2-4

Steps 2 through 4 are repeated until the end of the experiment. This algorithm generates many TEPR events of varying durations and magnitudes.

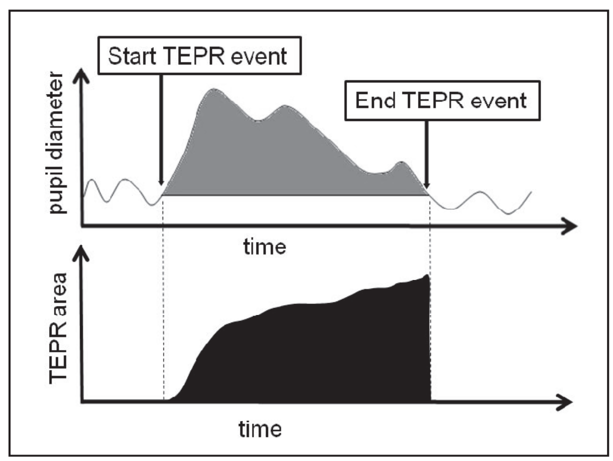

Figure 3. Illustration of Steps 2-5 of TEPR algorithm

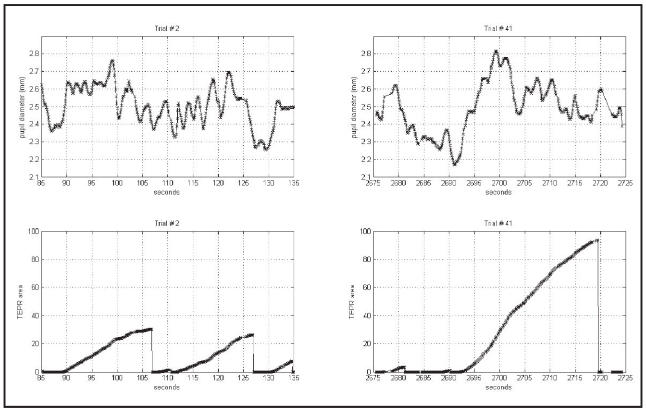

Figure 4. Data from a subject comparing pupil diameter (top row) and cumulative TEPR (bottom row) area during a low workload event, Trial #2 (left column), and a high workload event, Trial #41 (right column)

The top two graphs in Figure 4 show data from a subject's pupil diameter and the bottom two graphs show the cumulative TEPR areas. As expected the TEPR area during the higher workload trial is greater than during the low workload trial. The TEPR area is clearly larger during Trial #41 even though the pupil diameter averages for both trials are almost the same (see Table 2). The method we propose can also help identify the specific times when participants are starting to concentrate and focus more.

The TEPR area metric can be helpful in distinguishing the workload levels between trials. We conducted a within subjects ANOVA to determine whether our TEPR area metric was able to distinguish difficulty levels across the three UAV difficulty levels. Only the maximum TEPR area value for each trial was used, and the analysis focused on the heading calculation section of the experiment. Results of the analysis were not statistically significant (*p-value* = 0.20), possibly caused by the small sample size (power was only 0.25). Although not statistically significant, the results do show promise and we are currently running more subjects and will perform this analysis on a task with a simpler design and more distinct difficult levels.

Additionally, we tested the utility of this data in artificial neural networks (ANNs) because ANNs have been used successfully to develop predictive workload models with psychological data in previous studies (Van Orden, Limbert, Makeig, & Jung, 2001; Wilson, Estepp, & Davis, 2009; Wilson & Russell, 2003; Wilson & Russell, 2007). Given the specific algorithm parameters we used, the TEPR area metric does not show a significant increase in accurate classification rates when incorporated into ANN models (see Figure 5). We conducted a single factor ANOVA to assess the effect of the TEPR area metric on classification performance for eight subjects. Although the classification rates for three participants increased by five percent, overall classification improvement was not statistically significant (*p-value* = 0.8). Again, we believe that as we increase the sample size and refine our algorithms to better account for individual variability and data validity, this statistical significant will improve.

These results suggest that the TEPR algorithm can be a useful alternative method for detecting when a participant is experiencing increasing workload. However, more work is needed to refine the constraints and parameters governing the algorithm.

	Trial #2	Trial #41
TEPR area (max)	30.45	93.88
Pupil diameter (average)	2.51	2.49

Table 2. Comparing pupil diameter and TEPR metrics

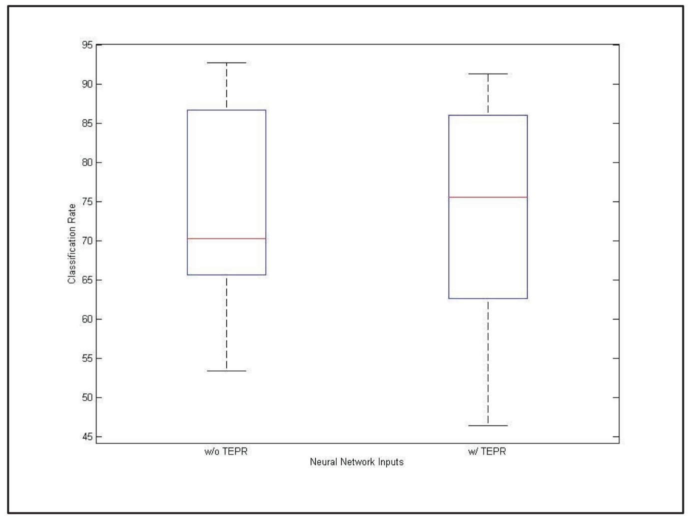

Figure 5. Classification rates of neural network models

DISCUSSION

In this paper, we proposed and developed a method that incorporates TEPR research into an automated search algorithm aimed at identifying when operators are under high workload. This research used previous TEPR studies as a framework in developing a method that detects and highlights pupil dilation signatures corresponding to specific TEPR characteristics. Because this method does not rely only on pupil diameter averages, it will be less impacted by fatigue. For example, a subject 50 minutes into an experiment will on average be more tired than when he or she started. Hence, the subject's pupil diameter averages for the later trials would be smaller than his or her pupil diameter averages during the first few trials even if the later trials are more difficult. The TEPR algorithm we propose will be better at addressing this issue because it is more dependent on pupil dilation and constriction rates.

