The Low/High Index of Pupillary Activity

Andrew T. Duchowski
School of Computing
Clemson University
duchowski@clemson.edu

Krzysztof Krejtz
Psychology
SWPS University of Social Sciences & Humanities
kkrejtz@swps.edu.pl

Nina A. Gehrer
Clinical Psychology
University of Tübingen
nina.gehrer@uni-tuebingen.de

Tanya Bafna
DTU Management
Technical Univ. of Denmark
taba@dtu.dk

Per Bækgaard
DTU Compute
Technical Univ. of Denmark
pgba@dtu.dk

ABSTRACT

A novel eye-tracked measure of pupil diameter oscillation is derived as an indicator of cognitive load. The new metric, termed the Low/High Index of Pupillary Activity (LHIPA), is able to discriminate cognitive load (*vis-à-vis* task difficulty) in several experiments where the Index of Pupillary Activity fails to do so. Rationale for the LHIPA is tied to the functioning of the human autonomic nervous system yielding a hybrid measure based on the ratio of Low/High frequencies of pupil oscillation. The paper's contribution is twofold. First, full documentation is provided for the calculation of the LHIPA. As with the IPA, it is possible for researchers to apply this metric to their own experiments where a measure of cognitive load is of interest. Second, robustness of the LHIPA is shown in analysis of three experiments, a restrictive fixed-gaze number counting task, a less restrictive fixed-gaze n-back task, and an applied eye-typing task.

Author Keywords

pupillometry; eye tracking; task difficulty

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI);** User studies

INTRODUCTION & BACKGROUND

Recent interest in the measurement of cognitive load has emerged from a variety of applied human factors settings, e.g., the automobile, flightdeck, operating room, and the classroom, to name a few [47, 5, 26, 22, 48, 29, 43, 59]. As noted by Fridman et al. [19], the breadth and depth of the published work highlights the difficulty of identifying useful measures of cognitive load that do not interfere with or influence behavior. Moreover, if the measure is based on pupil diameter, as a good deal of these metrics are, then it is also important to show that the metric is not susceptible to effects of luminance or off-axis distortion of the apparent pupil (e.g., as captured by the typically stationary camera) [15].

As introduced by Sweller [57, 58], cognitive load is a theoretical construct describing the internal processing of tasks that cannot be observed directly [41]. One of the most popular measures to assume indication of cognitive load is pupil diameter, originating with Hess and Polt [25] and later bolstered by Peavler [49], who showed correlation between pupil dilation and problem difficulty. It is generally considered that pupil diameter provides a "very effective index of the momentary load on a subject as they perform a mental task" [34].

Early studies of task-evoked pupillary response to cognitive load used specialized pupillometers to measure pupil diameter [1, 8, 7]. Because of their improved accuracy and reduced cost, eye trackers have become popular for the estimation of cognitive load via measurement of pupil diameter, which most eye trackers report as a matter of course [54, 11, 53]. The general approach to cognitive load estimation with eye-tracked pupil diameter relies on measurement relative to a baseline. Numerous examples of eye-tracked baseline-related pupil diameter measurements exist, focusing either on inter- [27, 38, 41, 36], or intra-trial baseline differences [54, 39, 30].

Besides pupil diameter, some eye-tracking users infer cognitive load from blink rate [12], while others consider blinks something of an eye-tracking by-product. When blinks occur, the eye tracker loses sight of the pupil, and often outputs some undefined value for gaze position. Other approaches to cognitive load measurement evaluate positional eye movements, including number of fixations [28], fixation durations [18, 32], and number of regressions [4], although these metrics could be considered indirect indicators of cognitive load. More recent approaches to cognitive load measurement use microsaccades (the component of miniature eye movements, along with tremor and drift, made during visual fixation [17]). For a review of the observed relationship between microsaccades and task difficulty, see Duchowski et al. [16]. For a detailed review of Cognitive Load Theory (CLT) and related measures, see Kelleher and Hnin [35], Duchowski et al. [15], and Cowley et al. [14].

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04…$15.00.
DOI: http://dx.doi.org/10.1145/3313831.3376394

We introduce the Low/High Index of Pupillary Activity, or LHIPA. The LHIPA is a wavelet-based algorithm inspired by Marshall's [45, 46] Index of Cognitive Activity (ICA), and Duchowski et al.'s [15] implementation of their IPA, introduced at CHI '18. Because the ICA and IPA are based on pupil diameter oscillation, they both reflect moment-to-moment pupil diameter changes and are thought to be insensitive to effects of luminance. Duchowski et al. reported sensitivity of the IPA under restricted conditions, namely during a fixed-gaze task based on mental arithmetic. Cognitive load, however, is associated with working memory capacity [57], and so the n-back task is often considered more appropriate for its manifestation. We show that the IPA fails to distinguish cognitive load during execution of an n-back (1- and 2-back) task with gaze fixed at specific screen coordinates (to test the effect of off-axis pupil distortion). The LHIPA yields significant results in the n-back task as well as in a less restrictive eye-typing task where the eyes move freely to effect key entry. Courtesy of Duchowski, we show that the LHIPA is also sensitive to task difficulty given their original number counting task.

Computation of the LHIPA is similar to the IPA except for one crucial difference: instead of counting the remnants of the thresholded modulus maxima of any particular wavelet frequency band (Duchowski et al. use the second), we count the thresholded modulus maxima of *the ratio of low and high frequency bands* contained within the wavelet decomposition of the pupil diameter signal, i.e., threshold of the low frequency/high frequency (LF/HF) ratio of pupil oscillation.

The low frequency/high frequency (LF/HF) ratio reflects changes in the parasympathetic and sympathetic components of the autonomic nervous system. Parasympathetic excitation and/or sympathetic inhibition result in pupil constriction, while sympathetic excitation and/or parasympathetic inhibition result in pupillary dilation [8, 56]. Peysakhovich [51] successfully used the LF/HF ratio of the pupillary power spectrum to measure cognitive load under varying light conditions.

Peysakhovich et al. [52] also suggest that the LF/HF ratio reflects the relationship between the Locus Coeruleus-Norepinephrine (LC-NE) system (functioning in two firing modes, tonic and phasic) and the pupil diameter. According to Adaptive Gain Theory (AGT) [3], and on the basis of the neuromodulatory effects of NE release on cortical processing, LC phasic firing typically occurs in response to task-relevant events during epochs of high performance and lower baseline LC activity (exploitation). In contrast, LC tonic firing is associated with elevated baseline firing rate, absence of phasic responses, and degraded task performance (exploration) [20]. Peysakhovich [51] suggested that *tonic pupil diameter* and *phasic pupil response* correspond to Beatty's [7] baseline diameter and the Task-Evoked Pupillary Response (TEPR), respectively. Measurement of the LF/HF ratio thus yields a baseline-related measure, similar to traditional baseline-related pupillometric measures of cognitive load (for a review of the latter, see Krejtz et al. [40]).

Paper Overview and Contributions

The novelty of our contributed LHIPA computation comes from the spectral analysis that is afforded by the (discrete) wavelet transform. That is, the LHIPA relies on the LF/HF ratio computed from the low and high frequency bands found in the (dyadic) wavelet decomposition. Because of the wavelet transform's local support (short convolution filters), the decomposition could in principle be carried out in real-time, without sacrificing temporal resolution as might be required by traditional power spectral analyses that typically require the complete signal for analysis.

Below we begin by providing details of the LHIPA implementation and then review three experiments where we show the measure's greater sensitivity to cognitive load (task difficulty) than the IPA under various conditions.

