A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions

Bastian Pfleging[1,3], Drea K. Fekety[2], Albrecht Schmidt[3], Andrew L. Kun[4]
[1]University of Munich (LMU), Munich, Germany
[2]Clemson University, Clemson, SC, USA
[3]University of Stuttgart, Stuttgart, Germany
[4]University of New Hampshire, Durham, NH, USA
bastian.pfleging@ifi.lmu.de, dfekety@g.clemson.edu,
albrecht.schmidt@vis.uni-stuttgart.de, andrew.kun@unh.edu

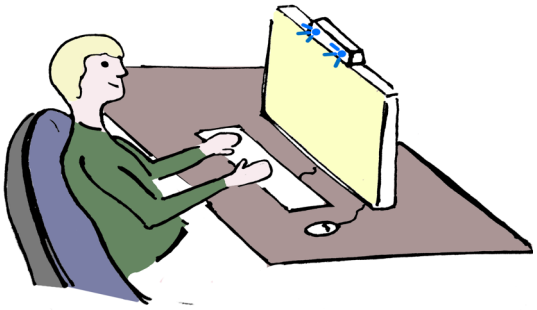
ABSTRACT
In this paper, we present a proof-of-concept approach to estimating mental workload by measuring the user's pupil diameter under various controlled lighting conditions. Knowing the user's mental workload is desirable for many application scenarios, ranging from driving a car, to adaptive workplace setups. Typically, physiological sensors allow inferring mental workload, but these sensors might be rather uncomfortable to wear.

Measuring pupil diameter through remote eye-tracking instead is an unobtrusive method. However, a practical eye-tracking-based system must also account for pupil changes due to variable lighting conditions. Based on the results of a study with tasks of varying mental demand and six different lighting conditions, we built a simple model that is able to infer the workload independently of the lighting condition in 75 % of the tested conditions.

Author Keywords
Estimation of mental workload; eye-tracking; task-evoked pupillary response; cognitive workload; compensation of pupillary light reflex; adaptive user interfaces; psychophysiology; lighting.

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces;

INTRODUCTION
A large body of literature exists on the topic of estimating a person's mental workload while engaged in cognitively-demanding tasks. Although mental workload is not something that can be measured directly, in recent years many different technologies have been used to gather data that can accurately infer a person's mental workload. Measurements in this area can be in the form of pupil diameter (collected from remote as well as head-worn eye-trackers), electroencephalography (EEG), heart rate (HR), heart rate variability (HRV), skin conductance (GSR), skin temperature, and respiration rate, to name a few. These methods have their strengths and weaknesses, but all have been used in some capacity to estimate how a person uses mental resources to process information and where a person's upper limits of cognitive capabilities lie. One important advantage of remote eye-trackers in comparison to other established physiological



	Light (large impact)		Task (small impact)	
	low	high	difficult	easy
Pupil diameter	↗	↘	↗	↘

Model: $PD = PD_{light} + PD_{task}$

$PD_{task} = PD - PD_{light}$

Estimate: $mental\ workload \approx f(PD_{task})$

Figure 1: For many applications (e.g., online learning), knowing the user's workload is beneficial for instance to adapt the interface. With a model extracted from our study, we are able to estimate mental workload based on pupil diameter retrieved through remote eye-tracking under different controlled lighting.

CHI'16, May 07 - 12, 2016, San Jose, CA, USA

ACM 978-1-4503-3362-7/16/05...$15.00
DOI: http://dx.doi.org/10.1145/2858036.2858117

measurements of mental workload is that they are rather unobtrusive and do not require the user to be tethered to measurement equipment.

The study outlined in this paper uses pupil diameter data to estimate mental workload (Figure 1). A large number of studies have confirmed that, in an effect called the task-evoked pupillary response (TEPR), pupil diameter increases when the cognitive difficulty of a task increases [3]. This change in pupil diameter occurs rapidly, both when the pupil dilates and when it contracts, making eye trackers attractive in efforts to estimate mental workload. However, pupil diameter also depends strongly on lighting conditions: the pupillary light reflex constricts the pupil in response to increased light levels and vice versa. Pupil size can change by several millimeters due to the pupillary light reflex. In contrast, changes related to mental workload are an order of magnitude smaller–usually between 0.1 mm and 0.5 mm [3]. These small changes can be overwhelmed by the large changes due to the pupillary light reflex. This work addresses this issue by modeling the effects of TEPR and the pupillary light reflex on pupil size, and by using this model to separate the two effects.

We propose a simple model as a proof of concept that we extracted from our experiment data on pupil size changes under different known levels of light reaching the eye, and under different levels of task difficulty. We then use this model to estimate pupil diameter changes occurring as a result of cognitive load, and ultimately to estimate the task difficulty that resulted in the measured pupil diameter. In this work we constrain the model to cases when both the light reaching the eye and the mental task difficulty are constant during a certain condition.

The pupillary light reflex will respond to both ambient light (such as room illumination) and visual target luminance (such as the light emitted from a computer screen). We demonstrate these effects in the different lighting conditions of the experiment.

CONTRIBUTION STATEMENT

The contribution of this paper is an investigation of the combined influence of light and mental workload on the user's pupil size. Previous studies investigating eye-tracking and mental workload often fixed lighting conditions to one level in order to eliminate the effects of the pupillary light reflex. Instead, we conducted a study in which users performed different tasks and were exposed to six different (and controlled) lighting conditions. We used the results of this study to create a simple model of pupil diameter change that incorporates both the pupillary light reflex (for our controlled conditions), and the pupil diameter change that is due to changes in mental workload. We see this model as a first step towards pupil-based illumination-independent workload estimation.

Estimating the user's workload is helpful for many situations where people interact with computing devices or machines. One example is the automotive domain where it is important not to overload the driver as they interact with in-vehicle devices [28] or communicate with remote conversants [21, 38], since performance degrades with increasing workload [6]. With automated driving modes and non-driving-related activities [37], workload estimation can be useful to track and support the driver's reengagement in the driving task [20]. In a desktop setup, one application domain would be online learning where the estimation of mental workload is used to adapt the content and/or interface to improve the learner's performance.

RELATED WORK

Measuring the user's mental workload has been the subject of various studies.

Eye-Tracking Data Used as Indices of Mental Workload

Eye gaze data has been established as an effective method for measuring mental workload in response to a cognitively demanding task, by focusing on certain parameters of autonomically-driven eye behavior. Pupils tend to dilate in response to greater mental workload (often called the task-evoked pupillary response). A number of early studies supported this relationship between pupillary changes and workload, and in a way established the foundation for this approach to psychophysiology [15, 16]. A dated but thorough review of the literature surrounding task-evoked pupillary responses outlined the relationship between changes in pupil size and mental workload [3].

Other early research with eye-tracking indicated a greater time interval between blinks for a visual identification task compared to a visual detection task [8]. Blink duration and frequency were also found to decrease in response to increased difficulty of a visual identification game [46]. Nunes & Recarte [32] found increased pupil diameter and higher blink rates when participants were engaged in more difficult secondary tasks while driving an automobile. Ahlstrom & Friedman-Berg [1] found that air traffic controllers exhibited greater pupil diameter and shorter blink duration in response to increased task difficulty. In response to performing an auditory workload task while driving, Tsai et al. [45] found that participants' blink frequencies and pupil diameters significantly increased, though measures of blink duration did not reflect these results. Benedetto and colleagues [4] also recently found that increased dual-task difficulty was accompanied by shorter blink duration and longer duration between blinks in a simulated driving environment, though the former provided more reliable data. The contextual qualities of these selected studies suggest that inferences about mental workload indicated by eye-tracking data are robust and not necessarily domain-specific.

