Cognitive Workload, Pupillary Response, and Driving: Custom Applications to Gather Pupillary Data

Thomas M Gable
Georgia Institute of Technology
654 Cherry Street Atlanta GA 30332
001.404.894.2680
thomas.gable@gatech.edu

Bruce N Walker
Georgia Institute of Technology
654 Cherry Street Atlanta GA 30332
001.404.894.8265
bruce.walker@psych.gatech.edu

Ashley G Henry
Georgia Institute of Technology
654 Cherry Street Atlanta GA 30332
001.404.894.2680
ashleyghenry@gatech.edu

ABSTRACT
Drivers often increase their cognitive workload (CW) through the use of in-vehicle technologies for communication, information, or entertainment. Previous work has attempted to measure CW in the driving domain through the use of performance, subjective, and physiological measures, however few have attempted to use pupillary response to estimate CW. The present work discusses a method of analyzing previously collected eye tracking data despite the eye tracking device's lack of pupil diameter (PD) analysis abilities and the validity of such a measure in the driving context. A custom made parser program was developed to gather the data from the eye tracking files and then put through another program to split and organize the data into the correct blocks and averages. The paper also addresses the difficulties in using such custom application for PD analysis as well as how to address issues of light induced pupillary response and a short discussion of standardization of pupillary response for CW in driving.

Categories and Subject Descriptors
H.5.2 [**Information Interfaces And Presentation (e.g., HCI)**]: User Interfaces –graphical user interfaces (GUI), interaction styles (e.g., commands, menus, forms, direct manipulation), user-centered design; I.6.7 [**Simulation and Modeling**]: Simulation Support Systems

General Terms
Measurement, Reliability, Human Factors, Standardization, Verification.

Keywords
Driving, Cognitive Workload, Pupillary Response, Eye Tracking

1. INTRODUCTION
The definition of cognitive workload (CW), also referred to as cognitive demand or cognitive load, has long been debated in the psychological community. For the purpose of this paper CW will be defined similarly to how Mehler, Reimer, and Coughlin [1] defined cognitive demand based on De Warrd's [2] book: load or demand referring to the features of a task an individual performs and workload meaning the affect on the individual due to his or her performance of the task. An area where researchers focus heavily on CW within psychology is in that of driving, more specifically the area of driving and secondary tasks.

A prominent secondary task performed while driving is the use of in-vehicle technologies (IVTs). This use of IVTs while driving has been found to increase the CW of a driver [3]. This additional CW has been found to significantly decrease a driver's sensitivity to road events as well as lower their confidence in detection [4]. Furthermore, as the CW of a task increases, the risk of the user making an error before completing the task increases [5]. Within an attention-demanding task such as driving, increases in CW can make a big difference in safety. Measuring CW, whether during research or in real time, without any interaction from the driver, is an important area of research within the driving domain.

1.1 Measuring Cognitive Workload
Measuring CW can be achieved through subjective, performance, and physiological assessments [1,2,6]. Subjective measures, rating scales that convey the user's perception of the CW after each task, are an easy method of measuring workload. A common tool used for this measurement is the NASA-Task Load Index (TLX) assessment [8]. While this self-assessment tool provides highly validated measures of CW it is a subjective measure, which can be confounded, and it does not offer a real-time assessment without driver disruption, not ideal for measuring CW in a dynamic environment such as driving [1,7,9].

Since driving is most often performed in a dynamic environment, successful performance of the task necessitates driver attention. This highlights the need to measure CW in real-time without disrupting the user's performance [1]. To this end, unobtrusive measures such as driving performance or physiological responses can be used. In driving, performance measures are based on how well the user completes a particular part of the driving task (e.g., lane keeping, speed or speed variance, and steering wheel angle variance) [5]. While these performance measures may yield a correlation to CW, they also measure actual driving ability and can confound the data. Additionally, as every task measured in driving could be different, performance measures have to be tailored to each specific task, thus limiting the generalizability of the results [7].

Physiological measures can be used to interpret the user's cognitive state in real-time and unlike performance measures, do not have to be customized to each specific task, allowing for more flexibility and comparison across studies. Variations of these measures (e.g., heart rate, heart rate variability, respiration, eye position, and skin conductance) have been shown to correlate in an almost stepwise fashion with levels of induced cognitive load [1,7,10]. Mehler, Reimer, Coughlin and Dusek [10] found a correlation between heart rate and skin conductance and the introduction of secondary cognitive tasks, increasing the cognitive load of the drivers. However, while engaging in a secondary task, emotional and physical workload factors (body movement, temperature, stress, etc.) can also contribute to increases of these measures [7].

