Cognitive Load during Eye-typing

Tanya Bafna
taba@dtu.dk
Technical University of Denmark
Kongens Lyngby

John Paulin Hansen
jpha@dtu.dk
Technical University of Denmark
Kongens Lyngby

Per Bækgaard
pgba@dtu.dk
Technical University of Denmark
Kongens Lyngby

ABSTRACT

In this paper, we have measured cognitive load during an interactive eye-tracking task. Eye-typing was chosen as the task, because of its familiarity, ubiquitousness and ease. Experiments with 18 participants, where they memorized and eye-typed easy and difficult sentences over four days, were used to compare the difficulty levels of the tasks using subjective scores and eye-metrics like blink duration, frequency and interval and pupil dilation were explored, in addition to performance measures like typing speed, error rate and attended but not selected rate. Typing performance lowered with increased task difficulty, while blink frequency, duration and interval were higher for the difficult tasks. Pupil dilation indicated the memorization process, but did not demonstrate a difference between easy and difficult tasks.

CCS CONCEPTS

• **Human-centered computing** → **Accessibility systems and tools**; **Heuristic evaluations**; *HCI theory, concepts and models*; *Empirical studies in HCI*; Laboratory experiments.

KEYWORDS

cognitive load, eye-typing, pupil size, working memory, blinks

ACM Reference Format:
Tanya Bafna, John Paulin Hansen, and Per Bækgaard. 2020. Cognitive Load during Eye-typing. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers), June 2–5, 2020, Stuttgart, Germany.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3379155.3391333

1 INTRODUCTION

Detection of cognitive load during the execution of a task is useful in several applications like education, armed forces, driving [Lieberman et al. 2002; Lohani et al. 2019; Sweller 2011], etc. Cognitive load, defined by the task difficulty, influences the performance on the task, in addition to psychophysiological measures like pupil size and blinks [Appel et al. 2018; Bækgaard et al. 2016; Haapalainen et al. 2010; Kahneman and Beatty 1966; Tanaka and Yamaoka 1993]. Tasks commonly used to detect cognitive load include mental arithmetic, n-back, visual search tasks, or application based tasks like driving, code comprehension, etc [Begel and Vrzakova 2018; Fridman et al. 2018; Lohani et al. 2019].

With eye-tracking becoming ubiquitous not only in wearable technology for virtual and augmented reality but also on laptops and mobile devices [Papoutsaki et al. 2016], it is of great interest to detect cognitive load in tasks that involve eye-based interactions. Since eye-typing is a task that can be implemented and can find application in all technologies that use eye-tracking, including virtual and augmented reality [Hansen et al. 2018; Ma et al. 2018; Polacek et al. 2017], we have designed experiments on cognitive load detection during eye-typing tasks.

Cognitive load detection using physiological measures during eye-typing has not been a topic of research as yet, and is a novel direction of research. We have examined working memory load of remembering sentences, the difficulty of which has been manipulated to control the cognitive load, and then eye-typing them from memory, a task we have called typing-from-memory task. Applications for this can be to not only design keyboards [Sengupta et al. 2017], but also design interfaces involving key selection [Blattgerste et al. 2018], develop adaptive interfaces [Katidioti et al. 2016], detect cognitive load for people with neuro-muscular problems who use eye-typing systems on a daily basis [Kane and Morris 2017] and overall replace self-reporting.

Eye-typing is a key method for people with neurological disorders to communicate. However, its uses can be extended to hands-free typing in other situations like when using augmented reality for some training or surgical procedures [Webel et al. 2013]. One of the first methods for eye-typing is dwell-time selection, where each key is focused on for a particular amount of time, called the dwell-time, in order to select that key [Majaranta 2009]. The most common keyboard used for eye-typing is the QWERTY keyboard, due to its familiarity for an ever increasing digital world [Kristensson and Vertanen 2012; Polacek et al. 2013; Räihä and Ovaska 2012; Salvucci 2000; Tuisku et al. 2013].

Eye-metrics using eye-tracking is promising for non- invasive cognitive load detection. Pupil size, blink frequency and blink duration and blink interval are some of the commonly used eye-metrics to indicate cognitive load. Task-evoked pupillary response (TEPR) is proposed as an index for task difficulty inducing cognitive load [Kahneman and Beatty 1966], with more difficult problems resulting in increased pupil dilations.

Derived measures from blinks also have a relationship with cognitive load. Blink frequency is expected to reduce with increasing task difficulty, as for visual tasks like reading, blinks are inhibited at a greater extent for difficult tasks. Blink duration, also an indicator for cognitive load, reduces in length with increasing task difficulty [Bækgaard et al. 2016; Kosch et al. 2018]. Blink interval increases with increase in task difficulty [Ryu and Myung 2005].

ETRA '20 Full Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7133-9/20/06…$15.00
https://doi.org/10.1145/3379155.3391333

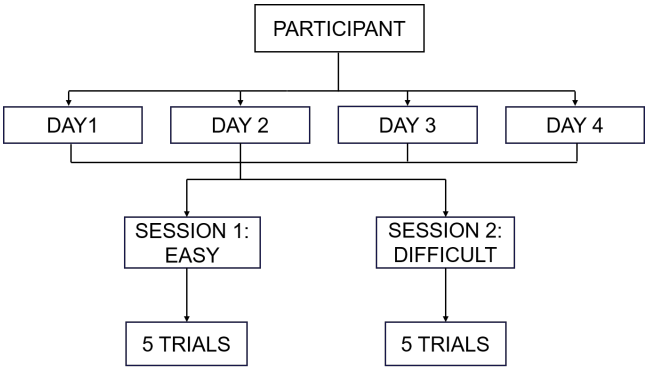


Figure 1: Schematic of the experimental procedure

Typing speed, error rate — corrected and uncorrected — and read text events rate are the commonly used performance measures for typing tasks. Read text events are the events of reading the text already typed. Higher uncertainty about the typing method results in increased events per character typed. Additionally, attended but not selected rate is a performance measure used in eye-typing, where a key is focused on for a short time, but not selected [Majaranta 2011].