A key issue with using eye-tracking data to infer mental workload is that physiological changes in the eye are influenced both by lighting conditions and the difficulty of the task a person is engaged in. In creating algorithms to predict eye gaze changes based on mental workload, researchers have to parse out the effects of changes in lighting and mostly

do so by choosing constant lighting conditions for experiments. A series of recent preliminary studies have addressed this issue in further detail, in the context of a simulated driving environment [23, 34, 35]. Methods have also been developed to gather eye gaze data in sub-optimal lighting conditions [49, 50]. Furthermore, Marshall proposed the Index of Cognitive Activity, which attempts to estimate cognitive load [24] independent of lighting, based on rapid fluctuations of pupil size. However, Marshall's proprietary algorithm by default outputs a sequence of estimates once a second, which can obscure interesting changes in cognitive activity that occur at a quicker pace, such as those observed in spoken interactions [14, 22, 36]. Using TEPR and compensating for lighting conditions should allow us to detect such changes.

Presentation Modalities of Workload Tasks

In many studies where eye-tracking data were used as indices of mental workload, the task used to induce workload was often a visually-based stimulus [7]. This highlights the need to identify and isolate the separate influences of lighting and workload on gaze data. One potential solution to this issue could be to introduce an auditory-based workload task and have participants focus on a single point in a visually neutral or uninteresting stimulus [18]. Ideally, this would allow for the measurement of gaze data in response to a cognitively-demanding task without the interference of changing lighting conditions.

Tasks used to induce mental workload, such as the N-back test, delayed digit recall test, or the PASAT [12, 17, 27] can be presented as either a visual stimulus or as an auditory stimulus. The auditory N-back task has been successfully used to induce workload, with different physiological measures (e.g., [25, 39]). This protocol has been used in a number of studies in the realm of automobile driving. Mehler and colleagues found that heart rate and skin conductance were reliably indicative of driver workload estimates in simulated driving tasks [25, 26]. In an on-road driving study they also found similar changes in heart rate and eye gaze measurements with respect to task difficulty [48]. The same research group showed that pupil diameter can be used to estimate workload for drivers engaged in the N-back task [47]. More recently Gable et al. found evidence that in driving simulator studies pupil diameter might be a more sensitive measure of cognitive load changes than heart rate [9].

Klingner et al. [19] found that the difficulty of a task may change between different perceptual modalities. However, the peer-reviewed literature has yet to fully explore the differences between these modalities. For our study, we chose to use an auditory presentation, working memory mental workload task to minimize the effects of extraneous visual stimuli which may influence eye-tracking measurements.

Other Physiological Measures of Mental Workload

Physiological measures such as functional near-infrared spectroscopy (fNIRS) [11, 44], electroencephalography EEG [5, 10, 33, 42], heart rate [31, 40] and heart rate variability [31, 41], and skin conductance [13, 43], have also been used to track mental workload while engaged in a cognitively demanding task. One advantage of using these physiological measures instead of eye-trackers is that they are not dependent upon the lighting conditions of the environmental configuration used for data collection. However, these methods can be rather obtrusive since they require the user to wear or connect sensor technology and also reveal health-relevant personal data. Furthermore, these types of solutions can invoke unnatural participant behaviors which confound the data (e.g., jaw clenching with EEG). Remote eye-tracking is advantageous in this domain as it can maintain reliable recordings while alleviating participant discomfort, which could otherwise negatively influence experimenters' data yield.

Summary: Measuring Mental Workload

As outlined in this section, various methods have already been used to measure and estimate mental workload. With regard to unobtrusiveness, measuring the pupil diameter through a (remote) eye-tracker seems to be one of the most promising approaches. In contrast, many of the approaches that employ physiological sensors require the user to wear specific sensor hardware.

For eye-tracking, one drawback is the concurrent influence of the pupillary light reflex and the task-evoked pupillary reflex on pupil diameter. That is why previous work often required strongly controlled lighting setups. With the experiment presented in this paper, we want to broaden the flexibility of this approach by investigating different lighting conditions and generating a model that allows for workload estimation in different cases.

EXPERIMENT: ANALYZING THE INFLUENCE OF LIGHT AND WORKLOAD ON PUPIL DIAMETER

Our study explores how we can separate the effects of light and mental workload on pupil diameter. We therefore conducted an experiment where we explored the combined effect of lighting (variable ambient light or visual target luminance) and cognitive load on pupil diameter.

Variable Light Conditions

We used six different lighting conditions in order to demonstrate pupil changes in typical situations of varying ambient light (such as differences from office lighting); and target illumination (such as screen content). The six lighting conditions consisted of three conditions (part 1) where the environmental illumination was varied using different lamps (1 lamp, 2 lamps, 3 lamps) while keeping the target constant, and three conditions (part 2) where the target illumination (screen content) was varied (25%, 50%, 75 % grey) while keeping the environmental illumination constant (3 lamps).

Auditory Delayed Digit Recall Task

In the experiment we utilized an auditory delayed digit recall task [27] as the primary task to induce workload. For each trial of the delayed digit recall task, participants heard a randomized set of 20 digits ranging between 0 and 9 spoken by a computerized voice, with a 1.5-second interval between each spoken number. In total, each trial was 30 seconds long.

You hear:	6	5	2	7	4	4
You say (0-back):	6	5	2	7	4	4
You say (1-back):		6	5	2	7	4
You say (2-back):			6	5	2	7

Table 1: Example numbers spoken by a computer voice, followed by correct participant responses for the three task difficulties of the auditory delayed digit recall task. Blank cells indicate that the correct response for the participant is to say nothing.

Our pilot testing revealed this rate of presentation to be optimal for participants engaged in our workload test (i.e., no secondary tasks) after sufficient practice. For each trial, participants were instructed to verbally repeat the number that they heard N numbers ago. The number N corresponded to the difficulty level of the task they were currently performing, i.e. 0-, 1-, or 2-back. Table 1 shows examples of the tasks performed. The same delayed digit recall tasks were employed throughout the whole experiment, and numbers were randomly generated for each trial.

Participants were allowed one minute of resting time between the end of one trial and the start of the next trial. This was done to allow for a recovery to a baseline state of workload before the next trial. In situations where lighting conditions changed between the end of one trial and the start of the next trial, the experimenters allowed a 2.5-minute recovery period in order to allow the participants' eyes to appropriately adjust to the new lighting conditions.

Design

The experiment was designed as a within-subject experiment. Participants completed three trials with different levels of difficulty (0-back, 1-back, 2-back) for each of the six lighting conditions. This results in a total of 3 x 6 = 18 trials with level of difficulty (3 levels) and lighting (6 levels) as independent variables. Given that the pauses between tasks of different difficulty were between one minute and 2.5 minutes, we did not expect any need to counterbalance the order of task presentation. After all, prior research shows that after the end of a cognitively-demanding task, pupil diameter returns to its pre-task level within a few seconds (see for instance [3, 22]). Nevertheless, we introduced two task orders, such that in each lighting condition participants started with the 0-back task, while the order of the next two tasks (1-back and 2-back) was counterbalanced between participants.