IMPLEMENTATION OF THE LOW/HIGH IPA

Computation of the LHIPA relies on wavelet decomposition of the pupil diameter, $x(t)$, in all experiments taken to be the average of the left and right pupil diameter as reported binocularly by each eye tracker, and its wavelet analysis [45, 10]. As per Duchowski et al.'s IPA implementation, we choose the *symlet*-16 wavelet. As with computation of the IPA, the LHIPA proceeds identically via a dyadic series representation of $x \in L^2(\mathbf{R})$: $x(t) = \sum_{j,k=-\infty}^{\infty} c_{j,k}\psi_{j,k}(t)$, $j,k \in \mathbf{Z}$, with $\{c_{j,k}\}$ the wavelet coefficients, and the *mother wavelet function* $\psi_{j,k}(t)$ expressed by $\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)$, with wavelet dilation and translation parameters j,k, respectively. It is important to emphasize that integral powers of 2 are used in the decomposition of the wavelet $\psi_{j,k}(t)$ via binary dilation (2^j), and dyadic translation ($k/2^j$) of a single function ψ. The importance of the use of integral powers of 2 are evident when indexing wavelet coefficients at different scales, i.e., when obtaining the pointwise low/high coefficient ratio, as given below.

Using the Discrete Wavelet Transform (DWT) to analyze the pupil diameter signal at multiple levels of resolution, the wavelet coefficients are found directly:

$$x_{\psi}^{j-1}(t) = \sum_{k} g_k x_{\psi}^{j}(2t+k),$$

where $\{g_k\}$ is the one-dimensional high-pass wavelet filter. Level j can be chosen arbitrarily to select either high- or low-frequency wavelet coefficients.

For the high frequency component, we choose $j = 1$ and for the low, $j = \frac{1}{2}\log_2(n)$, the mid-level frequency octave where $\log_2(n)$ is the number of octaves. The LF/HF ratio is thus:

$$x_{\psi}^{1/2\log_2(n)}(t) / x_{\psi}^{1}(2^{1/2\log_2(n)}t)$$

where the term $(2^{1/2\log_2(n)}t)$ in the denominator is the scale factor of the index into the high frequency signal when iterating over the (shorter) low frequency component. Remember that the length of the wavelet coefficient signal (array) at each lower resolution level is half that of the previous higher level of resolution. Hence, while iterating over the lower frequency signal, the index must be multiplied by raising 2 to the power

of frequency octave level being processed. Python implementation of the LHIPA is given in Listing 1 (see Appendix).

Note that the LHIPA is a ratio of low to high frequency, with high frequency response expected with increased cognitive load (task difficulty), thus ***LHIPA is expected to decrease with increased cognitive load***, the reverse of the IPA response.

The last stage of LHIPA computation is almost identical to what is used in the computation of the IPA, except for the mode of thresholding. Specifically, we first find the sharp points of variation in the LHIPA signal via modulus maxima detection as for the IPA, then again use universal thresholding. Instead of 'hard' thresholding, we use 'less' thresholding wherein data is decimated if above the threshold while lesser values pass untouched. This is because unlike the IPA, we expect a smaller LHIPA response for greater cognitive load. Finally, we count the number of remaining coefficients to produce the LHIPA.

EXPERIMENT 1: FIXED-GAZE NUMBER COUNTING

To compare the sensitivity of the LHIPA and IPA to task difficulty, we first used the dataset provided by Duchowski et al. [15]. We include only the relevant methodological aspects of their study for context. All details of the experiment are identical to their original study and have not been altered in any way.

Experimental Setting and Apparatus

An SR Research EyeLink 1000 eye tracker was used to record eye movements binocularly at a sampling rate of 500 Hz. Each participant's head was stabilized with a chin rest during the entire experimental procedure. Eye tracker accuracy is reported by the manufacturer as 0.25–0.5° visual angle on average.

The experimental procedure was controlled by a personal computer connected to the eye-tracking computer. Visual stimuli were displayed on a computer screen with 1920×1080 resolution. The procedure was written in Python with the use of the PsychoPy package [50]. Responses made by participants were performed on a standard numerical keyboard connected to the stimuli presentation computer and placed at the side of the participant's dominant hand.

The experimental laboratory was devoid of windows limiting the amount of ambient light during the study. Ambient luminance in the laboratory was 520 lux, with luminance of 120-130 lux at the computer screen with the fixation point at screen center during the main part of the procedure.

Experimental Procedure

Three types of number counting trials, Difficult, Easy, and Control, were grouped into 6 blocks, yielding 18 trials total. Each block started with the Control, followed by the Easy and Difficult trials in counterbalanced order. Between each block of trials, participants were asked to take a short break lasting 2-5 minutes. Each block started with the instruction screen. After each block of trials, the NASA Task Load Index [23] (NASA-TLX) questionnaire was completed as a self-reported cognitive load assessment. The raw NASA-TLX index was used for the experimental manipulation check.

In the Difficult trials, participants were asked to mentally count backwards, as fast and accurately as possible, in steps of 17 starting at one of the following 4-digit numbers drawn randomly from this set: {1375, 8489, 5901, 5321, 4819, 1817}.

The Easy and Control trials were constructed similarly to Difficult trials, but differed in task performance and initial instructions. In the Easy tasks, participants were instructed to mentally count forward, as fast and accurately as possible, in steps of 2 starting at one of the following 3-digit numbers drawn randomly from this set: {363, 385, 143, 657, 935, 141}. In the Control trials, participants had no mental task assigned.

During each trial, participants were prompted four times to enter their current number (in Difficult and Easy trials) or any 3-digit number (in Control trials) in a text box on the screen. A limit of 9 seconds was given for providing the entry. Three prompts appeared at random times during each trial, and the fourth at the very end of the trial. The gap between prompts was a minimum of 15 seconds and a maximum of 80 seconds.

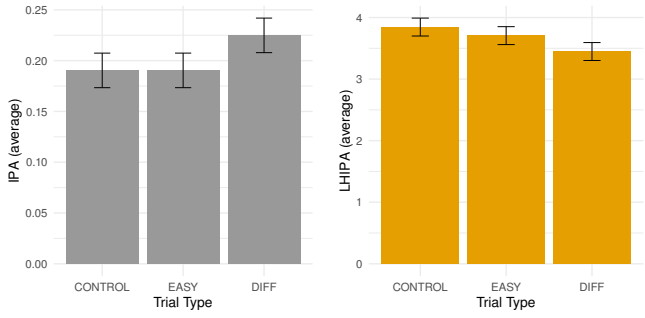
When performing the mental calculations, participants were asked to gaze at the fixation point appearing at screen center. Whenever their gaze shifted 3° visual angle away from the fixation point a warning beep sounded.

Participants

Volunteers ($N=17$) for the study were recruited verbally and by social media. Due to problems with eye tracker calibration or misunderstanding of the task, data from 4 were discarded giving a final sample of $N=13$ (7 M, 6 F with ages in range [20:40] years old, $M=29.77$, $SD=7.15$). All were right-handed with normal, uncorrected vision.

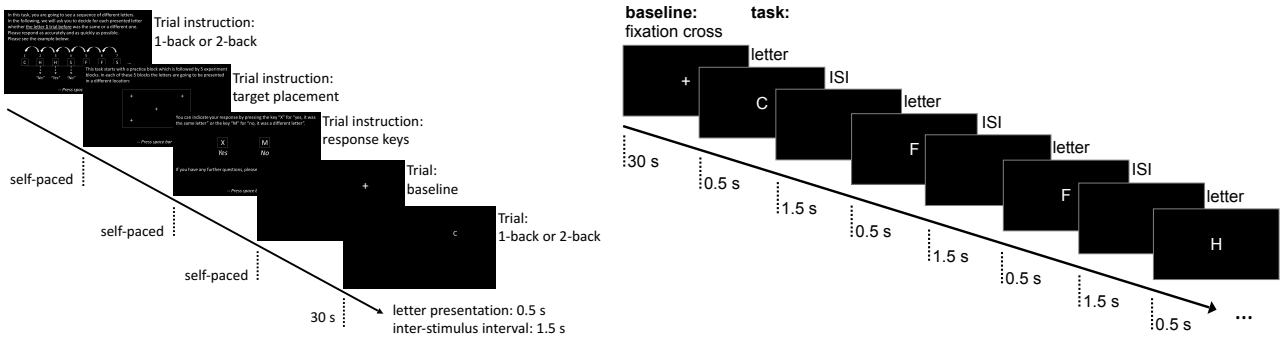
Results & Brief Discussion

Before testing hypotheses, a manipulation check was performed with the raw NASA-TLX score as a dependent variable. Within-subject one-way ANOVA was performed on this score with task difficulty as a fixed factor. As expected, results showed a significant main effect of task difficulty, $F(1.40, 16.76) = 46.81, p < 0.001, \eta^2 = 0.39$ (c.f. results reported by Duchowski et al. [15]). The following pairwise comparisons with Tukey HSD correction revealed that Difficult tasks were evaluated with significantly ($p < 0.001$) higher



(a) IPA vs. trial. (b) LHIPA vs. trial.