1.2 Pupil Diameter and Cognitive Workload

One physiological measure that has been found to react to changes in CW but not widely referred to in the driving domain is pupil diameter (PD), an effect also known as task evoked pupillary response (TEPR) [11,12]. This form of CW measurement is less intrusive than other physiological measures and could still provide an assessment in real-time. Pupillary response indicates levels of CW at each moment, communicates differences in processing load during different tasks, and conveys variance within the same task [13]. While showing that PD varies with emotional stimulation, Partala and Surakka [14] noted the difficulty of voluntarily varying PD as an advantage of the measure.

Pupil dilations occur as soon as it processes load and quickly returns to baseline state thus creating a sensitive measure of CW [9,5]. Palinko, Kun, Shyrokv, and Heeman [5] observed the dilation and contraction of pupils as driver's attention was divided by a word game. PD was found to increase as the driver thought of a word, peaked when the word was uttered, and gradually decreased before the next word. This indicated an increase in CW as words were recalled which were similar to results of Granholm, Asarnow, Sarkin, and Dykes [15] in a digit span recall task. They also found that PD increased as processing load fell below the resource limits of the cognitive task, was stationary once processing load was reached, and decreased when the user disengaged active processing, displaying that PD responds to varying levels of processing load [15].

Iqbal, Zheng, and Bailey [9] measured percentage change in PD (PCPS) by subtracting the baseline PD from the size at each task and dividing the result by the baseline size. The authors then averaged PCPS of participants performing visual tasks on the computer and results showed a significant difference in average PCPS between easier and more complex tasks. PD was also found to correlate with changes in cognitive load that varied over hierarchical tasks, indicating its validity as a measure for CW. Palinko et al. [5] also looked at mean PD and mean PD change rate when focusing on CW changes, results suggesting the measures' usefulness.

The application of PD has been attempted and seems to hold up in the driving environment as well. Recarte and Nunes [16] found that when participants were performing a secondary task while driving they had significantly larger PDs than when performing only the driving task. In a later experiment on detecting targets while performing mental tasks, Recarte and Nunes [17] noted lower percentages of detected targets causing poorer performance as a result of an increase of mental tasks and participants' workload. This increased workload while performing tasks was also shown by pupillary dilations. Similar results were seen in a simulation study where the driver performed a lane-changing task and a visual search task [18]. As the visual task was introduced, driving performance decreased and PD increased. This correspondence between performance and PD has been attributed to their convergence, but assessing using PD still creates a finer form of measurement of CW [5].

While research in this domain has included the use of TEPR to estimate CW, some researchers' technologies do not allow for access to the data required or are not currently supporting this application of the devices used. In order to increase the availability of this data and allow for researchers to use TEPR as another tool for measuring cognitive load the process of gathering and analyzing the data must be more easily completed. The current paper discusses the use of Tobii mobile eye trackers to measure CW through PD, including the creation of necessary programs to get the data from the eye trackers to an analyzable state. The paper also compares the results of the PD differences between conditions to differences seen in subjective workload measures in the same study. The data used in this report is taken from a study previously analyzed, written up, and accepted for publishing [19]. However, at the time of the research design and analysis, PD data was not an available measure due to the lack of these programs and was therefore also not gathered in a way to be analyzed specifically in this way. This lack of planning to analyze the data may have created some of the noise seen in the analysis and is discussed along with the support found for the application of this technique in future work.

2. METHODS

2.1 Participants

The participants in this analysis were 24 students at a large research university in the United States. All participants had valid drivers licenses and had normal or corrected to normal hearing and vision. The 17 males and 7 females were an average of 20.17 years old and had a mean of 4.54 years driving experience. Not all of the participants included in the initial study are included in the current analysis due to technical issues with some data files.