As dependent variables, we recorded workload task performance (percentage of correct responses), pupil diameter (as well as other eye-tracking measurements), and eye behavior (blinks, saccades, and fixations). Due to time restrictions, subjective workload ratings (NASA TLX) were collected only at the end of the experiments to get a summative rating for each level of difficulty.

Figure 2: Experiment setup. The eye-tracker was placed between participant and wall/ TV; the computers to log all measurements were located out of the participant's sight next to the cubicle. Lamps as illumination sources were placed on tripods behind the participants about 40 cm above the participants' heads. This figure shows the configuration as it was used for part 1 of the experiment.

Figure 3: In part 2 of the study participants sat in the 3-walled white 'cubicle' with the LCD TV in front of them.

Participants

In total, 24 participants (19 male, 5 female) aged between 19 and 42 (M = 25.00 years, SD = 6.16 years) took part in this study. Participants were recruited from students and faculty members of the University of Stuttgart. All participants were compensated for their time by receiving a small give-away. Student recruits also received course credit for their participation in the study. Eye tracking data from 3 participants were excluded from our analysis due to technical difficulties. We also excluded data from one other participant because he did not follow instructions in completing the delayed digit recall task. Thus, our sample size is N=20.

Apparatus and Data Collection

The study took place in a windowless room. The participant's seat was placed in a white-walled "cubicle" in which three (temporary) poster walls surrounded the participant.

These walls were covered with large white sheets of paper (see Figure 2). The participants' eye movement and pupil data were sourced from an SMI RED250 eye-tracker[1] and recorded using the iView X software from SMI.

The lighting intensity variable for part 1 of the experiment was manipulated such that the participant's cubicle was illuminated by 1, 2, or 3 professional-grade studio lamps (Product number 400894 from TecTake.de) as only sources of illumination with 55-watt white fluorescent bulbs. The TV as gaze target was covered with white paper in this part 1 of the study to have a homogeneous, constant target.

For part 2 of the experiment, we used a 55-inch FullHD (1920x1080 pixels) Philips 55PFL7606/K02 television (see Figure 3 for equipment setup) displaying a full-screen view of one of three gray images with different brightness. The major device settings were: Contrast: 90, Brightness 76, Color 67, Sharpness 5, Color Temperature Cold, Dynamic Contrast: Minimum, Light Sensor: Off, Ambilight: Off. In this part, the environmental illumination was kept constant (all three lamps on).

The lamps and the TV screen were the only light sources in the windowless room. At the beginning of each trial, we measured the amount of light reaching the participant's eyes with a lux meter (iClever digital lux meter LX1330B[2]) placed at the participant's forehead. The amount of illumination reaching the participant's eye (in lux) for the three different levels of environmental illumination (part 1) were M=133.50, SD=5.45 (1 lamp); M=247.55, SD=23.55 (2 lamps); M=387.14, SD=22.81 (3 lamps). The amount of illumination reaching the participant's eye (in lux) for the three levels of different target brightness (part 2) were M=255.19, SD=31.51 (25%); M=308.10, SD=28.69 (50%); M=364.76, SD=35.07 (75%).

Procedure

When the participants arrived at our lab, we first introduced them to the topic and procedure of our experiment. Next, the participants read and signed the consent forms and filled out an initial demographic questionnaire. The participants were then seated in the experiment room. Participants remained seated for the duration of the experimental session, and were asked to visually focus on a small crosshair at a centrally located position in front of them.

Before the start of the experiment, we introduced the participants to the cognitive workload task (the auditory delayed digit recall test). Participants were given up to 12 practice sessions of non-scored trials at each level of task difficulty in order to fully familiarize themselves with the different difficulty levels of the task. Also, we presented them an empty copy of the NASA TLX questionnaire as a preparation for those sheets that should be filled out at the end of the experiment. Afterwards, the eye-tracker was calibrated using a nine-point calibration method that was provided with the eye-tracking software.

As explained before, the experiment itself was then conducted in two parts. The difference between the two parts was the way in which the participant experienced lighting conditions. We chose to split the experiment into these parts, since technical reasons prevented us from switching easily between these setups. The order between the two experiment parts was counter-balanced such that half of the participants started with part one and the other half with part two.

For each scored delayed digit recall trial in both parts, we asked participants to focus on a small "+"-shaped target located on the white board / TV set seen in front of the participant. This was done to minimize the possibility of eye movements interfering with eye-tracking data while the participant performed the task.

Part 1: Controlling Environmental Illumination

In this part, we controlled the amount of environmental illumination that reaches the participant's viewing perspective. We manipulated ambient room illumination such that the influence of illumination and workload on pupil diameter could be examined in different lighting conditions.

Part 2: Controlling Large Field-of-View Stimulus Brightness

In this part, we controlled the intensity of target luminance the participant was viewing. Participants were seated in the same location (with illumination provided by all three lamps at the same time) and viewed a homogeneous single-color gray image on the TV screen while engaged in the same delayed digit recall test. These gray images were created such that they represented 25%, 50%, and 75% brightness relative to a full-white color image. The TV screen covered much of the participants' field of view (see Figure 3). All other factors between the two experiments were kept constant.

Both independent variables were manipulated in the same order as in experiment part 1. Because all data were collected from participants in the same session, participants experienced the conditions of experiment part 2 either immediately before or immediately after experiment part 1 (depending on the condition they were assigned).

At the end of the experiment, the participants were asked to fill out three NASA TLX questionnaires, one for each level of the cognitive workload task.

DATA SET

We compiled the recordings made during the experiment into a data set that is publicly available under an open data license. Details about the full dataset are available at http://www.hcilab.org/research/workload-pupil-diameter/.

Apart from the analysis in this paper, we believe that this data set will be helpful for the research community in various purposes. Especially, it will be helpful towards the development

[1] http://www.smivision.com, last access 2016-01-07

[2] http://goo.gl/qNCdNB, last access 2016-01-07

		Part 1: Environmental Illumination								
		1 lamp ("low")			2 lamps ("medium")			3 lamps ("high")		
Metric	**Units**	**0-back**	**1-back**	**2-back**	**0-back**	**1-back**	**2-back**	**0-back**	**1-back**	**2-back**
Left pupil diameter	**mm**	3.889 (0.492)	4.044 (0.483)	4.210 (0.578)	3.406 (0.382)	3.610 (0.416)	3.742 (0.478)	3.139 (0.317)	3.206 (0.313)	3.328 (0.367)

Table 2: Means and standard deviations of pupil diameter measurements of workload under each condition in Part 1

		Part 2: Stimulus Brightness								
		25% ("low")			50% ("medium")			75% ("high")		
Metric	**Units**	**0-back**	**1-back**	**2-back**	**0-back**	**1-back**	**2-back**	**0-back**	**1-back**	**2-back**
Left pupil diameter	**mm**	3.719 (0.434)	3.947 (0.482)	4.100 (0.467)	3.086 (0.298)	3.203 (0.342)	3.318 (0.321)	2.903 (0.283)	3.010 (0.316)	3.099 (0.313)

Table 3: Means and standard deviations of pupil diameter measurements of workload under each condition in Part 2

of generalizable models that allow for workload detection based on eye-tracking data independent from the current illumination situation.