[Biswas et al. 2016] is one of the first papers to combine gaze interaction with investigation of cognitive load, and have compared using eye-tracking and mouse for the level of cognitive load, via pupil dilation and blinks, during a pointing and selecting task. [Sengupta et al. 2017] have used electroencephalography-based cognitive load detection during eye-typing, with the goal of designing the keyboard layout that induces the least cognitive load on the users. However, this paper is the first attempt at detecting cognitive load during an eye-typing task using eye-based measures.

The contributions of this present paper are to test an eye-typing experimental paradigm for cognitive load detection and to explore the usability of the existing performance measures and eye-metrics for cognitive load detection during an interactive eye-tracking task.

2 METHODS

The goal of the study was to detect the effect of task difficulty and cognitive load, induced using memorization of easy and difficult sentences, on eye-based physiological variables during a complex task of gaze-interaction, namely eye-typing. There is no standardized experiment design for cognitive load detection using eye-typing, and so the experiment described here had a basis in longitudinal eye-typing experiments [Mott et al. 2017], to allow for the measurement of the cognitive load, once the learning effect of the keyboard and the task had taken place.

2.1 Experiment Design

The experiment was repeated on 4 days for every participant, and is depicted in Figure 1. The multi-day experiment was designed so as to allow the participants to get accustomed to the method of eye-typing and memorization. The procedure of the experiment was as follows: after the participants signed the consent form, they were given written instructions on the experimental procedure. They were instructed to type as fast as they could, while being as accurate as possible. The participants were provided an incentive of extra compensation for every day in which they had the highest typing speed.

The experiment consisted of two sessions on each day - easy and difficult, resulting in a total of eight sessions throughout the experiment for one participant. A session was composed of five typing-from-memory trials The order of the easy and difficult sessions were counter-balanced for each participant. At the beginning of the first day, they performed a trial session with two typing-from-memory trials.

The complete experiment could be performed in Danish or English, and the language could be chosen by the participant on the first day.

The typing-from-memory task is defined here as eye-typing sentences after reading and memorizing them. The procedure of each trial was composed of reading and memorizing the sentence, followed by eye-typing the sentence and thereafter answering the *effort* question from the NASA-TLX questionnaire, as shown in Figure 2. The scale was adapted from [Mott et al. 2017] to be between 1 and 7. The inter-trial time was set to 5 s.

The sentences were taken from the Lepzig corpus of mixed sources of sentences [Goldhahn et al. 2012]. The difficulty level of the sentences was determined by the Readability score, as computed using the Lasbarheitsindex (LIX) score [Björnsson 1968]. The LIX score defines the readability of the text, based on the number of words (A), number of periods (B) and the number of long words (C), containing more than 6 letters in the text:

$$LIX = A/B + (Cx100)/A \quad (1)$$

For the sentences selected for the experiment, B was always taken as 1. The sentences were divided by using an upper cutoff of 30 on the LIX score for easy sentences and a lower cutoff of 60 for difficult sentences. An example of an easy sentence was *"I would like you to have a look at this, but if you get stuck example 1 is provided."*. An instance of the difficult sentence was *"The large-scale orchestral introduction contains the famous chromatic 'yearning' theme at the beginning, which becomes the thematic essence of the whole opera."*.

The keyboard layout was as shown in Figure 1. The top row provided four word suggestions, to aid completion of the current word.

2.2 Experiment Setup

The experiments were performed using Tobii Eye Tracker 4C (sampling frequency: 90 Hz). The eye-typing interface used was OptiKey, an open-source on-screen gaze-based keyboard developed as Windows Presentation Foundation (WPF) application. An experimental version of the keyboard, which allowed us to display the text to be eye-typed as well as log the text typed, was used for the experiment. The experimental procedure was incorporated into the keyboard and displayed to the participants on a computer with 1920 x 1080 resolution.

The experimental setting was a laboratory with no entry of natural light and luminance varying between different days and participants, to $25 - 60$ lux at the computer screen. However, the luminance was kept constant for both sessions in a day for each participant.