2.2 Apparatus and Procedure

To see an extended description of the apparatus and procedure of the study this data was taken from see Gable, Walker, Moses and Chitloor [19]. In short, participants were asked to wear Tobii eye tracking glasses while performing a dual task situation by driving the lane change task and executing a search task on a touchscreen smartphone. The participants completed 6 conditions during the experiment, 1 being a control of only performing the driving task and 5 dual task conditions. Of the 5 dual task conditions 4 had auditory cues and 1 was performed with no auditory cues.

2.3 Design and Analysis

The analysis of the data in the current report as compared to previously collected CW data is the focus of this paper. During the initial stages of the previous study and during the analysis and write-up of the data, the PD measurements were not accessible using the software available to the researchers. Recently, however, we created two custom programs in our lab that made the pupil data obtainable.

Of the two programs used in the analysis, one was created to pull the data out of the native Tobii files and the other to separate the data into blocks and give averages. First the program called TobiiReader, written in C#, extracts the values for each frame for all metrics from the proprietary Tobii projects (".gfp"s) and outputs these into tab-separated documents. This document then contains every frame value for all possible values from the eye trackers, including any frames where the eye trackers could not read the pupil due to the either tracking error or participants looking outside of the rim of the glasses. In this instance the file reports the pupillary response as 0, which can skew the data and should be addressed by anyone recreating this process. A second, command line application that was written in Ruby, called TobiiParser is then used to interpret the PD by finding averages over specific ranges of time. These time ranges are gathered by hand based on the timestamps separating blocks or conditions in Tobii Studio and input into the command line. TobiiParser then outputs the average PD with and without missing values, as well the number of missing values. The program could be modified to output other information if needed.

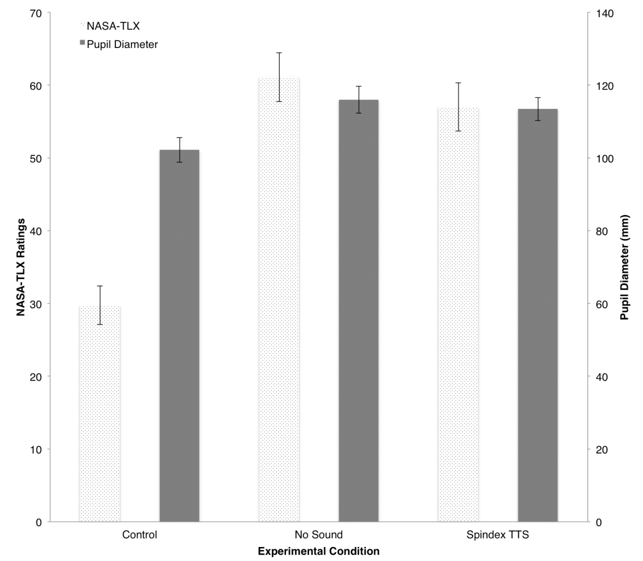

Figure 1. A bar chart displaying the mean TLX ratings and PDs of the three driving conditions: driving only, driving + search task no sound, and driving + search task spindex TTS.

This data was then entered into a spreadsheet along with the previously collected TLX ratings for the same participants for each of the blocks. Instead of comparing all of the conditions from the original study it was decided to only compare 3 of the conditions in an effort to save time while still examining the abilities of the eye trackers to measure changes in PD. The conditions that were chosen included: a control of only performing the driving task, the condition that according the TLX created the least amount of cognitive load; driving plus the secondary task with no sound, the condition that created the highest level of cognitive load according to TLX; and driving while performing the dual task with the auditory cue of spindex TTS (an advanced auditory cue, see [19] for more information), the cue that seemed to diminish the cognitive load on the drivers the most out of all the dual task conditions.

3. RESULTS

Figure 1 displays the mean combined TLX ratings and PDs for each of the three experimental conditions included in this analysis. A similar trend can be seen between the two measures for the three conditions, with the control condition having a much lower value and then increasing for the no sound search task condition before slightly decreasing in the audio condition of spindex TTS. In an effort to investigate this trend a one-tailed Pearson's r correlation test was performed, the scatterplot of which can be seen in Figure 2. Results of the test showed a moderate positive correlation between the TLX ratings and the PD (mm), $r = 0.339$, $n = 72$, $p = 0.002$.