Participants performed a total of 18 delayed digit recall tasks across the two experiment parts, and a counter-balanced design controlled the order in which participants experienced the conditions of each trial (i.e., some combination of lighting conditions, and task difficulty). While we excluded data for four of the 24 participants from the dataset, and subsequently used data from 20 participants, the loss of data did not skew our counter-balanced design; approximately equal sample sizes were maintained between groups. Additionally, main effects of trial order on task performance were not present. Means and standard deviations for pupil diameters for all participants can be found in Table 2 and Table 3.

Eye-Tracking Data

The eye-tracking data was recorded using the SMI RED 250 remote eye-tracker. This time-stamped data was recorded at 120 Hz. The measurements include all available information the SMI software provided. This includes pupil diameter for both eyes, and gaze position (on the screen). Also, details about the current eye behavior, i.e., blinks, saccades, and fixations were recorded and added using the SMI-internal software using default parameters for the event detection.

Hand-Coded Data

We collected NASA-TLX scores, the participants' error rates on the delayed digit recall task, and measurements of the amount of light reaching the participants' head in each lighting condition.

NASA TLX

After both experiments, the participants were asked to also complete NASA TLX questionnaires. Due to time restrictions of the experiment, we could not ask for subjective ratings after each condition. In order to get at least some subjective feedback, we did an overall assessment for each of the levels of the digit recall task (rating across all conditions with the corresponding level).

Participant Performance / Error Rates

We hand coded the participants' responses to the delayed digit recall task on paper. We later transcribed these scores and stored them in electronic form.

Illumination Measurements

Whenever the lighting conditions were changed we measured and recorded the illumination (in lux) reaching the participant's eye.

Pupil Diameter Data Preprocessing

For each of the 20 participants we found the average pupil diameter for each of the experimental condition. Given that the participants experienced 3 task difficulty levels at each of the 6 lighting levels, we calculated 20 x 3 x 6 = 360 pupil diameter averages.

To calculate the pupil diameter averages we started with the raw measurements from the SMI software. Since some of the data from the SMI software was provided with non-uniform sampling rates, we first used the "interp1" Matlab function to resample the raw measurements to attain a uniform sampling rate of 120 Hz. Next, we excluded samples for which the raw measurement values remained constant for at least 100 ms, as these samples were generated when the eye tracker failed to accurately track pupil diameter. Finally, we plotted pupil diameter changes over the 30 second periods delayed of the digit recall tasks for each of the 360 measurements. By visual inspection we identified three cases where the pupil diameter data was noisy for the majority of the experimental condition, and we rejected this data. We then calculated averages for 360 – 3 = 357 cases. Finally, we used a single imputation method [2] to fill in the remaining 3 averages: in our method the missing average is set equal to an average calculated for the same participant, and same lighting conditions, but a different task difficulty. Note that this method is conservative in that it does not artificially improve our chances of confirming our expectation that pupil diameter will increase with increased task difficulty.

DATA ANALYSIS AND DISCUSSION

Lighting Conditions

We analyzed whether the settings chosen for our lights created measurable differences in the stimulus and room lighting. At the same time, we analyzed potential order effects related to the counter-balanced lighting conditions. A one-way ANOVA comparing the lighting conditions of each trial to the lighting condition order showed no significant differences in lighting conditions based on order effects ($p>.05$).

For the purposes of this analysis, we treated all 6 lighting conditions as levels in one variable. Consequently, we compared the average illumination reaching the participants' eyes between all 6 lighting conditions using a one-way repeated measures ANOVA.

Mauchly's test failed to detect a violation of the sphericity assumption in our illumination data, $\chi^2(14)=14.672$, $p>.05$, therefore no corrections to degrees of freedom are needed. The results indicate a significant main effect of lighting manipulations on illumination reaching the eye, $F(5,100)=461.987$, $p<.001$, $\eta^2=.959$. Least Significant Difference (LSD) post-hoc analyses within lighting conditions revealed significant differences ($p<.001$) in average illumination reaching the eye between all levels, except for the "2 lamps" vs. "50%" conditions ($p>.05$).

Task Performance

Participants' performance on each delayed digit recall task was measured in terms of percentage of correct responses. Percentage correct was used instead of a measure of frequency of correct responses, because the 3 different difficulty levels of the task are designed in such a way that yields a different number of responses from the participant. In other words, the 2-back task (if performed correctly) yields fewer responses from the participant compared to the 1-back and 0-back tasks, because the participant has no response for the first two numbers they hear. This effectively means that 0-back tasks produce 20 responses from participants, where 1-back tasks produce 19 and 2-back tasks produce 18 responses.

A one-way ANOVA revealed statistically significant differences in average task performance among difficulty levels, $F(2,63)=27.300$, $p<.001$, $\eta^2=.464$. A Least Significant Difference (LSD) post-hoc analysis showed statistically significant differences between all pairwise comparisons of difficulty level except for the 0-back-to-1-back comparison ($p=.247$). This suggests our participants were able to perform equally well on the easiest and middle-difficulty tasks featured here. However, it is important to remember that the focus of this paper is estimating users' workload and not their ability to perform tasks. Thus our measurements of TEPR inferring workload should have more value to our model than task performance in this situation.

Subjective Workload

Subjective workload was measured after the end of the experiment, where participants were asked to fill out a NASA-TLX form for each task difficulty level. Each subjective workload rating (one per difficulty level) represents the participants' average workload from 6 trials of that difficulty level. Unweighted subjective workload scores for each participant were calculated by averaging participants' ratings on a scale of 1-20 over the six dimensions of the NASA-TLX questionnaire [29].

A one-way ANOVA revealed statistically significant differences in average subjective workload among task difficulty levels, $F(2,63)=94.660$, $p<.05$, $\eta^2=.750$. A Least Significant Difference (LSD) post-hoc analysis showed statistically significant differences between all pairwise comparisons of difficulty level such that 0-back<1-back<2-back, $p<.001$.

There are moderate, negative correlations between participants' task performance (percentage of correct responses) and their subjective ratings of workload (NASA-TLX) for the 1-back task ($r=-0.445$, $p<.05$) and the 2-back task ($r=-0.466$, $p<.05$). In other words, as participants performed better on the workload task, they felt that the task was less workload-inducing. However, a weak positive correlation exists between task performance and subjective workload for the 0-back task, although this relationship is not significant ($r=0.154$, $p>.05$).