Figure 1. Pupil response to task difficulty, with mean IPA and LHIPA versus task; error bars represent ± 1 SE.



(a) Procedure schematic showing instruction and baseline.

(b) Procedure schematic showing trial timing.

Figure 2. Experimental procedure schematic and trial timing.

cognitive load (M = 12.26, SE = 1.04) than Easy tasks (M = 8.39, SE = 1.04) and Control tasks (M = 5.19, SE = 1.04). The difference between Easy and Control Tasks was also statistically significant ($p < 0.001$).

For analyzing pupillary response to task difficulty we first pre-processed the pupil diameter signal removing data 200 ms before the start of, and 200 ms following the end of a blink, as identified by the eye tracker, following Engbert and Kliegl [17]. After this pre-processing step, we then compute the IPA and LHIPA on the raw pupil diameter signal, $x(t)$.

We hypothesized that both IPA and LHIPA should indicate differences in cognitive load, distinguishing between Difficult, Easy, and Control tasks. We expected the IPA to be directly proportional and the LHIPA to be inversely proportional to task difficulty and for each to distinguish between the Easy, Difficult, and Control conditions.

In order to test this hypothesis each of the IPA and LHIPA served as the dependent variable in two one-way within-subject ANOVA tests with task difficulty as a fixed factor. Both ANOVAs were followed with pairwise comparisons with Tukey correction for comparing a family of 3 estimates. Before running statistical tests both IPA and LHIPA were tested for outliers. Outliers were defined as 1.5 times the interquartile range above the upper quartile and below the lower quartile. For IPA one (1) and for LHIPA six (6) outliers were identified. They were replaced with .05 or .95 percentile values of the variable depending on whether they were below lower or above upper thresholds.

IPA Sensitivity

The IPA appears to increase significantly with task difficulty, $F(2,24) = 4.124, p = 0.03, \eta^2 = 0.07$, see Figure 1(a). Pairwise comparisons showed that the IPA differed ($p = 0.051$) between the Difficult (M = 0.23, SE = 0.02) and each of the Easy (M = 0.19, SE = 0.02) and Control tasks (M = 0.19, SE = 0.02). The difference between the Easy and Control tasks was not significant ($p = 1$).

LHIPA Sensitivity

ANOVA of the LHIPA revealed a statistically significant main effect of task difficulty, $F(2,24) = 3.737, p = 0.04, \eta^2 = 0.10$. Pairwise comparisons showed that the LHIPA differed significantly ($p = 0.03$) between the Difficult (M = 3.45, SE = 0.15) and Control tasks (M = 3.85, SE = 0.15). Differences between the Easy task (M = 3.71, SE = 0.15) and each of the Control and Difficult tasks were not significant ($p = 0.62$, $p = 0.21$, resp.), see Figure 1(b).

Brief Discussion

Results indicate that the LHIPA, like the IPA, responds to task difficulty in the fixed-gaze number counting task. Both measures distinguish between the difficult task and either of the easy and control tasks, but are not sensitive enough to distinguish between the easy and control conditions. It is also worth noticing that the effect size of task difficulty on LHIPA is slightly greater ($\eta^2 = 0.10$) than on IPA ($\eta^2 = 0.07$).

EXPERIMENT 2: FIXED-GAZE N-BACK TASK

Experiment 1 demonstrated the sensitivity of the IPA and LHIPA to task difficulty during mental calculation when pupillary response was captured with gaze fixed at the center of the screen. In Experiment 2 to better establish a connection between task difficulty and active working memory load [57], the manipulation task was changed to an n-back task [37]. To evaluate potential effects of the distortion of the apparent off-axis pupil, the gaze fixation point was shown at five (5) different locations on the screen (see details below).

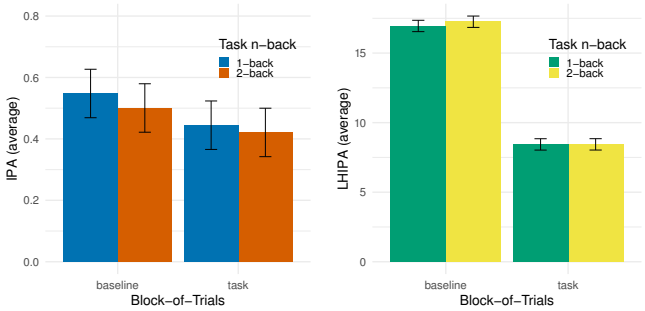
Experimental Procedure

Experiment 2 followed Appel et al.'s [2] n-back experimental protocol who used Marshall's [45] ICA as a feature of their cross-subject classification of cognitive load.

The n-back task consists of a randomly generated sequence of letters from the set $L = \{C, F, H, S\}$ shown one after another, each for a duration of 0.5 seconds with an inter-stimulus duration of 1.5 seconds (see Figure 2). Each letter and inter-stimulus period thus constituted a 2 second trial. Within each trial, participants had to state whether the currently shown letter was the same as the one n trials before by pressing one of two keyboard buttons.

Participants underwent a training phase for each of the n-back tasks they started. They were presented with a block of 15

Figure 3. Experimental setup with participant performing n-back task.



(a) IPA vs. baseline and task. (b) LHIPA vs. baseline and task.

Figure 4. Pupil response to task difficulty, with mean IPA and LHIPA vs. baseline and task; error bars represent ± 1 SE.

trials from the task they were training for and repeated the training block until they reached an accuracy of at least 60%.

Each of the 1- and 2-back tasks was preceded by a baseline task with a fixation cross positioned at the point where the stimulus letters would appear and lasted for 30 s.

Every participant completed five blocks of each of the 1- and 2-back tasks. Each of the five blocks displayed the target stimulus (letter) at one of 5 position on the screen, center, top-left, top-right, bottom-right, or bottom-left. With the exception of the center block, the remaining 4 blocks were at 10° visual angle from center along each of the x- and y-axes. The order of the 1- and 2-back tasks was counterbalanced. Within the tasks, the order of the 5 locations was randomized. Each block consisted of 60 trials, lasting 2 minutes.

Participants

Volunteers (N = 19) for the study were recruited verbally. Data from 1 participant were discarded due to their accuracy falling well below the 60% level required during training, giving a final sample of N = 18 (15 M, 3 F with ages in range [21:29] years old, M = 22.77, SD = 2.26). All had normal, or corrected to normal vision.

Experimental Setting and Apparatus

Similar to the first experiment, we used an SR Research Eye-Link 1000 eye tracker to record eye movements binocularly at a sampling rate of 500 Hz. Each participant's head was stabilized with a chin rest during the entire experimental procedure.

The experimental procedure was controlled by a personal computer connected to the eye-tracking computer. Visual stimuli were displayed on a 24 in wide computer screen with 1920×1200 resolution. The procedure was written in Python with the use of the PsychoPy package [50]. Responses made by participants were performed on a standard numerical keyboard connected to the stimuli presentation computer and placed at the side of the keyboard.