Paired t-tests showed that the control condition had significantly lower TLX ratings ($M = 29.8$, $SD= 16.5$) than either the no sound condition ($M = 61.1$, $SD= 18.0$), $t(23) = -8.22$, $p < .001$, and the spindex TTS condition ($M = 57.0$, $SD= 15.5$), $t(23) = -9.13$, $p < .001$. No significant difference was seen between the two search conditions. Similar results were found for the PD measure with the control condition having significantly lower average PD ($M = 102.2$, $SD= 13.0$) than the no sound condition ($M = 116.0$, $SD= 16.5$), $t(23) = -10.28$, $p < .001$, and the spindex TTS condition ($M = 113.4$, $SD= 16.2$), $t(23) = -6.20$, $p < .001$, with no significant difference between the search conditions.

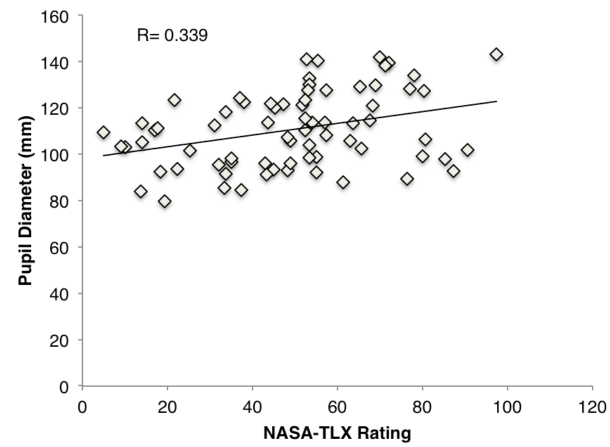

Figure 2. A scatterplot displaying PD and corresponding TLX ratings for all the participants in the three conditions.

4. DISCUSSION

The creation of these custom applications will allow for the measurement of pupillary response in future work and could easily be shared with other academics who have encountered the same issues with this software lacking in PD analysis abilities. The lack of differences between the two search conditions for pupil size was possibly due to a lack of trials since no difference was seen with TLX ratings, but not at the fault of PD as a measure of CW. While the correlation between the TLX and PD was only moderate, it does give merit to using this method for estimating CW, or using the measure along with other physiological, subjective, and performance measures in a multivariate analysis. The lack of a stronger correlation could be affected by multiple factors, particularly of interest being the issue of looking between the driving and secondary task, causing differences in luminance.

Kun, Palinko, and Razumenić [12] reviewed this topic of luminance and the obscuring effect it can have on data when measuring CW through pupillary response. They discuss that while CW can have an affect of pupillary dilation, the major contributor to the size of an individual's pupil is the pupillary light reflex (PLR). This reflex can confound data if luminance in part of a scenario is darker than others when using either a simulator or an on road study. Additionally, and particularly important in the current analysis, when participants are interacting with secondary tasks while driving their visual attention can move between the screen or road area, and the IVT. This visual movement from very separate luminance areas of outside and inside the vehicle or simulated cab could have large impacts on the pupillary response. Kun, Palinko, and Razumenić [12] discuss a possible way of addressing this issue through the use of a weighting function. Another option given by the authors is creating a scenario with minimal changes in target luminance, however this could be difficult to do when a study involves IVTs.

Although these custom applications work, it would be to our advantage to continue to make the process more efficient. The need to use multiple applications to get the data into the correct format is time consuming. Through merging the applications this would decrease the complexity of the process and hopefully make the process faster as well. The ability to input time blocks through some sort of script would also be a helpful addition to the application, as in its current form the process must be done one block at a time. Additionally the two programs used in this process run on different operating systems (Windows and Macintosh) so both types are necessary to complete the process.

Overall this expansion in abilities will allow for more measures of CW and offer variables to researchers where they were not possible before. However, before using these applications or any like them to analyze previous eye tracking data, the effect of not planning a study to measure pupillary response data should be considered due to the effects of PLR discussed above. If possible the research using this type of data and these custom applications should consider the effects of PLR and plan a way to gather the necessary data to create a weighting function. The next study investigating the effectiveness of the Tobii eye trackers and these custom applications will need to address this confounding factor to look at the relationship of pupillary response and TLX with less noise. Additionally researchers that would like access to these applications should understand the applications are not yet refined to a commercial level and the analysis remains time consuming.