Eye-Tracking: Pupil Diameter

We conducted a 6 (lighting level) x 3 (task difficulty) repeated measures ANOVA to examine the separate and combined influences of these factors on pupil diameter. For the remainder of this section, we only consider the left eye's pupil diameter because we observed fewer inaccurate/missing data points in this eye throughout our data set. Mauchly's tests detected a violation of the sphericity assumption in our pupil data for the main effect of lighting ($\chi^2(14)=63.728$, $p<.001$, $\varepsilon=.478$), the main effect of task difficulty ($\chi^2(2)=11.162$, $p<.05$, $\varepsilon=.684$), and the interaction between lighting and task difficulty ($\chi^2(54)=107.262$, $p<.001$, $\varepsilon=.436$). Therefore, Greenhouse-Geisser corrections are reported here. We observed a significant main effect of lighting level on pupil diameter, $F(2.391,45.424)=167.143$, $p<.001$, $\eta^2=.898$. Post-hoc LSD comparisons revealed significant differences in pupil diameter among all lighting levels ($p<.001$), except for the comparisons between "1 lamp" vs. "25%" and "3 lamps" vs. "50%" ($p>.05$). This omnibus test also revealed a significant main effect of task difficulty on pupil diameter, $F(1.368,25.990)=71.903$, $p<.001$, $\eta^2=.791$. Further, post-hoc LSD comparisons showed significant differences in pupil diameter among all task difficulty levels, $p<.001$.

Another result observed within this omnibus test was the significant interaction between lighting level and task difficulty on pupil diameter, $F(4.356,82.767)=3.633$, $p<.05$, $\eta^2=.161$. Post-hoc LSD comparisons of pupil diameter also revealed significant differences among all task difficulty levels within each of the 6 lighting conditions ($p<.05$).

TOWARDS A MODEL FOR WORKLOAD ESTIMATION

The eye-tracking results indicate that average pupil diameter increases with increased task difficulty, and that we can detect this under all lighting conditions. These results are also in agreement with the performance results and the subjective workload ratings.

But how well can individual measurements of pupil diameter be used to assess workload? To answer this question we first plotted all of the pupil diameter measurements (18 data points for 20 participants) as shown in Figure 4. On the x-axis of Figure 4 we vary the task difficulty – this results in 3 groups of data points, one for each delayed digit recall task. Within each group we also vary the lighting condition – this results in the six columns of measurements, one for each lighting condition.

Two trends are visible in the data in Figure 4. First, the average pupil diameter clearly changes with light. Second, the average pupil diameter also increases with task difficulty. However, the data shows that, if we want to use individual pupil diameter measurements to assess the user's current workload we have to address two problems. First, we must account for lighting conditions. As the data in Figure 4 demonstrates, the same pupil diameter can be the result of different workloads under different lighting conditions. Second, pupil diameter measurements are noisy. Thus, even if we know the lighting conditions we might not be able to always correctly identify workload.

To address these issues we introduce a model that serves as a proof-of-concept for the idea that if we know the lighting conditions then we can use pupil diameter values to identify the task difficulty. Thus, for each participant p, we model pupil diameter PD_p as a sum of two contributing factors:

$$PD_p = PD_{p,light} + PD_{p,task}$$

Equation 1 Modelling pupil diameter.

In Equation 1, $PD_{p,light}$ is the pupil diameter for participant p given lighting condition *light*, while $PD_{p,task}$ is the normalized pupil diameter for that participant's pupil diameter given task difficulty *task*.

For a given level of *light* we calculate $PD_{p,light}$ as the average pupil diameter for all three levels of *task*. Expressing $PD_{p,task}$ as the difference between PD_p and $PD_{p,light}$, we get the normalized pupil diameter data presented in Figure 5.

Next, we classify the normalized pupil diameter readings as indicating one of the three task difficulties. We perform the classification by comparing the normalized pupil diameter to the lower and upper limit of the 1-back region (shown as blue lines in Figure 5): pupil diameters below the lower limit are classified as indicating a 0-back task, diameters between the limits as indicating a 1-back task, and those above the upper limit as indicating a 2-back task. In this paper we identify the upper and lower limits using a simple heuristic approach: we first find the mean of the normalized pupil diameter values

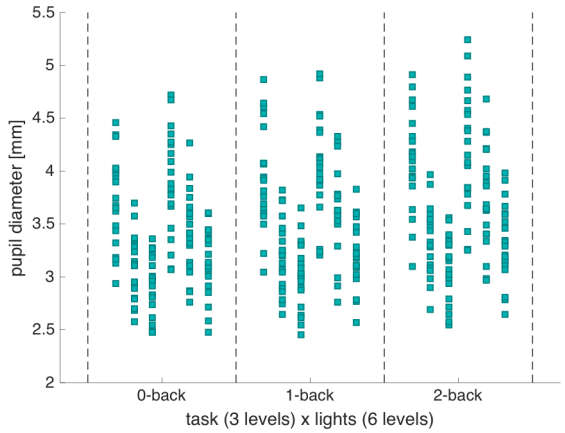


Figure 4: Pupil diameter measurements. Each data point shows the average pupil diameter (y-axis, in mm) of one participant during one 30 second-long trial. The trials on the x-axis are grouped by task (0-, 1-, and 2-back) and lighting (1, 2, and 3 lamps, and 25%, 50%, and 75% brightness).

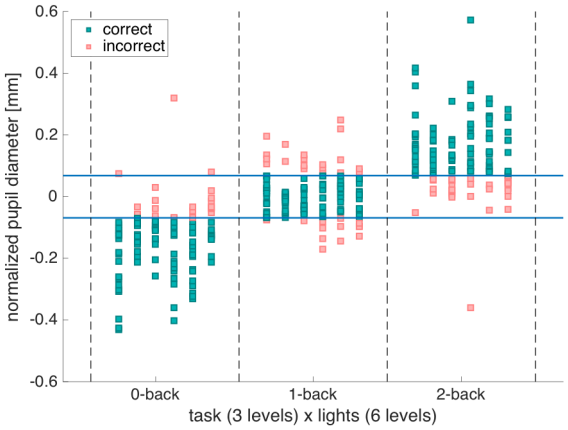


Figure 5: Using Equation 1, we can calculate the normalized pupil diameter by subtracting the participant's average pupil diameter for a given level of light ($PD_{p,light}$) from the pupil diameter of the trial. This diagram shows the normalized pupil diameter (y-axis) using the pupil diameters shown in Figure 4, for all participants during the different trials (x-axis) separated by task and illumination. Blue lines indicate the thresholds we used to classify the underlying workload: data above the top line was classified as workload due to 2-back task, between the two lines as due to 1-back, and below bottom line as due to 0-back. For correctly classified workloads we marked the data points in green, and for incorrectly classified workloads we marked them in red.

for the 1-back task, and set the upper and lower limits one standard deviation away from this mean. We visually demonstrate the results of this approach in Figure 5. Here we use all of the 360 data points to calculate the limits, and then apply the limits to the same 360 data points, and the correctly classified data points are shown in green. We evaluated this approach using k-fold cross-validation. With k = 4, we get

result of 75% correct classification. In our k-fold cross validation we found the threshold values based on data from ¾ of the participants (15 of 20), and tested them with data from ¼ of the participants (5 of 20).

GENERAL DISCUSSION

In this study we explored the combined effects of light and cognitive load on pupil diameter and we proposed a model that allows us to estimate task difficulty based on pupil diameter measurements. Having such an estimate of task difficulty would be useful in a number of situations. In software environments it might allow us to identify when users are confused or overwhelmed, which in turn could trigger a change in the interaction approach for the user interface. In learning environments, the estimate of task difficulty might help us find the optimal way to present the material based on the user's current state.