The experimental laboratory was devoid of windows limiting the amount of ambient light during the study (see Figure 3). Ambient luminance in the laboratory was 30 lux at the screen and 8 lux at eye distance during baseline and trial stimulus presentation, measured with a Mastech Professional Lux Meter (model LX1010B).

Results & Brief Discussion

As a manipulation check, the accuracy of responses between experimental conditions was compared by a one-way within-subject ANOVA where task (1-back vs. 2-back) was treated as a fixed factor. The ANOVA showed a significant main effect of task, $F(1,18) = 46.859, p < 0.001, \eta^2 = .16$. As expected, accuracy in the 1-back task was significantly greater (M = .91, SE = .03) than in the 2-back task (M = .80, SE = .03).

We pre-processed the eye movement signal similarly to that of the first experiment, removing data 200 ms before the start of, and 200 ms following the end of a blink, as identified by the eye tracker. After this pre-processing step, we then computed the IPA and LHIPA on the raw pupil diameter signal, $x(t)$.

We hypothesized that both IPA and LHIPA should indicate differences in cognitive load, distinguishing between both 1- and 2-back tasks and the baseline task. We expected the IPA to be significantly greater for the 2-back tasks and the reverse for LHIPA. We expected neither indicator to vary with gaze distance from center (off-axis position).

To test the hypotheses two three-way within-subjects ANOVA tests were conducted with IPA and LHIPA as dependent variables. In both tests Block-of-Trials (baseline vs. task), Task Type (1-back vs. 2-back) and Off-Axis Position (top-left vs. top-right vs. center vs. bottom-right vs. bottom-left) were treated as fixed factors. Results are reported with Greenhouse-Geisser correction when the sphericity assumption was not met. Both ANOVAs were followed by pairwise comparisons with Tukey correction for comparing a family of 3 estimates. Prior to ANOVA, as in Experiment 1, we employed a similar strategy for identification and removal of IPA and LHIPA outliers, with 13 outliers identified for LHIPA (3.61%), and 33 outliers for IPA (9.7%).

IPA Sensitivity

The ANOVA for IPA as a dependent variable revealed only a marginally significant main effect of Block-of-Trials, $F(1,17) = 4.20, p = 0.056, \eta^2 = 0.01$. Moreover, contrary to expectations, the IPA for n-back tasks was lower (M = 0.43, SD = 0.07) than for the baseline (M = 0.52, SD = 0.07), see Figure 4(a). For the effect of Task (1-back vs. 2-back),

neither main effect, $F<1$ nor interaction, between Block-of-Trials and Task, $F<1$ was statistically significant.

Contrary to the hypothesis, off-axis gaze position significantly influenced the IPA, $F(2.78,47.21)=5.19, p=0.004, \eta^2=0.05$. Pairwise comparisons showed that the IPA obtained from gaze at bottom-left was significantly greater than the IPA from gaze at top-left ($p=0.003$), top-right ($p=0.004$), and the screen center ($p=0.037$). Pairwise comparisons of the IPA and other off-axis positions were not significant. See Table 1 for detailed descriptive statistics.

Neither interaction effect of off-axis position with Block-of-Trials, $F(2.95,50.21)=1.41, p=0.25$ nor with Task Type $F<1$ was statistically significant. The three-way interaction between Off-Axis Position, Block-of-Trials and Task Type was not significant, $F<1$.

LHIPA Sensitivity

ANOVA of the LHIPA revealed a statistically significant main effect of Block-of-Trials, $F(1,17)=648.93, p<0.001, \eta^2=0.81$. In line with hypotheses, LHIPA from task trials was significantly lower (M = 8.44, SD = 0.38) than from the baseline trials (M = 17.10, SD = 0.38), see Figure 4(b).

The analysis also showed that neither the Task main effect, $F<1$ nor interaction between Block-of-Trials and Task, $F(1,17)=1.01, p=0.33$ was statistically significant. The main effect of Off-Axis Position was not significant, $F(2.28,38.83)=1.23, p=0.31$ neither was any of the remaining interaction terms in which Off-Axis Position was involved (in all of these the interaction effects showed $F<1$).

Brief Discussion

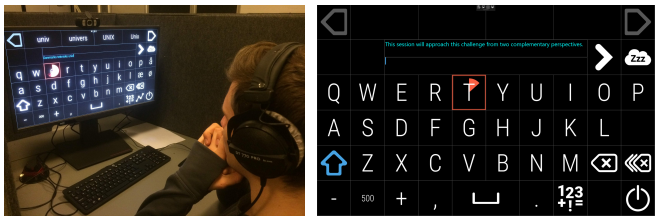
Results indicate that the LHIPA was not sensitive to the off-axis positions of the letter targets, but that the IPA was. For the IPA, this is an undesirable result, since it shows that off-axis distortion, otherwise known as the Pupil Foreshortening Effect (PFE) [24], may decrease the IPA's reliability in detecting cognitive load. As in the fixed-gaze number counting task, neither the IPA nor LHIPA could distinguish between the tasks themselves, but the LHIPA could distinguish between task and baseline, whereas the IPA could not, possibly due to the PFE.

The reason for the LHIPA's inability to distinguish between tasks (1- and 2-back) may be twofold. First, Experiment 2 was run in fairly low light conditions (30 lux compared to 520 lux of Experiment 1). Thus, participants' pupil diameter was probably already somewhat dilated. Therefore, the relative phasic change in pupil diameter could have been much smaller due to a relatively large tonic signal to begin with. Had the lights been turned on during the study, the tonic pupil diameter would have been smaller allowing for greater changes in dilation, i.e., greater gain in phasic response could have been observed, according to Adaptive Gain Theory. Peysakhovich put it this way: according to the *law of initial value* [42, 31], a large tonic pupil diameter would imply a smaller phasic pupil response, potentially resulting in a restriction of pupillary dynamic range, i.e., when the pupil is already large, it cannot dilate further.

Second, another reason for LHIPA's lack of sensitivity to trial difficulty is that actual cognitive load did not differ greatly between the 1- and 2-back tasks, where task accuary was recorded as M = .91 and M = .80, respectively. Perhaps using a 1- and 3-back task would have elicited greater differences. However, increasing difficulty may lead to cognitive overload at which point phasic response could start to diminish.

Table 1. IPA scores for different off-axis positions.

Off-axis position	Mean	Standard Error
bottom-left	0.65	0.08
bottom-right	0.55	0.08
center	0.44	0.08
top-left	0.38	0.08
top-right	0.39	0.08



(a) Eye-typing experiment setup. (b) OptiKey onscreen keyboard.

Figure 5. Experimental setup of a participant performing the eye-typing task and the on-screen keyboard OptiKey.

EXPERIMENT 3: EYE TYPING WITH OPTIKEY

To evaluate the LHIPA in an unrestricted environment, we used a text-copy task in an eye-typing experiment. We modified the task by involving the working memory aspect of cognitive load and asking the participants to memorize the text. Task difficulty was varied by the complexity of the text to be memorized. The study followed a within-subjects design with task difficulty as a fixed factor.

Experimental Setting and Apparatus

A Tobii Eye Tracker 4C (sampling frequency 90 Hz) was used to record the eye movements binocularly during the experiment, with an experimental version of the open-source keyboard *OptiKey*, a Windows Presentation Foundation (WPF) application.[1] The experimental version of the on-screen keyboard [13] allows us to provide and control the text to be eye-typed in the interface.

The experimental procedure, incorporated into the keyboard, was presented to the participants on a computer screen with 1920×1080 resolution. The logging method, also programmed in the keyboard, allowed us to obtain eye movements as well as the typing procedure during the experiment.

The experiment was performed in a laboratory with no access to natural light. Luminance was kept low, varying with participants and days between 25-60 lux at viewing distance, but was kept constant for each day and participant.