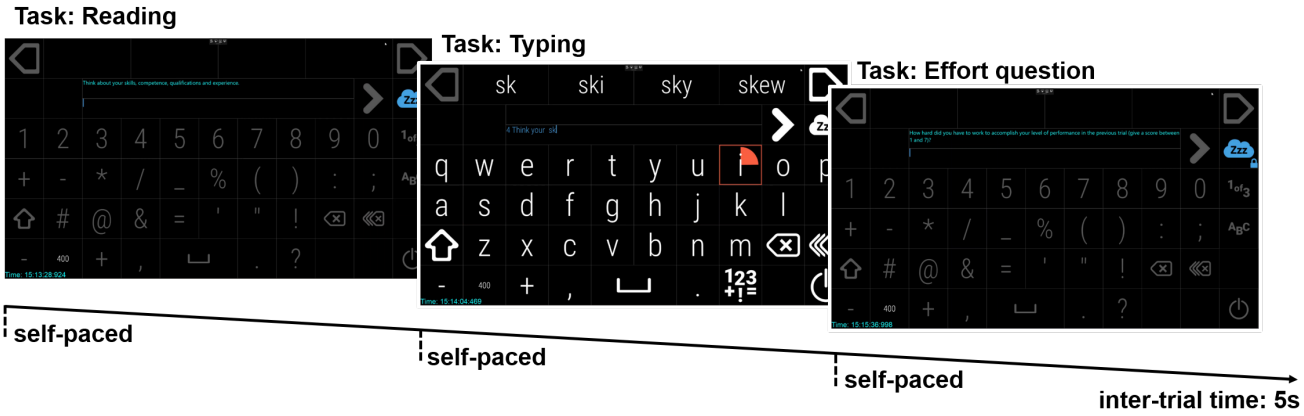


Figure 2: Schematic of the trial procedure

2.3 Participants

The experiments were conducted with 19 volunteers from Technical University of Denmark. A participant dropped out after only one day, and so the data from 18 participants (9M, with age in the range of [21:31] years, $M = 25.5, SD = 2.38$) is presented in this paper. Ten participants performed the experiment in English, and the rest in Danish.

2.4 Data Analysis

The data obtained from the eye-tracker was logged using Tobii Pro SDK. Blinks were defined as missing data from the eye-tracker which have durations between 75 *ms* and 500 *ms*. Data 200 *ms* before and after the blinks was removed, and the data was linearly interpolated, followed by application of hampel filter with a length of five samples before and after the current sample, and values greater than three times the standard deviation treated as an outlier and modified. For the results on eye-metrics, sessions that did not match a threshold of 0.75 correlation between the right and left pupil size were removed.

Eye-typing data was logged from the on-screen keyboard in form of five different logging files. The raw data log files were used to analyze and obtain various features of pupils, blinks and performance. The independent variables between each participant was the language chosen by the participants for the experiment, and within each participant was session difficulty, day number and session number on each day. The dependent variables are divided into four categories — subjective scores, pupil size, blinks and typing performance.

The relationships between the dependent variables and independent variables were analyzed using linear mixed effects model (LMM), due to the within-subject experiment design. The fixed effects were the session difficulty and language as categorical variables, day number as a numerical variable, session number nested under the day number, also as a numerical variable. Additionally, four time periods during the trial were used as a reference to track the progress of the pupil size. These time periods were obtained for the following phases of the trials — when reading started, reading ended, writing started and writing ended. These phases were considered to be a fixed effect for the linear models with relative pupil size as the dependent variable. Fixed effects were considered only when $p < 0.05$. Random intercepts and slopes were included for each participant for the independent variable session difficulty unless the model failed to converge, and in that case, only the random intercept for the participants were included in the models. The analysis was performed using the packages lme4[Bates et al. 2014] and lmerTest[Kuznetsova et al. 2017] in R [R Core Team 2013]. Effect sizes were computed using the package r2glmm [Jaeger 2017] using the Nakagawa and Schielzeth approach [Nakagawa and Schielzeth 2013].

Subjective scores consists of the score on the *effort* question of NASA-TLX, asked after every trial, which we have termed *perceived difficulty* scores.

Relative pupil size was computed using subtractive baseline correction following [Mathôt et al. 2018], where baseline pupil size was computed during the 5 *s* inter-trial time. The section on pupil size in the results is based on the mean relative pupil size for a duration of 300 *ms* during the designated period — when participants started reading the sentence, when participants finished reading the sentence, when participants started eye-typing and when they ended eye-typing. The duration for mean relative pupil size during these phases was adjusted for blinks, by finding a continuous duration of 300 *ms* that did not include interpolation of blinks. This blink-adjusted method for 300 *ms* was also used for computation of the baseline pupil size within the 5 *s* inter-trial period. Weighted mean of the pupil size was computed for the relative pupil size, with the weights computed from the normalization of inverse of the rolling standard deviations of the left and right pupil size.

The blink category of results consists of blink frequency - the number of blinks in the duration of typing, blink duration - average of the duration of the blinks during the trial and blink interval - average time between consecutive blinks in the trial.

Performance measures consist of typing speed, corrected and uncorrected error rate, read text events rate and attended but not selected rate. Typing speed is reported using normalized words per

minute (WPM), where one word consists of five characters, including space. Since word suggestions were activated and they might skew the letter count, each selection of suggestion was counted as 1 letter, and the typing speed reported depends on the key selection speed. For the sake of simplicity, we continue to call it typing speed. Uncorrected error rate was computed using the Levenshtein distance at character and word levels, with equal weighting to each of them. Corrected error rate was the ratio of the number of times *Backspace* key was selected to the number of characters typed. Read text events rate was computed using the ratio of the time spent reading the already eye-typed text to the total time of the trial. Attended but not selected rate was computed by counting the number of keys where the selection was started but not completed and selected.

3 RESULTS

A total of 729 trials were recorded from 18 participants performing 10 trials each day for 4 days. An extra 9 trials occurred as a result of a setting in the experimental keyboard. The results presented for eye-metrics — blinks and pupil size, are from 709 trials, as 4 sessions had to be removed due to low correlation between the right and left pupil size.