Our results are encouraging. We explored the combined effect of lighting and task difficulty on pupil diameter for indoor lighting conditions commonly found in offices and homes, and for plausible task difficulty levels. We found that we must account for the effects of light if we intend to use pupil diameter as a measure of task difficulty. This is demonstrated in Figure 6, which shows pupil diameter measurements for one participant at two different lighting levels: 25% gray scale brightness, and 2 lamps. Note that the participant's average pupil diameter is 3.484 mm when engaged the 0-back task and viewing the screen at 25% gray scale brightness, and it is 3.496 mm when engaged in the 1-back task with 2 lamps as the light source. In practice the two pupil diameters are indistinguishable from each other. This situation demonstrates that in general we cannot estimate task difficulty from pupil diameter without separating the effects of task difficulty and light.

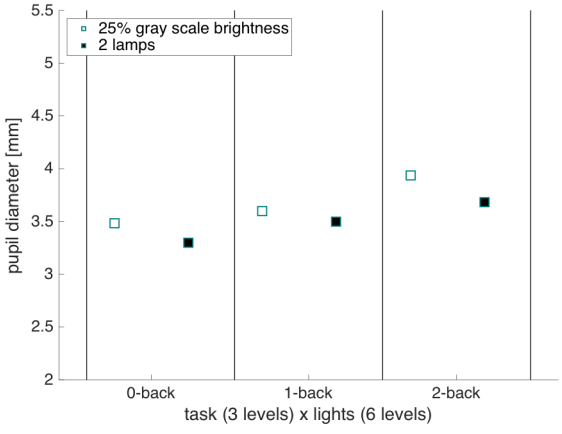


Figure 6: Pupil diameter data for one participant for two different lighting conditions. Note that raw pupil diameter measurements result in indistinguishable results for the 0-back task performed while looking at the LCD screen set to 25% gray scale brightness, and for the 1-back task performed with 2 lamps.

To be able to separate these effects we introduce a simple model, which treats the effects of light and cognitive load as additive parts of the overall pupil diameter. We show that we can use the model to classify pupil diameter readings according to the underlying task difficulty if we have two pieces of information available: the current lighting level and the average size of the user's pupil for that lighting level. Our results demonstrate that it is possible to use the model to estimate the difficulty of the underlying task with about 75% accuracy. As an example, if we apply our model to the data in Figure 6 the resulting normalized pupil diameters are -0.188 mm and 0.005 mm, respectively. Our model correctly classifies these values as resulting from 0-back and 1-back task difficulty, respectively.

Our simple model builds on the work of Palinko and Kun [34], in which they proposed a model of pupil diameter that incorporated both the effects of visual target luminance and task difficulty. However, the parameters of their model [34] were manually derived to demonstrate the idea of separating the effects of light and task difficulty for a small portion of the data collected in that experiment. There was no attempt to systematically evaluate their model's performance using all of the data from the experiment. In contrast, the model presented in this paper was created and evaluated using the entire dataset collected in this experiment, allowing us to quantify what we might expect to achieve in estimating task difficulty based on pupil diameter measurements.

The results also provide indication that the system might be able to use training data from existing users in order to accommodate new users. Specifically, in our k-fold validation we found thresholds for our model based on data from 15 of the 20 participants, and tested the model on data from the remaining 5 participants. With this approach we were able to correctly classify the workload based on pupil diameter measurement in 75% of the cases.

Using the current model as input for adaptive UIs could be challenging since this accuracy could (negatively) impact the user in case of misclassifications. Thus, further work is required to increase accuracy and investigate the required level of accuracy. However, our approach cannot only be used for UI adaptation, but also as a tool to highlight challenges of the interaction during UI design, and provide ideas how/what to modify. For UI adaptation as one of the long-term goals we see our approach as one of the steps towards this goal but agree that further aspects need to be considered (e.g., [30]).

Finally, our data also demonstrates that using pupil diameter measurements to estimate task difficulty is complicated by measurement noise. Noise can result from eye blinks, head motion, and other natural user actions and environmental conditions. We can expect such noisy measurements to occur with remote eye trackers, although the situation might improve if we use head-worn eye trackers. Head-worn eye trackers might become widespread with the advent of devices similar to Google Glass, or the upcoming Microsoft

HoloLens (although neither of these two devices has eye tracking capabilities at this time).

PRACTICALLY ESTIMATING WORKLOAD

Looking ahead, based on the findings presented in this paper it becomes possible to continuously estimate workload for new users and new tasks. To apply our model in an office or home environment, for example to track a user's task difficulty while coding, or playing a game, we would need three types of information: the user's pupil diameter, the current lighting conditions, and the average pupil diameter for those lighting conditions.

Pupil diameter measurement will likely become inexpensive soon. Even today we can purchase eye trackers at about $200, and we can imagine that in the near future eye tracking applications will be able to utilize a high-resolution webcam (and infrared illumination) integrated into a computing device. And, at least for applications where the user's visual attention is focused on a screen, we should be able to easily assess lighting conditions: e.g., for a user writing code on a computer we can track the user's gaze, and assess the luminance of the screen around the user's visual focus as well as the environmental illumination. We can also assess ambient light – mobile devices can already measure the intensity of ambient light to control their display luminance. Finally, assessing the average pupil diameter for a given lighting level would require a calibration process: this might be as simple as instructing the user to look at different areas of the screen and it collecting pupil diameter data as a function of screen luminance. In this process the pupil size would be measured at expected lighting levels (e.g. lighter and darker application windows), and at extremes (e.g., a white and a black screen), while there is no workload present. One could calculate, based on these data, an illumination-dependent average pupil diameter for the user. Using these values for the pupil size and knowing the current lighting level (e.g., based on what is shown on the screen) the proposed model can be used to estimate the workload while engaged in arbitrary tasks.

LIMITATIONS

One limitation of our work is that our model was implemented and tested using discrete levels of lighting and task difficulty. Furthermore, we avoided transitions between these discrete levels by introducing long rest periods between experimental conditions. Finally, the time resolution of our estimate is low, as we estimate workload for 30-second segments. We averaged measurements from the eye-tracker to reduce the effect of measurement noise. For future work, we plan to separate the two effects for each measurement point (at the frequency of the eye-tracker). However, we see this model as a first step towards high time-resolution workload estimation through pupil diameter, independent of lighting.

In a natural setting, a number of variables will change continuously in time and value, including the luminance and color of visual targets, ambient lighting, as well as task modality and difficulty. Thus, future work will have to address modeling pupil diameter changes in these realistic environments. If we can model pupil diameter changes due to factors other than cognitive workload, then we expect that we can use this model to estimate the changes that are due to cognitive workload. Our plan is to first explore this question in the context of interactions with user interfaces presented on screens, both small and large.

Another limitation is that our calculations are completed after all of the data has been collected, and not in real-time. While this approach would be useful in the design phase of a user interface, more work is needed to create a model that can be used to provide real-time feedback to UI algorithms.

Furthermore, in this work we use a simple heuristic classifier. It is possible that we could attain better classification with a more sophisticated classifier. However, the classifier we implemented serves the intended purpose of providing a proof of concept that separating the effects of light and task difficulty is possible. One of the next steps is to investigate continuously changing levels of light and task difficulty and present a system that can do this in real-time. With this step we will also introduce a sophisticated new classifier.