[1] https://github.com/OptiKey/OptiKey/wiki

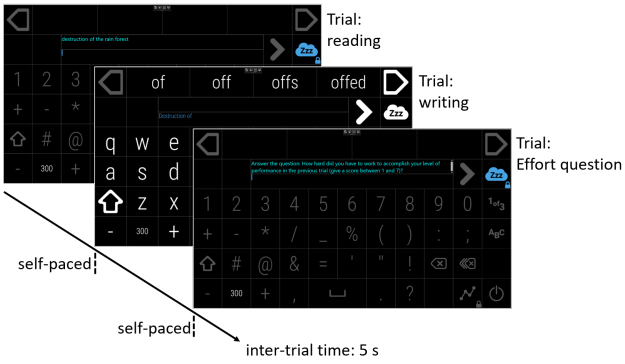
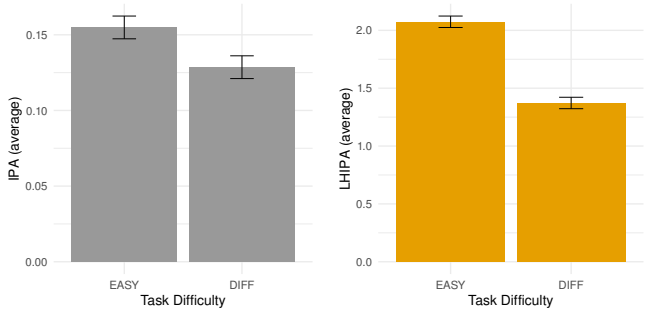


Figure 6. Procedure schematic showing the sequence of one trial.



(a) IPA vs. task difficulty. (b) LHIPA vs. task difficulty.

Figure 7. Pupil response to task difficulty, with mean IPA and LHIPA versus eye-typing trial difficulty; error bars represent ± 1 SE.

Experimental Procedure

Two types of text to be copied, Easy and Difficult, were separated into 4 blocks with 5 trials of each kind, resulting in a total of 40 trials. Different blocks were performed by the participants on different days. Each block consisted of 5 easy or difficult trials, followed by 5 trials of the opposite kind. The order of the easy and difficult trials were counterbalanced. Participants took a small break between each set of trials of a difficulty level.

A trial in this self-paced experiment started with the sentence presented to the participants in the space above the keyboard, shown in Figure 5, to be memorized and eye-typed. After reading and memorizing, the participants were asked to type the sentence. On entering the first character, the sentence would disappear, forcing the participants to rely on their memory to eye-type the rest of the sentence. A dwell-time typing technique with auditory and visual feedback was used for eye-typing [44]. The trial ended with the participants being asked to rate their perception of the difficulty level of the trial (termed perceived difficulty), using the *effort* question from the NASA-TLX questionnaire, on a scale of 1 to 7. The inter-trial time was 5 seconds and the sequence of the trial is shown in Figure 6.

The difficulty level of the sentence to be memorized and typed, used to induce cognitive load, was operationalized by the text complexity score, which in turn was defined by the LIX readability score [9]:

$$\text{LIX} = A/B + (C \times 100)/A$$

where A is the number of words, B is the number of periods (defined by period, colon, or capital first letter, and which was always 1 in our case), and C is the number of long words (more than 6 letters). Sentences with a LIX score greater than 60 were categorized as difficult and sentences with a LIX score less than 30 were categorized as easy. The sentences were taken from the Leipzig corpus [21].

Participants were not restricted in any way, neither in terms of head mobility nor asked to fix their gaze location during the experiment.

Participants

Volunteers ($N = 19$) were recruited via announcements in courses and the university's social media groups. One participant dropped out after one block, and so the data from $N = 18$ (9M, 9F with ages in range [21:31] years, $\text{M} = 25.5$, $\text{SD} = 2.38$) is presented here.

Results & Brief Discussion

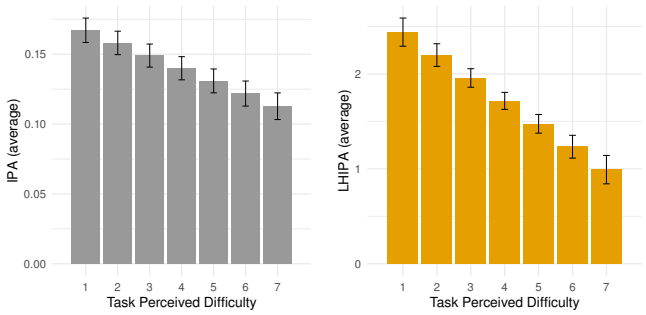
We start the results description with a manipulation check. In this experiment we implemented two types of manipulation check based on subjective evaluation of *effort* (from the NASA-TLX questionnaire) of each trial as well as two performance measures: error rate and the typing speed. The error rate, measured in %, was computed to be equally weighted character- and word-level Levenshtein distance, and the typing speed was calculated in normalized words per minute (WPM), where 1 word consists of 5 characters. The effort evaluation was treated as an indicator of self-assessed perceived task difficulty. All three measures were compared between easy and difficult trials with the use of within-subject one-way ANOVA tests.

First, the ANOVA on perceived effort showed a significant effect of the task difficulty, $F(1,12) = 144.24, p < 0.001, \eta^2 = 0.57$. As expected, the subjective effort evaluation for easy trials was lower ($\text{M} = 3.03, \text{SE} = 0.22$) than for difficult trials ($\text{M} = 4.83, \text{SE} = 0.22$).

Second, the ANOVA test on error rate showed a significant main effect of task difficulty $F(1,17) = 43.15, p < 0.001, \eta^2 = 0.29$. As expected, error rate for easy trials was lower ($\text{M} = 0.09, \text{SE} = 0.03$) than for difficult trials ($\text{M} = 0.23, \text{SE} = 0.03$).

Finally, the ANOVA for typing speed gave a significant effect of the task difficulty, $F(1,17) = 15.67, p = 0.001, \eta^2 = 0.02$. Following expectation, the typing speed was higher for easy tasks ($\text{M} = 9.94, \text{SE} = 0.67$) than for difficult tasks ($\text{M} = 9.20, \text{SE} = 0.67$).

The pupillary data were pre-processed in the same way as in the two previous experiments, by detecting blinks and removing 200 ms before the start and after the end of the blinks. We computed the IPA and LHIPA on the raw pupil size data.



(a) IPA vs. perceived difficulty. (b) LHIPA vs. perceived difficulty.

Figure 8. Mean IPA and LHIPA compared to perceived difficulty; error bars represent ± 1 SE.

We hypothesized that the IPA and LHIPA should differentiate between the cognitive load generated in performing easy and difficult tasks. We expected the IPA to be higher for the difficult trials than for easy trials and LHIPA to be lower for the difficult trials than for easy trials. The hypothesis was tested using repeated measures one-way ANOVA with trial difficulty as the independent variable and IPA and LHIPA as the dependent variable.

We also hypothesized that the IPA and LHIPA should distinguish between subjective perceived effort. We expected the IPA to be higher for a higher perceived effort and the LHIPA to be lower for a higher score of perceived difficulty score. Both hypotheses were tested with linear mixed models (LMMs) with perceived effort as a within-subject predictor. All models were estimated using the *lme4* package in R [55]. We fit models with perceived effort as the independent variable using a step-up approach based on Akaike Information Criterion (AIC), starting from the model with only random intercepts.

IPA Sensitivity

One-way repeated measures ANOVA showed a significant effect of trial difficulty, $F(1,17) = 26.61$, $p < 0.001$, $\eta^2 = 0.15$. Contrary to expectations, IPA was significantly lower for difficult trials (M = 0.13, SE = 0.01) than for easy trials (M = 0.16, SE = 0.01), see Figure 7(a).