3.1 Subjective Scores

As manipulation check, to ensure that the cognitive load imposed with the difficult sentences was perceived to be difficult, before testing hypotheses on eye metrics, we compared the perceived difficulty given after every trial to the two categories of session difficulty — easy and difficult.

Marginal means of the perceived difficulty scores for the session difficulty levels (easy: $M = 2.95, SE = 0.179$, difficult: $M = 4.72, SE = 0.180$)) decreased over the days ($M_{diff} = 0.338$) per day, $SE = 0.182$). As expected, we found a significant main effect of session difficulty on the perceived difficulty scores ($\chi^2(1) = 257.41, p < 0.001, \eta^2 = 0.26$) increasing by about 1.77 ($SE = 0.10$) for difficult sessions. Crucially, we also found a significant effect of the days on the difficulty scores ($\chi^2(1) = 55.271, p < 0.001, \eta^2 = 0.06$) reducing over the days by about 0.34 ($SE = 0.05$) per day, showing a learning effect of the memorization task (see Figure 3).

3.2 Eye typing and Dwell time

The participants could change the dwell time setting from the keyboard, which was initially set to 800 *ms*. Additionally, there was an activation time, set to a constant 250 *ms*, which was the minimum amount of time required to activate a key before the visual feedback would be available, and which could not be changed by the participants. The dwell time presented here is the total time the participants had to dwell on a key to select it, which was the sum of the activation time and the additional time they could set on the keyboard. It is calculated as a weighted average of the dwell times used during a session by each participant, with the weights of the time duration for which the dwell time was chosen.

The average dwell time was 768.26 *ms* ($SE = 274.17$), with 12 out of 18 participants reducing the dwell time from the initial setting. Out of those, five ended up with a dwell time between 350 – 450 *ms*. Their average typing speed and corrected error rate were 14.659 *WPM* ($SE = 0.342$) and 9.543% ($SE = 2.163$). There were two participants who reduced the dwell time to below 350 *ms* and typed with a dwell time of 250 *ms*, which is the same as the activation time. The session with the highest typing speed was the one with a weighted average of 256.59 *ms* dwell time, resulting in a typing speed of 17.424 *WPM* but also a second highest corrected error rate of 59.499%. There were two participants who increased the dwell time, and only later found out, that they had done so by mistake.

The dwell time range of 250 – 350 *ms* resulted in the highest typing speed, see Figure 4, however, the corrected error rate for the same was also the highest and more than 3 times that of the dwell time range of 350 – 450 *ms*. Based on this, a dwell time in the range of 350 – 450 *ms* appears to give the best performance of eye-typing in our setup.

3.3 Performance

The following performance measures are analyzed and reported in this section — typing speed, uncorrected error rate and two

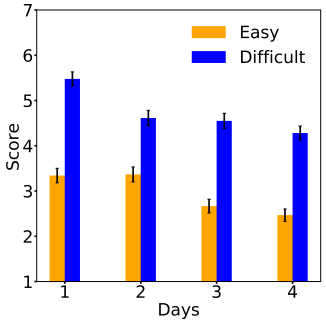
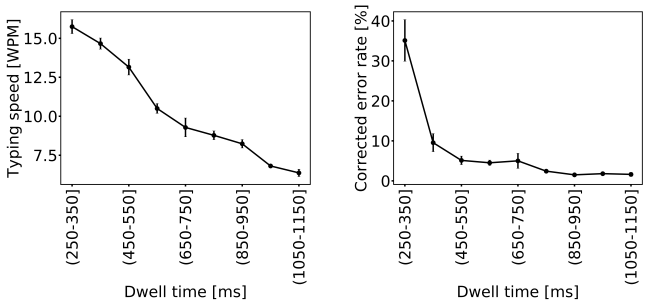
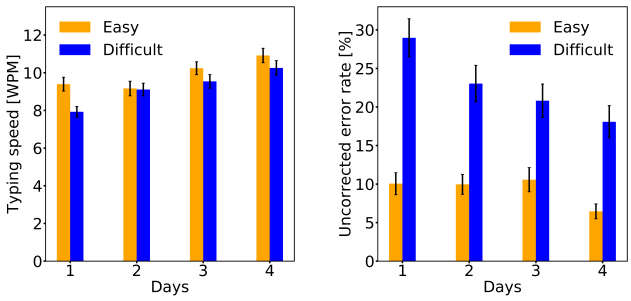


Figure 3: Subjective perceived difficulty scores for easy and difficult sessions over the 4 days of the experiment. Error bars indicate standard errors



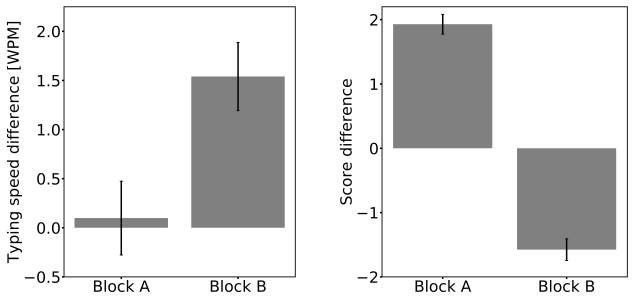
(a) Typing speed **(b) Corrected error rate**

Figure 4: Performance measures for the dwell times selected. Error bars indicate standard errors



(a) Typing speed (b) Error rate

Figure 5: Performance measures for easy and difficult sessions over the 4 days experiment. Error bars indicate standard errors



(a) Typing speed (b) Perceived difficulty scores

Figure 6: Difference of typing speed and perceived difficulty scores between session 2 and session 1 for blocks A and B, where:
block A = Session 1: easy, Session 2: difficult and
block B = Session 1: difficult, Session2: easy
Error bars indicate standard errors given by the expression for error propagation: $\sqrt{SE^2_{Session1} + SE^2_{Session2}}$

standard measures of eye-typing; read text events rate and attended but not selected rate [Majaranta 2011].