CONCLUSION

With the increasing proliferation of eye-tracking devices, it becomes more and more feasible to rely on this technology when interacting with computers and intelligent devices. One desire is to use eye-tracking to estimate the user's mental workload. This work seeks to tackle potential limitations of using eye-tracking methods in estimating mental workload such as the influence of illumination. By conducting a fully controlled experiment under different lighting and workload conditions, we are able to provide a dataset and preliminary model that can be used to estimate mental workload based on eye-tracking employing the user's pupil diameter.

ACKNOWLEDGMENTS

This work was funded in part by the NSF under grant IIA-1358096. We thank Micah Lucas, Michael Nguyen, and Rudra Timsina for their help in conducting the study and processing the data. We also thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310/2) and within the projects C02 of SFB / Transregio 161 "Quantitative Methods for Visual Computing" at the University of Stuttgart.

REFERENCES

1. Ulf Ahlstrom and Ferne J. Friedman-Berg. 2006. Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomic*s 36, 7 (Jul. 2006), 623-636. DOI: http://dx.doi.org/10.1016/j.ergon.2006.04.002
2. Amanda N. Baraldi and Craig K. Enders. 2010. An introduction to modern missing data analyses. *Journal of School Psychology* 48, 1 (Feb. 2010), 5-37. DOI: http://dx.doi.org/10.1016/j.jsp.2009.10.001
3. Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin* 91, 2 (Mar. 1982), 276-292. DOI: http://dx.doi.org/10.1037/0033-2909.91.2.276
4. Simone Benedetto, Marco Pedrotti, Luca Minin, Thierry Baccino, Alessandra Re, and Roberto Montanari. 2011. Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behavior* 14, 3 (May 2011), 199-208. DOI: http://dx.doi.org/10.1016/j.trf.2010.12.001
5. Anne-Marie Brouwer, Maarten A. Hogervorst1, Jan B. F. van Erp, Tobias Heffelaar, Patrick H. Zimmerman, and Robert Oostenveld. 2012. Estimating workload using EEG spectral power and ERPs in the n-back task. *Journal of Neural Engineering* 9, 4, Article 045008 (Jul. 2012), 045008, 14 pages. DOI: http://dx.doi.org/10.1088/1741-2560/9/4/045008
6. Joseph F. Coughlin, Bryan Reimer, and Bruce Mehler. 2009. *Driver Wellness, Safety & the Development of an AwareCar*. MIT AgeLab White Paper. Massachusetts Institute of Technology, Cambridge, MA. http://web.mit.edu/reimer/www/pdfs/coughlin_wellness_2009.pdf
7. Andrew T. Duchowski. 2002. A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers* 34, 4 (Nov. 2002), 455-470. DOI: http://dx.doi.org/10.3758/BF03195475
8. Christine Fogarty, and John A. Stern, 1989. Eye movements and blinks: Their relationship to higher cognitive processes. *International Journal of Psychophysiology* 8, 1 (Sep. 1983), 35-42. http://dx.doi.org/10.1016/0167-8760(89)90017-2
9. Thomas M. Gable, Andrew L. Kun, Bruce N. Walker, and Riley J. Winton. 2015. Comparing heart rate and pupil size as objective measures of workload in the driving context: initial look. In *Adjunct Proceedings of the 7th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '15)*. ACM, New York, NY, USA, 20-25. DOI: http://doi.acm.org/10.1145/2809730.2809745
10. Alan Gevins, and Michael E. Smith. 2003. Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science 4*, 1-2 (2003), 113-131. DOI: http://dx.doi.org/10.1080/14639220210159717
11. Audrey Girouard and Erin Treacy Solovey and Robert J.K. Jacob. 2013. Designing a passive brain computer interface using real time classification of functional near-infrared spectroscopy. *International Journal of Autonomous and Adaptive Communication Systems* 6,1 (2013), 26-44. DOI: http://dx.doi.org/10.1504/IJAACS.2013.050689
12. D. M. A. Gronwall. 1977. Paced Auditory Serial-Addition Task: A Measure of Recovery From Concussion. *Perceptual and Motor Skills* 44 (1977), 367-373. DOI: http://dx.doi.org/10.2466/pms.1977.44.2.367
13. Jennifer A. Healey, and Rosalind W. Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Trans. Intelligent Transportation Systems* 6, 2 (June 2005), 156-166. DOI: http://dx.doi.org/10.1109/TITS.2005.848368
14. Peter A. Heeman, Tomer Meshorer, Andrew L. Kun, Oskar Palinko, and Zeljko Medenica. 2013. Estimating cognitive load using pupil diameter during a spoken dialogue task. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (AutomotiveUI '13). ACM, New York, NY, USA, 242-245. DOI: http://dx.doi.org/10.1145/2516540.2516570
15. Daniel Kahneman, D., and Jackson Beatty. 1966. Pupil Diameter and Load on Memory. *Science* 154, 3756 (Dec. 1966), 1583-1585. DOI: http://dx.doi.org/10.1126/science.154.3756.1583
16. Daniel Kahneman, Linda Onuska, L., and Ruth E. Wolman. 1968. Effects of pupillary response in a short-term memory task. *The Quarterly Journal of Experimental Psychology* 20, 3 (1968), 309-311. DOI: http://dx.doi.org/10.1080/14640746808400168
17. W. K. Kirchner. Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology* 55, 4 (Apr. 1958), 352-358.
18. Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. 2008. Measuring the task-evoked pupillary response with a remote eye tracker. In *Proceedings of the 2008 symposium on Eye tracking research & applications* (ETRA '08). ACM, New York, NY, USA, 69-72. DOI: http://dx.doi.org/10.1145/1344471.1344489
19. Jeff Klingner, Barbara Tversky, and Pat Hanrahan. 2011. Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology* 48, 3 (Mar. 2011), 323-332. DOI: http://dx.doi.org/10.1111/j.1469-8986.2010.01069.x
20. Andrew L. Kun, Susanne Boll, and Albrecht Schmidt. 2016. Shirting gears: User interfaces in the age of autonomous driving. *IEEE Pervasive Computing* 15, 1