The LMM analysis retained perceived effort as a significant predictor of IPA, ($\chi^2(2) = 16.65$, $p < 0.001$, AIC = -2170.70, BIC = -2143.34, Pseudo $-R^2 = 0.12$, observation number = 700). The model included a perceived effort random intercept and slope grouped by participant, see Figure 8(a). The intercept was significantly greater than zero, (est. = 0.18, SE = 0.01, $t(28.09) = 19.80$, $p < 0.001$). The slope was also significant, (est. = -0.01, SE = 0.00, $t(28.53) = 7.46$, $p < 0.001$).

LHIPA Sensitivity

One-way repeated measures ANOVA showed a significant effect of task difficulty, $F(1,17) = 256.79$, $p < 0.001$, $\eta^2 = 0.75$. Following expectations, LHIPA was significantly lower for difficult trials (M = 1.37, SE = 0.049) than for easy trials (M = 2.07, SE = 0.049), see Figure 7(b).

Linear-Mixed Models analysis of LHIPA with perceived effort revealed a significant effect of perceived effort ($\chi^2(2) = 14.76$, $p < 0.001$, AIC = 1309.20, BIC = 1336.50, Pseudo $-R^2 = 0.38$, observation number = 700). Similar to LMM analyses for IPA, the model for LHIPA included a perceived effort random intercept and slope grouped by participant. The obtained intercept was significantly greater than zero, (est. = 2.68, SE = 0.16, $t(8.00) = 16.99$, $p < 0.001$). In line with hypotheses, the slope was negative and significantly different from a flat line, (est. = -0.24, SE = 0.03, $t(7.18) = 7.05$, $p < 0.001$). Higher perceived effort of the eye-typing task significantly lowers the LHIPA score, see Figure 8(b).

IPA and LHIPA Over Fixed Trial Duration Time

Due to the self-paced nature of the experimental design, the trial duration could vary between trials. The LHIPA computation involves a factor of time, i.e., the number of detected modulus maxima peaks over threshold are divided by the trial duration. To account for the variability in time duration, we decided to investigate further by using an equal trial duration for all trials. The fastest trial was 12 s, and so we used the first 12 s of every trial to compare the IPA and LHIPA. We expected the effect of task difficulty on the IPA and LHIPA to not change, even if only the first 12 s were used to compute the IPA and LHIPA, since memory load would be highest at the beginning, when participants started typing the sentence, and would gradually reduce as the participants continued typing. Recalling the sentence by typing it, similar to reporting the digit and word recall in the study by Kahneman and Beatty [33], could have led to reduction of cognitive load, exhibited by a decreasing pupil size during the reporting part of the experiment.

Trial difficulty did not have a significant effect on IPA for equally timed trial duration, $F(1,17) = 4.05$, $p = 0.06$, $\eta^2 = 0.05$, and, contrary to expectations, the IPA for more difficult tasks was lower (M = 0.22, SE = 0.01) than for easy tasks (M = 0.24, SE = 0.01). For the LHIPA, the effect was significant, $F(1,17) = 7.12$, $p < 0.05$, $\eta^2 = 0.01$, and, as expected, LHIPA produced significantly lower values for difficult tasks (M = 6.37, SE = 0.15) than for easy tasks (M = 6.50, SE = 0.15).

Brief Discussion

Results indicate that the LHIPA responds to task difficulty in an unrestricted (i.e., head and eye movement) applied setting when eye typing memorized sentences of varying complexity under variable light conditions (25-60 lux). As expected, LHIPA was inversely proportional to task difficulty (easy or difficult). Interestingly, a similar inversely proportional relation was exhibited by the LHIPA for perceived task difficulty, aligning well with self-reported task difficulty scores.

Contrary to hypothesis, results indicate that the IPA failed to produce a directly proportional response to either task difficulty or subjective perceived difficulty. Instead, the IPA produced the opposite, suggesting decrease in task difficulty and thus cognitive load. Why this should be is puzzling, although reasons for its failure may be rooted in unrestricted eye movement (off-axis distortion of the apparent pupil, the Pupil Foreshortening Effect), or variable luminance levels.

On average, response of LHIPA to task difficulty manifested significance early on, as early the first 12 seconds into the trials, based on our analysis of fastest eye-typing task completion. The IPA failed to respond significantly within this short time frame. This suggests that the LHIPA may require relatively little time to indicate load, which bodes well for potential future real-time applications.

GENERAL DISCUSSION

In all three experiments, performance measures (e.g., speed, accuracy) indicate manipulation of task difficulty, i.e., easy, difficult, and/or baseline. The IPA effectively reflected task difficulty in only one situation, where gaze was fixed at screen center. The LHIPA responded to task difficulty reliably in all three experiments, each successively less restrictive in terms of head and eye movement and luminance conditions.

It appears that high frequency pupil diameter oscillation, the artifact measured by the IPA, is, by itself, not a reliable indicator of task difficulty, i.e., cognitive load. In contrast, although the LHIPA performs a similar measurement of pupil diameter oscillation, it does so via measurement of the LF/HF ratio. The low frequency component provides a type of built-in baseline measurement that the IPA lacks. That is, the IPA, while measuring the pupil phasic response (HF), ignores the tonic component (LF). In using both, the LHIPA appears to provide a more reliable indicator of task difficulty, i.e., cognitive load.

LIMITATIONS

Because the IPA is based on measurement of pupillary oscillation, it was thought immune to effects of lighting (and sampling rate, e.g., as indicated by Bartels and Marshall's [6] Index of Cognitive Activity, or ICA). However, results from Experiment 2 suggest that the IPA may be susceptible to the *law of initial value* [42, 31], when the pupil is already somewhat dilated in very low-light (e.g., scotopic, <70 lux) conditions. When the pupil is already dilated (high tonic signal), the relative change in pupil diameter (phasic change) may be difficult to detect. Adaptive Gain Theory suggests that cognitive load via pupillometric measurement may in general be better detected under daylight (low mesopic–photopic, ≥ 70 lux) conditions. The response of eye-tracked pupillometric measures of cognitive load under varying luminance requires further investigation.

IMPLICATIONS FOR INTERACTION DESIGN

There is still a pressing need for non-invasive measures of cognitive load, so that interactive systems do not overload users, especially if the measurement can be made in real time.

Because of the local support of the wavelet transform, computation of the LHIPA shows that it may be possible to use a short-term filter to compute the low/high frequency ratio (LF/HF) of pupil diameter in real time.

The LF/HF ratio is thought to give an indication of phasic activity of the autonomic nervous system relative to tonic activity. This may in turn obviate the need for a separate baseline pupil diameter measurement as evidenced by the LHIPA response to task difficulty in the unrestricted eye-typing task.

```
import math, pywt, numpy as np

def lhipa(d):
  # find max decomposition level
  w = pywt.Wavelet('sym16')
  maxlevel = \
    pywt.dwt_max_level(len(d),filter_len=w.dec_len)

  # set high and low frequency band indeces
  hif, lof = 1, int(maxlevel/2)

  # get detail coefficients of pupil diameter signal d
  cD_H = pywt.downcoef('d',d,'sym16','per',level=hif)
  cD_L = pywt.downcoef('d',d,'sym16','per',level=lof)

  # normalize by 1/sqrt(2^j)
  cD_H[:] = [x / math.sqrt(2**hif) for x in cD_H]
  cD_L[:] = [x / math.sqrt(2**lof) for x in cD_L]

  # obtain the LH:HF ratio
  cD_LH = cD_L
  for i in range(len(cD_L)):
    cD_LH[i] = cD_L[i] / cD_H[((2**lof)/(2**hif))*i]

  # detect modulus maxima, see Duchowski et al. [15]
  cD_LHm = modmax(cD_LH)

  # threshold using universal threshold λ_univ = σ̂ sqrt(2 log n)
  # where σ̂ is the standard deviation of the noise
  λ_univ = \
    np.std(cD_LHm) * math.sqrt(2.0*np.log2(len(cD_LHm)))
  cD_LHt = pywt.threshold(cD_LHm,λ_univ,mode="less")

  # get signal duration (in seconds)
  tt = d[-1].timestamp() - d[0].timestamp()

  # compute LHIPA
  ctr = 0
  for i in xrange(len(cD_LHt)):
    if math.fabs(cD_LHt[i]) > 0: ctr += 1
  LHIPA = float(ctr)/tt

  return LHIPA
```

Listing. 1. LHIPA implementation.