Given the criterions used in the algorithm, the results and effectiveness of the TEPR metric differed across individuals. For some individuals, the addition of the TEPR metric was helpful in developing better predictive neural network models. For other, the classification performance of their models either remained the same or slightly decreased. This could be caused by the conservative data validation constraints we set in the TEPR algorithm. This is an area that requires further analysis and investigation. The classification results could also be caused by random seeds associated with developing neural networks and the time increments of the inputs. Additional saliency analysis and Monte Carlo simulations can be used to assess if the TEPR metrics will significantly improve overall neural network performances across subjects.

It is apparent that more work needs to be done to improve the adaptability of this algorithm to different individuals. We will look at additional ways to determine the pupil diameter criteria, allowing the criteria to change with time to better account for experimental factors. Furthermore, additional research is needed to understand the sensitivity of the TEPR metric when subjects look at different screen locations with varying brightness and contrast levels.

Although the methodology proposed has similarities with the wavelet analysis researched by Marshall (2002), we believe that this approach is more intuitive and can be implemented easier. The algorithm is transparent and the steps are fairly simple. The parameters for this algorithm can also be adjusted and customized to individual subjects and tasks.

In this paper we presented a method for analyzing pupil diameter data for specific TEPR event signatures. The algorithm we developed is based on previous TEPR studies and observations. It assumes that TEPR events can be characterized by rapid pupil dilations followed by pupil constrictions where the rate of pupil constriction is inversely proportional to the workload level experienced.

The TEPR metric can distinguish between trials from different workload blocks and can provide additional benefits to pupil diameter averages when the pupillometry data is valid. Further work is needed to make the algorithm more robust and generlizable across individuals. Although this method is currently applied post-hoc, our goal is to, after ensuring the method's validity, adapt it to real-time analysis.

REFERENCES

Ahern, S., and Beatty, J. Pupillary Responses During Information Processing Vary with Scholastic Aptitude Test Scores. *Science*, 205 (1979), 1289-1292.

Geacintov, T., and Peavler, W. S. Pupillography in industrial fatigue assessment. *Journal of Applied Psychology*, 59, 2 (1974), 213-216.

Iqbal, S.T., Zheng, X.S., and Bailey, B.P. Task-Evoked Pupillary Response to Mental Workload in Human-Computer Interaction. In *Proc. CHI 2004*, ACM Press (2004), 1477-1480.

Klingner, J., Kumar, R., and Hanrahan, P. Measuring the Task-Evoked Pupillary Response with a Remote Eye Tracker. *Proc. Symposium of Eye-Tracking Research and Applications 2008*, ACM Press (2008), 69-72.

LeDuc, P. A., Greig, J. L., and Dumond, S. L., Involuntary Eye Responses as Measures of Fatigue in U.S. Army Apache Aviators. *Aviation, Space, and Environmental Medicine*, 76 (2005), C86-C91.

Marshall, S.P. The Index of Cognitive Activity: Measuring Cognitive Workload. *Proc. Human Factors*, IEEE (2002), 5-9.

Marshall, S. P. Identifying Cognitive State from Eye Metrics. *Aviation, Space, and Environmental Medicine.* 78, (2007), B165-B175.

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

Nakayama, M., and Shimizu, Y. Frequency Analysis of Task Evoked Pupillary Response and Eye-movement. *Proc. Symposium of Eye-Tracking Research and Applications 2004*, ACM Press (2004), 71-76.

Palinko, O., Kun, A.L., Shyrokov, A., and Heeman, P. Estimating Cognitive Load Using Remote Eye Tracking in a Driving Simulator. *Proc. Symposium of Eye-Tracking Research and Applications 2010*, ACM Press (2010), 141-144.

Sweller, J. Discussion of Emerging Topics in Cognitive Load Research: Using Learner and Information Characteristics in the Design of Powerful Learning Environment. *Applied Cognitive Psychology*, 20 (2006), 353-357.

Van Orden, K. F., Limbert, W., Makeig, S., and Jung, T.P. Eye Activity Correlates of Workload during a Visuospatial Memory Task. *Human Factors 43,* 1 (2001), 111-121.

Wilson, G.F., Estepp, J., and Davis, I. A Comparison of Performance and Psychophysiological Classification of Complex Task Performance. *Proc. Human Factors and Ergonomics Society 2009*, 141-145.

Wilson, G. F., and Russell, C. A. Real-Time Assessment of Mental Workload Using Psychophysiological Measures and Artificial Neural Networks. *Human Factors 45,* 4 (2003), 635-643.

Wilson, G. F., and Russell, C. A. Performance Enhancement in an Uninhabited Air Vehicle Task Using Psychophysiologically Determined Adaptive Aiding. *Human Factors 49,* 6 (2007), 1005-1018.