Marginal means for typing speed estimated for different levels of session difficulty were — (Easy: $M = 9.921\ WPM, SE = 0.187$, Difficult: $M = 9.207\ WPM, SE = 0.177$). There was an increase in typing speed each day ($M_{diff} = 0.633\ WPM$ per day, $SE = 0.358$) and an overall increase in the typing speed during the second session from the first, within each day ($M_{diff} = 0.810\ WPM, SE = 0.257$). Moreover, when typing an easy session first, the typing speed changed only marginally in the second, difficult session ($M_{diff} = 0.099\ WPM, SE = 0.375$), but when typing the difficult session first, the typing speed during the second, easy session, was higher ($M_{diff} = 1.539\ WPM, SE = 0.347$), see Figure 6a. On performing a linear mixed model analysis, there were found to be significant effects of session difficulty ($\chi^2(1) = 13.084, p < 0.001, \eta^2 = 0.007$), increasing by 1.107 WPM ($SE = 0.294$) for easy sessions. Additionally, day number had an effect ($\chi^2(1) = 12.764, p < 0.05, \eta^2 = 0.003$), increasing the typing speed by approximately 0.262 WPM ($SE = 0.093$) each day. Session number ($\chi^2(1) = 36.102, p < 0.001, \eta^2 = 0.013$) also had a significant effect on typing speed and increased by 0.342 WPM ($SE = 0.058$) in the second session within each day. Finally, there was a significant interaction between session difficulty and session number within each day ($\chi^2(1) = 4.901, p < 0.05, \eta^2 = 0.002$).

Uncorrected error rates were different for the levels of session difficulty, as estimated using the marginal means (Easy: $M = 9.283\%, SE = 0.670$, Difficult: $M = 22.714\%, SE = 1.155$). For each day, there was a reduction in the overall error rate ($M_{diff} = 2.305\%$ per day, $SE = 1.983$), and the error rate for the difficult sessions ($M_{diff} = 3.632\%$ per day, $SE = 3.205$). For easy sessions, there was no pattern observed over the four days, see Figure 5. LMM revealed significant effects of session difficulty ($\chi^2(1) = 22.447, p < 0.001, \eta^2 = 0.049$), lowering the uncorrected error rate by about 19.7% ($SE = 3.148$) for easy sessions. The day number had a significant effect on the uncorrected error rate ($\chi^2(1) = 19.074, p < 0.001, \eta^2 = 0.022$), lowering by 3.377% ($SE = 0.688$) each day. There was also a significant effect of their interaction ($\chi^2(1) = 6.498, p < 0.05, \eta^2 = 0.006$).

There was no difference obtained for the corrected error rate for the session difficulty (Easy: $M = 5.873\%, SE = 0.647$, Difficult: $M = 5.912\%, SE = 0.564$).

Read text events rate was not found to be affected by the session difficulty level (Easy: $M = 0.062, SE = 0.002$, Difficult: $M = 0.057, SE = 0.002$).

The rate of keys attended but not selected were higher for the difficult session, as seen by the estimated marginal mean values (Easy: $M = 0.537, SE = 0.022$, Difficult: $M = 0.615, SE = 0.025$). The rate reduced each day ($M_{diff} = 0.126$ per day, $SE = 0.043$) and was lower for the second session of the day than the first ($M_{diff} = 0.087, SE = 0.815$). There was an effect of session difficulty ($\chi^2(1) = 4.271, p < 0.05, \eta^2 = 0.007$), where the rate of keys attended not selected lowered for easy sessions by about 0.073 ($SE = 0.033$). The rate of keys attended but not selected were affected by the day ($\chi^2(1) = 32.2, p < 0.001, \eta^2 = 0.023$), lowering by 0.094 ($SE = 0.016$) each day. Lastly, session number also affected the rate ($\chi^2(1) = 6.69, p < 0.01, \eta^2 = 0.005$), reducing by about 0.022 ($SE = 0.008$) in the second session from the first.

3.4 Pupil Size

Pupil size for the task of reading a sentence, memorization and eye-typing was examined. It was expected that the pupil dilates from the time the participants start reading the sentence to when they remember it, and then it constricts during the typing task, similar to the working memory test with numbers [Kahneman and Beatty 1966]. Furthermore, it was also expected that the pupil dilation during reading is higher for difficult sentences compared to easy sentences.

Mean pupil dilation across all trials is shown in Figure 7, divided into the phases of — reading starts, reading ends, writing starts and writing ends. The relative pupil size was calculated from the baseline pupil size measured during the 5s break between trials.

The pupil size was computed in a the period of 300 *ms* closest to the event reported during which no blink interpolation took place.