CONCLUSION

We derived a novel wavelet-based method for computing the low/high frequency ratio of pupil oscillation termed the Low/High Index of Pupillary Activity, or LHIPA. Empirical results show that the LF/HF ratio of pupillometric oscillation expressed by the LHIPA is robust in its sensitivity to the measurement of cognitive load, in response to the presence of task difficulty. The LHIPA is not sufficiently sensitive to necessarily distinguish between levels of task difficulty, but it is more effective at detecting load than the similar Index of Pupillary Activity. Moreover, for small angles (i.e., $\sim 10^\circ$ visual angle), off-axis distortion of the pupil diameter is negligible in its effect on LHIPA although it appears to negatively affect the IPA. Robustness of the LHIPA response to task difficulty was shown in an applied setting (eye-typing) which demonstrates its potential for eventual use in real-time interactive systems.

APPENDIX

Listing 1 gives the Python 2.7 implementation of the LHIPA.

Acknowledgments

This work is supported in part by the U.S. National Science Foundation (grant IIS-1748380), Polskie Ministerstwo Nauki i Skzolnictwa Wyższego (Regional Initiative of Excellence,

grant 012-RID-2018/19), and by Bevica Fonden, Denmark. We thank reviewers for their suggestions for improvement.

REFERENCES

[1] Sylvia Ahern and Jackson Beatty. 1979. Pupillary Responses During Information Processing Vary with Scholastic Aptitude Test Scores. *Science* 205, 4412 (1979), 1289–1292.

[2] Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci. 2018. Cross-subject Workload Classification Using Pupil-related Measures. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. ACM, New York, NY, Article 4, 8 pages. DOI: http://dx.doi.org/10.1145/3204493.3204531

[3] Gary Aston-Jones and Jonathan D. Cohen. 2005. An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience* 28, 1 (2005), 403—450. DOI:http: //dx.doi.org/10.1146/annurev.neuro.28.061604.135709

[4] Miyuki Azuma, Takehiro Minamoto, Ken Yaoi, Mariko Osaka, and Naoyuki Osaka. 2014. Effect of memory load in eye movement control: A study using the reading span test. *Journal of Eye Movement Research* 7, 5 (2014), 1–9.

[5] Brian P. Bailey and Shamsi T. Iqbal. 2008. Understanding Changes in Mental Workload During Execution of Goal-directed Tasks and Its Application for Interruption Management. *ACM Trans. Comput.-Hum. Interact.* 14, 4, Article 21 (Jan. 2008), 28 pages. DOI: http://dx.doi.org/10.1145/1314683.1314689

[6] Mike Bartels and Sandra P. Marshall. 2012. Measuring Cognitive Workload Across Different Eye Tracking Hardware Platforms. In *ETRA '12: Proceedings of the 2012 Symposium on Eye Tracking Research & Applications*. ACM, Santa Barbara, CA.

[7] Jackson Beatty. 1982. Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin* 91, 2 (1982), 276–292.

[8] Jackson Beatty and Brennis Lucero-Wagoner. 2000. The Pupillary System. In *Handbook of Psychophysiology* (2nd ed.), John T. Cacioppo, Louis G. Tassinary, and Gary G. Bernston (Eds.). Cambridge University Press, 142–162.

[9] Carl-Hugo Björnsson. 1968. *Läsbarhet*. Seelig, Solna, Sweden.

[10] Deborah A. Boehm-Davis, Wayne D. Gray, Leonard Adelman, Sandra Marshall, and Robert Pozos. 2003. *Understanding and Measuring Cognitive Workload: A Coordinated Multidisciplinary Approach*. Technical Report AFRL-SR-AR-TR-03-0407 (ADA417743); Grant #49620-97-1-0353. AFOSR, Arlington, VA.

[11] Siyuan Chen and Julien Epps. 2014a. Efficient and Robust Pupil Size and Blink Estimation from Near-Field Video Sequences for Human-Machine Interaction. *IEEE Transactions on Cybernetics* 44, 12 (2014), 2356–2367. DOI:http://dx.doi.org/10.1109/TCYB.2014.2306916

[12] Siyuan Chen and Julien Epps. 2014b. Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *Human-Computer Interaction* 29, 4 (2014), 390–413. DOI: http://dx.doi.org/10.1080/07370024.2014.892428

[13] Peter Øvergård Clausen. 2018. *Manual for the experimental version of OptiKey*. Technical University of Denmark / DTU, Anker Engelunds Vej 1, Building 101A, 2800 Kgs. Lyngby, Denmark. https://github.com/GazeIT-DTU/OptiKey

[14] Benjamin Cowley, Marco Filetti, Kristian Lukander, Jari Torniainen, Andreas Henelius, Lauri Ahonen, Oswald Barral, Ilkka Kosunen, Teppo Valtonen, Minna Huotilainen, Niklas Ravaja, and Giulio Jacucci. 2016. The Psychophysiology Primer: A Guide to Methods and a Broad Review with a Focus on Human-Computer Interaction. *Foundations and Trends in Human-Computer Interaction* 9, 3-4 (2016), 151–308. DOI:http://dx.doi.org/10.1561/1100000065

[15] Andrew T. Duchowski, Krzysztof Krejtz, Izabela Krejtz, Cezary Biele, Anna Niedzielska, Peter Kiefer, Martin Raubal, and Ioannis Giannopoulos. 2018. The Index of Pupillary Activity: Measuring Cognitive Load Vis-à-vis Task Difficulty with Pupil Oscillation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, Article 282, 13 pages. DOI: http://dx.doi.org/10.1145/3173574.3173856

[16] Andrew T. Duchowski, Krzysztof Krejtz, Justyna Żurawska, and Donald House. 2019. Using Microsaccades to Estimate Task Difficulty During Visual Search of Layered Surfaces. *IEEE Transactions on Visualization and Computer Graphics* 213 (2019). DOI:http://dx.doi.org/10.1109/TVCG.2019.2901881

[17] Ralf Engbert and Reinhold Kliegl. 2003. Microsaccades uncover the orientation of covert attention. *Vision Research* 43 (2003), 1035–1045.

[18] Paul M. Fitts, Richard E. Jones, and John L. Milton. 1950. Eye Movements of Aircraft Pilots During Instrument-Landing Approaches. *Aeronautical Engineering Review* 9, 2 (1950), 24–29.

[19] Lex Fridman, Bryan Reimer, Bruce Mehler, and William T. Freeman. 2018. Cognitive Load Estimation in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, Article 652, 9 pages. DOI: http://dx.doi.org/10.1145/3173574.3174226

[20] Mark S. Gilzenrat, Sander Nieuwenhuis, Marieke Jepma, and Jonathan D Cohen. 2010. Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, affective & behavioral neuroscience* 10, 2 (2010), 252—269. DOI:http://dx.doi.org/10.3758/CABN.10.2.252

[21] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (2012), 759–765. https://www.cancer.org/cancer/breast-cancer/about/how-does-breast-cancer-form.html

[22] Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-physiological Measures for Assessing Cognitive Load. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*. ACM, New York, NY, 301–310. DOI: http://dx.doi.org/10.1145/1864349.1864395

[23] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 50. 904–908.

[24] Taylor R. Hayes and Alexander A. Petrov. 2016. Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Reearch* 48 (2016), 510–527. DOI: http://dx.doi.org/10.3758/s13428-015-0588-x

[25] Eckhard H. Hess and James M. Polt. 1964. Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science* 143, 3611 (March 1964), 1190–1192.