While this paper does not directly discuss any standards, this is an important factor to consider when investigating pupillary response and CW. In its entirety, the measurement of driving distraction has a wide range of terms used for similar constructs and forms of measurement to estimate CW, and pupillary response within CW is no different. However, pupillary response along with some of the physiological measures of CW are still in the beginning stages of becoming a mainstream form of CW measurement due to technologies becoming more affordable and research supporting their application more widely available. As these somewhat recent measures of CW grow, authors should be able to find a way to know the correct way of referring to constructs and measuring these physiological factors. This need could be addressed through the community interested and active in this area of research to come together and decide what measures of CW will be used in which way and how the data should be gathered, organized (such as the issue with the missing frames in our data or the weighting function discussed by Kun et al., [12]), and reported. Whether the workshop on cognitive load should push this effort forward through the creation of an annual report or leave the standards papers to be written through government funded groups such as NHTSA is another decision. The decision however, must be made before it is too late so as to allow enough time to go by before the standards are released and new researchers in the area begin performing research and writing it up based on literature not addressing or using these standards.

5. ACKNOWLEDGMENTS

Thanks to Ben Stolvitz for creating the programs discussed. Portions of this work are supported by a National Science Foundation Graduate Research Fellowship (DGE-1148903).

6. REFERENCES

[1] De Waard, D. (1996). *The measurement of drivers' mental workload*. Groningen University, Traffic Research Center.

[2] Mehler, B., Reimer, B., & Coughlin, J. F. (2012). Sensitivity of Physiological Measures for Detecting Systematic Variations in Cognitive Demand From a Working Memory Task An On-Road Study Across Three Age Groups. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(3), 396-412.

[3] Engström, J., Johansson, E., & Östlund, J. 2005. Effects of Visual and Cognitive Load in Real and Simulated Motorway Driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8 (2), 97–120.

[4] Lee, Y., Lee, J. D., & Boyle, L. N. 2007. Visual Attention in Driving: The Effect of Cognitive Load and Visual Disruption. *Human Factors*, 49, 721-733.

[5] Palinko, O., Kun, A.L., Shyrokov, A., Heeman, P. 2010. *Eye Tracking Research and Applications Symposium,* 141-144.

[6] Charleton, S.G., & O'Brien, T. C. 2002. Measurement of Cognitive States in Testing and Evaluation. *Handbook of Human Factors and Evaluation,* 97-126.

[7] Ikehara, C.S., & Crosby, M.E. 2005. Assessing Cognitive Load with Physiological Sensors. *Proc. HICSS*, 295-303.

[8] Hart, S.G., & Staveland, L.E. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human Mental Workload*, by Meshkati, N., & Hancock, P.A., 239-250. Amsterdam: North Holland Press.

[9] Iqbal, S. T., Zheng, X. S., & Bailey, B. P. 2004. Task-Evoked Pupillary Response to Mental Workload in Human-Computer Interaction. *Proc. CHI'04.* Vienna: ACM 1477-1480.

[10] Mehler. B, Reimer, B., Coughlin, J. F., & Dusek, J. A. 2009. Impact of incremental increases in cognitive workload on physiological arousal and performance in young adult drivers. *Transportation Research Record: Journal of the Transportation Research Board*, *2138*(1), 6-12.

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

[12] Kun, A. L., Palinko, O., & Razumenić, I. (2012, October). Exploring the effects of size and luminance of visual targets on the pupillary light reflex. In *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (pp. 183-186). ACM.

[13] Hyönä, J., Tommola, J., & Alaja, A. M. 1995. Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *The Quarterly Journal of Experimental Psychology*, *48*(3), 598-612.

[14] Partala, T., & Surakka, V. 2003. Pupil size variation as an indication of affective processing. *International journal of human-computer studies*, *59*(1), 185-198.

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

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

[17] Recarte, M. A., & Nunes, L. M. 2003. Mental workload while driving: effects on visual search, discrimination, and decision making. *Journal of experimental psychology: Applied*, *9*(2), 119.

[18] Schwalm, M., Keinath, A., & Zimmer, H. D. 2008. Pupillometry as a method for measuring mental workload within a simulated driving task. *Human Factors for assistance and automation*, 1-13.

[19] Gable, T. M., Walker, B. N., Moses, H. R., & Chitloor, R. D. (2013, October). Advanced Auditory Cues on Mobile Phones Help Keep Drivers' Eyes on the Road. *Proceeding of the Fifth International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, Eindhoven, The Netherlands.