The relative pupil size was compared for all the phases (Reading starts: $M = 0.113, SE = 0.010$, Reading ends: $M = 0.100, SE = 0.011$, Writing starts: $M = 0.074, SE = 0.012$, Writing ends: $M = -0.076, SE = 0.011$) and session difficulty (Easy: $M = 0.046, SE = 0.016$, Difficult: $M = 0.062, SE = 0.019$). LMM revealed a significant effect of the phase ($\chi^2(3) = 47.899, p < 0.001, \eta^2_{readingends} = 0.002, \eta^2_{writingstarts} = 0.001, \eta^2_{writingends} = 0.006$) on the relative pupil size and there was interaction between the phases and the days ($\chi^2(3) = 121.6, p < 0.001, \eta^2_{readingstarts} = 0.001, \eta^2_{readingends} = 0.001, \eta^2_{writingstarts} = 0.001, \eta^2_{writingends} = 0.001$), however, the effect of the session difficulty was not significant ($\chi^2(1) = 2.028, p = 0.154$).

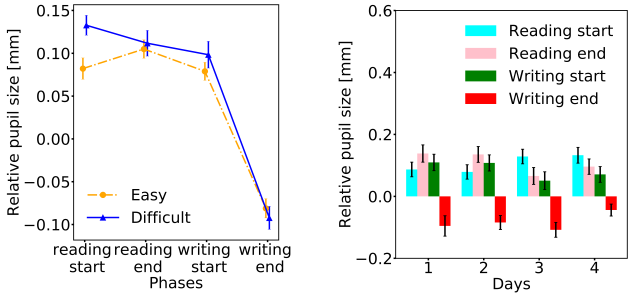
3.5 Blinks

The following blink properties were compared to the independent variables — blink frequency, blink duration, inter-blink duration.

Blink frequency was affected by the session difficulty level. The marginal means are estimated — (Easy: $M = 5.668, SE = 0.286$, Difficult: $M = 6.778, SE = 0.328$), and the effect of the session difficulty was significant ($\chi^2(1) = 4.088, p < 0.05, \eta^2 = 0.013$) using LMM, lowering the blink frequency by about $1.11/min$ ($SE = 0.435$) for easy sessions.

Blink duration was higher for difficult sessions than easy sessions, as estimated by the marginal means (Easy: $M = 0.117\ s, SE = 0.003$, Difficult: $M = 0.133\ s, SE = 0.002$). On analyzing using LMM, there was a significant effect of session difficulty ($\chi^2(1) = 7.228, p < 0.01, \eta^2 = 0.027$), reducing the blink duration by $0.02\ s$ ($SE = 0.01$)for easy sessions.

Blink intervals were higher for the difficult sessions, depicted using the estimated marginal means (Easy: $M = 7.280\ s, SE = 0.463$, Difficult: $M = 11.240\ s, SE = 0.539$). LMM showed that there was a significant effect of session difficulty ($\chi^2(1) = 10.901, p < 0.001, \eta^2 = 0.04$) on the blink interval, lowering them by $3.838\ s$ ($SE = 0.998$)for easy sessions.



(a) Relative pupil size over the phases of a trial **(b) Relative pupil size over days**

Figure 7: Relative pupil size for different phases and days. Error bars indicate standard errors

4 DISCUSSION

We have created an experimental design for cognitive load inducement and detection during an interactive eye tracking task. The eye tracking task selected was an eye-typing task in combination with a working memory task for extrinsic cognitive load. Eye and blink measures were analyzed for easy and difficult tasks, in addition to performance measures and subjective cognitive load ratings. Performance measures and subjective ratings detect a clear difference between easy and difficult tasks, indicating that the task design has an effect on the cognitive load.

The dwell time adjustment by the participants resulted in five out of eighteen participants coming down to a dwell time between 350 *ms* and 450 *ms* by day four. The average typing speed and corrected error rate for these, 14.659 *WPM* ($SE = 0.342$) and 9.543% ($SE = 2.163$), is comparable to the standard dwell-time eye-typing method in [Pi and Shi 2017]. The dwell-time of 256.59 *ms* resulted in a typing speed of 17.424 *WPM*, comparable to the typing speed obtained by probabilistic adjustment of dwell-time of 300 *ms* – 18.4 *WPM* in [Pi and Shi 2017].

According to cognitive load theory, two types of cognitive load were involved in this experiment - intrinsic and extrinsic. Intrinsic cognitive load, caused majorly by learning the method of eye-typing, clearly reduced over the days, as seen by the improvement in the typing speed for both easy and difficult sessions (see Figure 5a). The reduction in pupil dilation over the days, depicted in Figure 7, could also be an indication of this. Extrinsic cognitive load was caused by the task of memorization sentences of easy and difficult nature. A slight learning effect of the memorization task is also noticed by the reduction in the uncorrected error rate for the difficult sessions over the days. Reduction in the perceived difficulty scores over the days can be attributed to the overall improvement in performance.

The typing speed, session difficulty and the order of the sessions show an interesting pattern, as seen in Figure 6. Every time that the easy session was the first session of the day, the typing speed of the difficult session that followed did not increase significantly, although the perceived difficulty increased. However, when the difficult session was the first one in the day, the typing speed increased significantly during the second, easy, session, despite a reduction in the perceived difficulty. This could have an impact in the classrooms, where a teaching strategy could be formed, depending on the learning goals. If the goal would be to understand a difficult topic, teaching easy materials before difficult materials could result in slight increase in effort applied, but with the same performance on the difficult material as the easy one. However, if the learning goal would be to improve performance on a known or an easy task, completing a difficult task before performing the easy one could increase the performance on the easy task. Further research would be needed to elaborate on these results.

Pupil dilation was affected by the task of memorization, with a higher pupil dilation when the participants had finished memorization of the sentence, and pupil constriction as an effect of finishing typing of the sentence. This confirms the pupillary response during the word memory task tested by [Kahneman and Beatty 1966], where the pupil size increased as a function of listening to the digits to be remembered in a digit recall memory task, and the pupil size decreased as the digits were reported.