[26] Nina Hollender, Cristian Hofmann, Michael Deneke, and Bernhard Schmitz. 2010. Integrating cognitive load theory and concepts of human-computer interaction. *Computer in Human Behavior* 26, 6 (2010), 1278–1288.

[27] Jukka Hyönä, Jorma Tommola, and Anna-Mari Alaja. 1995. Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. *The Quarterly Journal of Experimental Psychology* 48, 3 (1995), 598–612.

[28] Robert J. K. Jacob and Keith S. Karn. 2003. Eye Tracking in Human-Computer Interaction and Usability Research: Ready to Deliver the Promises. In *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, Jukka Hyönä, Ralph Radach, and Heiner Deubel (Eds.). Elsevier Science, Amsterdam, The Netherlands, 573–605.

[29] Xianta Jiang, M. Stella Atkins, Geoffrey Tien, Roman Bednarik, and Bin Zheng. 2014. Pupil Responses During Discrete Goal-directed Movements. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, 2075–2084. DOI: http://dx.doi.org/10.1145/2556288.2557086

[30] Xianta Jiang, Bin Zheng, Roman Bednarik, and M. Stella Atkins. 2015. Pupil responses to continuous aiming movements. *International Journal of Human-Computer Studies* 83 (2015), 1–11.

[31] Putai Jin. 1992. Toward a Reconceptualization of the Law of Initial Value. *Psychological Bulletin* 111, 1 (1992), 176–184.

[32] Marcel Adam Just and Patricia A. Carpenter. 1976. Eye Fixations and Cognitive Processes. *Cognitive Psychology* 8, 4 (October 1976), 441–480.

[33] Daniel Kahneman and Jackson Beatty. 1966a. Pupil Diameter and Load on Memory. *Science* 154, 3756 (1966), 1583–1585.

[34] Daniel Kahneman and Jackson Beatty. 1966b. Pupillary Diameter and Load on Memory. *Science* 154 (1966), 1583–1585.

[35] Caitlin Kelleher and Wint Hnin. 2019. Predicting Cognitive Load in Future Code Puzzles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, Article 257, 12 pages. DOI: http://dx.doi.org/10.1145/3290605.3300487

[36] Peter Kiefer, Ioannis Giannopoulos, Andrew T. Duchowski, and Martin Raubal. 2016. Measuring Cognitive Load for Map Tasks through Pupil Diameter. In *Proceedings of the Ninth International Conference on Geographic Information Science (GIScience 2016)*. Springer International Publishing.

[37] Wayne K. Kirchner. 1958. Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology* 4, 58 (1958), 352—358. DOI: http://dx.doi.org/10.1037/h0043688

[38] Jeff Klingner, Rakshit Kumar, and Pat Hanrahan. 2008. Measuring the Task-Evoked Pupillary Response with a Remote Eye Tracker. In *ETRA '08: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications*. ACM, New York, NY, 69–72. DOI: http://dx.doi.org/10.1145/1344471.1344489

[39] 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 (2011), 323–332. DOI:http://dx.doi.org/doi:10.1111/j.1469-8986.2010.01069.x

[40] Krzysztof Krejtz, Andrew T. Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. 2018. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE* 13, 9 (September 2018), 1–23. DOI: http://dx.doi.org/10.1371/journal.pone.0203629

[41] Jan-Louis Kruger, Esté Hefer, and Gordon Matthew. 2013. Measuring the Impact of Subtitles on Cognitive Load: Eye Tracking and Dynamic Audiovisual Texts. In *Proceedings of the 2013 Conference on Eye Tracking South Africa (ETSA '13)*. ACM, New York, NY, 75–78. DOI:http://dx.doi.org/10.1145/2509315.2509333

[42] John I. Lacey. 1956. The Evaluation of Autonomic Responses: Toward a General Solution. *Annals of the New York Academy of Sciences* 67, 5 (1956), 125–163. DOI: http://dx.doi.org/10.1111/j.1749-6632.1956.tb46040.x

[43] Yongqiang Lyu, Xiaomin Luo, Jun Zhou, Chun Yu, Congcong Miao, Tong Wang, Yuanchun Shi, and Ken-ichi Kameyama. 2015. Measuring Photoplethysmogram-Based Stress-Induced Vascular Response Index to Assess Cognitive Load and Stress. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, 857–866. DOI: http://dx.doi.org/10.1145/2702123.2702399

[44] Päivi Majaranta, I. Scott MacKenzie, Anne Aula, and Kari-Jouko Räihä. 2003. Auditory and Visual Feedback During Eye Typing. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems*, Vol. 1. 1–3. DOI: http://dx.doi.org/10.1145/765978.765979

[45] Sandra P. Marshall. 2000. Method and Apparatus for Eye Tracking Monitoring Pupil Dilation to Evaluate Cognitive Activity. US Patent No. 6,090,051. (18 July 2000).

[46] Sandra P. Marshall. 2002. The Index of Cognitive Activity: Measuring Cognitive Workload. In *Proceedings of the 7th Human Factors Meeting*. IEEE.

[47] Sharon Oviatt. 2006. Human-centered Design Meets Cognitive Load Theory: Designing Interfaces That Help People Think. In *Proceedings of the 14th ACM International Conference on Multimedia (MM '06)*. ACM, New York, NY, 871–880. DOI: http://dx.doi.org/10.1145/1180639.1180831

[48] Oskar Palinko and Andrew L. 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

[49] W. Scott Peavler. 1974. Pupil Size, Information Overload, and Performance Differences. *Psychophysiology* 11, 5 (1974), 559–566. DOI: http://dx.doi.org/10.1111/j.1469-8986.1974.tb01114.x

[50] Jonathan W Peirce. 2007. PsychoPy–Psychophysics Software in Python. *Journal of neuroscience methods* 162, 1 (2007), 8–13.

[51] Vsevolod Peysakhovich. 2016. *Study of pupil diameter and eye movements to enhance flight safety*. Ph.D. Dissertation. Université de Toulouse, Toulouse, France.

[52] Vsevolod Peysakhovich, Fran cois Vachon, and Frédéric Dehais. 2017. The impact of luminance on tonic and phasic pupillary responses to sustained cognitive load. *International Journal of Psychophysiology* 112 (2017), 40–45. DOI:http://dx.doi.org/https://doi.org/10.1016/j.ijpsycho.2016.12.003

[53] Bastian Pfleging, Drea K. Fekety, Albrecht Schmidt, and Andrew L. Kun. 2016. A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, 5776–5788. DOI: http://dx.doi.org/10.1145/2858036.2858117

[54] Tepring Piquado, Derek Isaacowitz, and Arthur Wingfield. 2010. Pupillometry as a Measure of Cognitive Effort in Younger and Older Adults. *Psychophysiology* 47, 3 (2010), 560–569. DOI:http://dx.doi.org/doi:10.1111/j.1469-8986.2009.00947.x

[55] R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ ISBN 3-900051-07-0.

[56] Stuart R. Steinhauer, Greg J. Siegle, Ruth Condray, and Misha Pless. 2004. Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International Journal of Psychophysiology* 52, 1 (2004), 77–86. DOI:http://dx.doi.org/https://doi.org/10.1016/j.ijpsycho.2003.12.005 Pupillometric Measures of Cognitive and Emotional Processes.

[57] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science* 12, 2 (1988), 257–285.

[58] John Sweller. 1994. Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction* 4, 4 (1994), 295–312.

[59] Beste F. Yuksel, Kurt B. Oleson, Lane Harrison, Evan M. Peck, Daniel Afergan, Remco Chang, and Robert J. K. Jacob. 2016. Learn Piano with BACh: An Adaptive Learning Interface that Adjusts Task Difficulgty based on Brain State. In *Human Factors in Computing Systems: CHI '16 Conference Proceedings*. ACM, San Jose, CA.