(January-March 2016), 32-37. DOI: http://dx.doi.org/10.1109/MPRV.2016.14

21. Andrew L. Kun, and Zeljko Medenica. 2012. Video call or not, that is the question. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1631-1636. DOI: http://dx.doi.org/10.1145/2212776.2223684
22. Andrew L. Kun, Oskar Palinko, Zeljko Medenica, and Peter A. Heeman. 2013. On the Feasibility of Using Pupil Diameter to Estimate Cognitive Load Changes for In-Vehicle Spoken Dialogues. In INTERSPEECH, (2013), 3766-3770.
23. Andrew L. Kun, Oskar Palinko, and Razumenić, I. 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* (AutomotiveUI '12). ACM, New York, NY, USA, 183-186. DOI: http://doi.acm.org/10.1145/2390256.2390287
24. Sandra P. Marshall, 2002. The Index of Cognitive Activity: measuring cognitive workload. In *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants*. IEEE, 7-5 – 7-9. DOI: http://dx.doi.org/10.1109/HFPP.2002.1042860
25. Bruce Mehler, Bryan Reimer, Joseph F. Coughlin, and Jeffery A. Dusek. 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 (2009), 6-12. DOI: http://dx.doi.org/10.3141/2138-02
26. Bruce Mehler, Bryan Reimer, and Joseph F. Coughlin. 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* 54, 3 (2010), 396-412. DOI: http://dx.doi.org/10.1177/0018720812442086
27. Bruce Mehler, Bryan Reimer and Jeffery A. Dusek. 2011. *MIT AgeLab Delayed Digit Recall Task (n-back)* Working Paper 2011-3B. Massachusetts Institute of Technology, Cambridge, MA. http://agelab.mit.edu/system/files/Mehler_et_al_n-back-white-paper_2011_B.pdf
28. W. Thomas Miller, and Andrew L. Kun. 2013. Using speech, GUIs and buttons in police vehicles: field data on user preferences for the Project54 system. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (AutomotiveUI '13). ACM, New York, NY, USA, 108-113. DOI: http://dx.doi.org/10.1145/2516540.2516564
29. William F. Moroney, David W. Biers, Thomas Eggemeier, and Jennifer A. Mitchell. 1992. A comparison of two scoring procedures with the NASA Task Load Index in a simulated flight task. In *Proceedings of the IEEE 1992 National Aerospace and Electronics Conference, NAECON 1992*. IEEE, 734-740. DOI: http://dx.doi.org/10.1109/NAECON.1992.220513
30. Nesrine Mezhoudi, Iyad Khaddam, and Jean Vanderdonckt. 2015. Toward Usable Intelligent User Interface. In: *Proceedings of the 17th International Conference on Human-Computer Interaction (HCII '15), Part II*, LNCS 9170, Springer International Publishing, 459-471. DOI: http://dx.doi.org/10.1007/978-3-319-20916-6_43
31. Michael Myrtek, Doris Weber, Georg Brügner, and Wolfgang Müller. 1996. Occupational stress and strain of female students: results of physiological, behavioral, and psychological monitoring. *Biological Psychology* 42, 3 (Feb. 1996), 379-391. DOI: http://dx.doi.org/10.1016/0301-0511(95)05168-6
32. Luis Nunes and Miguel A. Rescarte 2002. Cognitive demands of hands-free-phone conversation while driving. *Transportation Research Part F* 5, 2 (Jun. 2002), 133-144. DOI: http://dx.doi.org/10.1016/S1369-8478(02)00012-8
33. Julie Onton, Arnaud Delorme, and Scott Makeig. 2005. Frontal midline EEG dynamics during working memory. *NeuroImage* 27, 2 (Aug. 2005), 341-356. DOI: http://dx.doi.org/10.1016/j.neuroimage.2005.04.014
34. Oskar Palinko and Andrew Kun. 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 (Driving Assessment 2011)*, University of Iowa, Iowa City, IA, 329-336. http://drivingassessment.uiowa.edu/sites/default/files/DA2011/Papers/048_PalinkoKun.pdf
35. Oskar Palinko and Andrew Kun. 2012. Exploring the effects of visual cognitive load and illumination on pupil diameter in driving simulators. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*, ACM, New York, NY, 413-416. DOI: http://dx.doi.org/10.1145/2168556.2168650
36. Oskar Palinko, Andrew L. Kun, Alexander Shyrokov, and Peter A. Heeman. 2010. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '10)*, ACM, New York, NY, 141-144. DOI: http://dx.doi.org/10.1145/1743666.1743701
37. Bastian Pfleging and Albrecht Schmidt. 2015. (Non-) Driving-Related Activities in the Car: Defining Driver Activitgies for Manual and Automated Driving. In Workshop on Experiencing Autonomous Vehicles: Crossing the Boundaries between a Drive and a Ride at

CHI '15. http://www.hcilab.org/wp-content/uploads/pfleging-2015-drivingrelatedactivities.pdf

38. Bastian Pfleging, Stefan Schneegass, and Albrecht Schmidt. 2013. Exploring user expectations for context and road video sharing while calling and driving. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '13)*. ACM, New York, NY, USA, 132-139. DOI: http://dx.doi.org/10.1145/2516540.2516547
39. Bryan Reimer, Bruce Mehler, Joseph F. Coughlin, Kathryn M. Godfrey, and Chauanzhong Tan. 2009. An on-road assessment of the impact of cognitive workload on physiological arousal in young adult drivers. In *Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Application (AutomotiveUI '09),* ACM, New York, NY, 115-118. DOI: http://dx.doi.org/10.1145/1620509.1620531
40. A. H. Roscoe. 1992. Assessing pilot workload. Why measure heart rate, HRV and respiration? *Biological Psychology* 34, 2 (Nov. 1992), 259-287. DOI: http://dx.doi.org/10.1016/0301-0511(92)90018-P
41. Dennis W. Rowe, John Sibert, and Don Irwin. 1998. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '98)*, ACM, New York, NY, 480-487. DOI: http://dx.doi.org/10.1145/274644.274709
42. Kilseop Ryu and Rohae Myung. 2005. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics* 35, 11 (Nov. 2005), 991-1009. DOI: http://dx.doi.org/10.1016/j.ergon.2005.04.005
43. Stefan Schneegass, Bastian Pfleging, Nora Broy, Frederik Heinrich, and Albrecht Schmidt. A data set of real world driving to assess driver workload. In *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI 2013)*. ACM, New York, NY, 150-157. DOI: http://dx.doi.org/10.1145/2516540.2516561
44. Erin T. Solovey, Paul Schermerhorn, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, and Robert J.K. Jacob. 2012. Brainput: Enhancing Interactive Systems with Streaming fNIRS Brain Input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 2193-2202. DOI: http://dx.doi.org/10.1145/2207676.2208372
45. Yi-Fang Tsai, Erik Viirre, Christopher Strychacz, Bradley Chase, and Tzyy-Ping Jung. 2007. Task performance and eye activity: Predicting behavior relating to cognitive workload. *Aviation, Space, and Environmental Medicine* 78, 5 (May 2007), B176-B185.
46. Karl F. Van Orden, Wendy Limbert, Scott Makeig, and Tzyy-Ping Jung. 2001. Eye activity correlates of workload during visuospatial memory task. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 43, 1 (Spring 2001), 111-121. DOI: http://dx.doi.org/10.1518/001872001775992570
47. Ying Wang, Bryan Reimer, Bruce Mehler, Jun Zhang, Alea Mehler, and Joseph F. Coughlin. 2010. The Impact of Repeated Cognitive Tasks on Driving Performance and Visual Attention. In *Proceedings of the 3rd International Conference on Applied Human Factors and Ergonomics (AHFE' 10)*, Miami, FL, USA.
48. Yan Yang, Bryan Reimer, Bruce Mehler,and Jonathan Dobres. 2013. A field study assessing driving performance, visual attention, heart rate and subjective ratings in response to two types of cognitive workload. In *Proceedings of the 7th International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design (Driving Assessment 2013)*. University of Iowa, Iowa City, IA, 397-403.
49. Yabo Yang, Keith Thompson, and Stephen A. Burns. 2002. Pupil location under mesopic, photopic, and pharmacologically dilated conditions. *Investigative Ophthalmology & Visual Science* 43, 7 (Jul. 2002), 2508-2512.
50. Zhiwei Zhu, Kikuo Fujimura, and Qiang Ji. 2002. Real-time eye detection and tracking under various light conditions. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications (ETRA 2002)*. ACM, New York, NY, 139-144. DOI: http://dx.doi.org/10.1145/507072.507100