Observing a difference in pupil dilation as a result of the task difficulty (which was sentence complexity in our case) was one of the goals of the experiment, with the expectation of a higher pupil dilation for the difficult sessions compared to the easy sessions. [Kahneman and Beatty 1966] also showed a difference between the peak pupillary dilation, when the participants were asked to remember four or five or six or seven digits, especially during the pause between presentation of digits and reporting. However, pupil dilation, in our experiment, did not indicate a difference between easy and difficult sessions. One explanation could be that the task of reading was self-paced, which allowed for enough time to also remember the difficult sentences after spending enough time on it. Another reason could be that the easy level was also quite difficult, and so the two levels of difficulty were not enough to model the pupil dilation. For an English sentence with LIX score ∼78, the relative pupil size at the beginning of typing was 0.218 mm ($SE = 0.058$) and for an English sentence with LIX score ∼3, the relative pupil size as the typing start was 0.412 mm ($SE = 0.091$), as obtained from four participants each. The lower pupil dilation for a more difficult sentence could indicate that the participants already thought they would not remember the sentence, and probably gave up before starting to type. Multiple levels of difficulty in between the two levels would help place the effort, pupil dilation and the performance better on the existing models like that of [Huang et al. 2006]. Modifying the experiment design in these two aspects could aid in indicating an effect on pupil dilation.

Although pupil dilation did not successfully indicate any difference between the easy and difficult sessions, blink interval was an indicator of difficulty levels in the way expected. However, blink frequency and blink duration did not follow the known convention of blink suppression with increase in difficulty. This could be due to the interaction method involving the use of eyes in the task.

After every session, the participants were given a chance to comment on the experiment and the eye-typing system. One of the comments, made by several participants and deemed useful for the improvement of eye-tracking communication systems, is *"to have different dwell time for the letters and the word suggestions, since a low dwell time of* 250 *ms would work for the keyboard letters but not for word suggestions, which require a bit more processing before selection"*.

The experiment and its results described in the paper will have serious implications when conducting research in the field of eye-tracking, combined with eye-interaction and cognitive psychology. Variables like blink frequency and blink duration are higher for more difficult tasks, due to the involvement of eyes in the task. Pupil dilation, although a strong indicator of cognitive load, might be not be a valid measure of cognitive load when the task takes longer and the tonic pupil dilation is being measured. Considering a task involving typing, the strongest indicator of cognitive load might still be performance measures like typing speed, error rate and the eye-typing specific measure of attended but not selected rate.

In summary, the results of this experiments indicate that eye data, in combination with performance measures, has immense potential for ubiquitous analysis to estimate users' cognitive load, in a range of possible applications. While data based purely on eye-metrics or performance alone is interesting [Appel et al. 2018], the combination of both, as described in this paper, would provide accurate as well as real-life applicable insights. Measurement of cognitive load for people using an eye-typing keyboard on a daily basis, to further investigate cognitive states like fatigue and attention, could help in better understanding their mental needs, for instance — adaptive keyboard layouts [Sengupta et al. 2017], assistance in word or phrase selection or requirements of breaks during cognitive overload to prevent frustration or fatigue. Cognitive load detection is a highly relevant aspect of other hands-free typing applications like augmented reality based training for skill development [Webel et al. 2013], to promote adaptive learning of the task. Extension of eye-typing to eye-tracking based key selection in virtual and augmented reality could find application in designing navigation interfaces and a deeper understanding of the user preferences for the same. Learning systems could vastly benefit from further research in similar areas as this paper, during implementation of online learning tools, by expanding on eye-based results during eye-typing to reading and understanding of the reading materials.

5 CONCLUSION

We have presented an experiment design and results on detection of cognitive load during an interactive eye-typing task, where the cognitive load was manipulated using a working memory task of memorizing sentences of varying complexity. The task difficulty was validated using subjective scores on perceived difficulty, which were significantly higher for the difficult trials. We have found significant differences between easy and difficult trials using performance measures and blink measures. Relative pupil size has shown correlation with the expected cognitive load of working memory, but the difference between easy and difficult trials was not found to be significant.

ACKNOWLEDGMENTS

This work is funded by the Bevica Fonden, Denmark and Technical University of Denmark.

REFERENCES

Tobias Appel, Christian Scharinger, Peter Gerjets, and Enkelejda Kasneci. 2018. Cross-subject workload classification using pupil-related measures. In *Proceedings - ETRA 2018: ACM Symposium on Eye Tracking Research & Applications.* https://doi.org/10.1145/3204493.3204531

Per Bækgaard, Michael Kai Petersen, and Jakob Eg Larsen. 2016. Assessing Levels of Attention Using Low Cost Eye Tracking. *International Conference on Universal Access in Human-Computer Interaction* 9737 (2016). https://doi.org/10.1007/978-3-319-40250-5

Douglas Bates, Martin Maechler, Ben Bolker, and Steven Walker. 2014. *lme4: Linear mixed-effects models using Eigen and S4.* R package version 1.1-7.

Andrew Begel and Hana Vrzakova. 2018. Eye Movements in Code Review. In *Proceedings - EMIP 2018: Eye Movements in Programming.* https://doi.org/10.1145/3216723.3216727

Pradipta Biswas, Varun Dutt, and Pat Langdon. 2016. Comparing Ocular Parameters for Cognitive Load Measurement in Eye-Gaze-Controlled Interfaces for Automotive and Desktop Computing Environments. *International Journal of Human-Computer Interaction* (2016). https://doi.org/10.1080/10447318.2015.1084112

C. H. Björnsson. 1968. Läsbarhet. *Lund: Liber* (1968).

Jonas Blattgerste, Patrick Renner, and Thies Pfeiffer. 2018. Advantages of eye-gaze over head-gaze-based selection in virtual and augmented reality under varying field of views. In *Proceedings - COGAIN 2018: Workshop on Communication by Gaze Interaction.* https://doi.org/10.1145/3206343.3206349

Lex Fridman, Bryan Reimer, Bruce Mehler, and William T. Freeman. 2018. Cognitive Load Estimation in the Wild. In *Proceedings - CHI 2018: CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3173574.3174226

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

Eija Haapalainen, SeungJun Kim, Jodi F. Forlizzi, and Anind K. Dey. 2010. Psycho-Physiological Measures for Assessing Cognitive Load. In *Proceedings - Ubicomp'12: 12th ACM international conference on Ubiquitous computing*. https://doi.org/10.1145/1864349.1864395

John Paulin Hansen, Vijay Rajanna, I. Scott MacKenzie, and Per Bækgaard. 2018. A Fitts' law study of click and dwell interaction by gaze, head and mouse with a head-mounted display. In *Proceedings - COGAIN 2018: Workshop on Communication by Gaze Interaction*. https://doi.org/10.1145/3206343.3206344

Weidong Huang, Seok-Hee Hong, and Peter Eades. 2006. Predicting graph reading performance: A cognitive approach. In *Proceedings - 2006 Asia-Pacific Symposium on Information Visualisation*. https://doi.org/10.1145/1151903.1151933

Byron Jaeger. 2017. *r2glmm: R Squared for Mixed (Multilevel) Models*. R package version 0.1.2.

Daniel Kahneman and Jackson Beatty. 1966. Pupil Diameter and Load on Memory. *Science* (1966). https://doi.org/10.1126/science.154.3756.1583

Shaun K. Kane and Meredith Ringel Morris. 2017. Let's Talk About X: Combining Image Recognition and Eye Gaze to Support Conversation for People with ALS. In *Proceedings - DIS '17 :2017 Conference on Designing Interactive Systems*. https://doi.org/10.1145/3064663.3064762

Ioanna Katidioti, Jelmer P. Borst, Douwe J. Bierens de Haan, Tamara Pepping, Marieke K. van Vugt, and Niels A. Taatgen. 2016. Interrupted by Your Pupil: An Interruption Management System Based on Pupil Dilation. *International Journal of Human-Computer Interaction* (2016). https://doi.org/10.1080/10447318.2016.1198525

Thomas Kosch, Mariam Hassib, Daniel Buschek, and Albrecht Schmidt. 2018. Look into my Eyes : Using Pupil Dilation to Estimate Mental Workload for Task Complexity Adaptation. *Proceedings - CHI 2018: CHI Conference on Human Factors in Computing Systems* (2018). https://doi.org/10.1145/3170427.3188643

Per Ola Kristensson and Keith Vertanen. 2012. The potential of dwell-free eye-typing for fast assistive gaze communication. *Proceedings - ETRA 2012: ACM Symposium on Eye Tracking Research & Applications* (2012). https://doi.org/10.1145/2168556.2168605

Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* (2017). https://doi.org/10.18637/jss.v082.i13

Harris R Lieberman, William J Tharion, and Barbara Shukitt-hale Karen L Speckman. 2002. Effects of caffeine , sleep loss , and stress on cognitive performance and mood during U . S . Navy SEAL training. *Psychopharmacology* (2002). https://doi.org/10.1007/s00213-002-1217-9

Monika Lohani, Brennan R. Payne, and David L. Strayer. 2019. A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving. *Frontiers in Human Neuroscience* (2019). https://doi.org/10.3389/fnhum.2019.00057

Xinyao Ma, Zhaolin Yao, Yijun Wang, Weihua Pei, and Hongda Chen. 2018. Combining Brain-Computer Interface and Eye Tracking for High-Speed Text Entry in Virtual Reality. *Proceedings - IUI 2018: Conference on Intelligent User Interfaces* (2018). https://doi.org/10.1145/3172944.3172988

Päivi Majaranta. 2009. *Text Entry by Eye Gaze*. Ph.D. Dissertation. https://doi.org/10.4018/978-1-61350-098-9.ch008

Päivi Majaranta. 2011. *Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies: Advances in Assistive Technologies*. IGI Global. https://doi.org/10.4018/978-1-61350-098-9.ch014

Sebastiaan Mathôt, Jasper Fabius, Elle Van Heusden, Stefan Van der Stigchel, Sebastiaan Mathot, Sebastiaan Mathôt, Jasper Fabius, Elle Van Heusden, and Stefan Van der Stigchel. 2018. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods* (2018). https://doi.org/10.3758/s13428-017-1007-2

Martez E Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. 2017. Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times. In *Proceedings - CHI 2017: CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3025453.3025517

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution* (2013). https://doi.org/10.1111/j.2041-210x.2012.00261.x

Alexandra Papoutsaki, Nediyana Daskalova, Patsorn Sangkloy, Jeff Huang, James Laskey, and James Hays. 2016. WebGazer: Scalable webcam eye tracking using user interactions. In *Proceedings - IJCAI International Joint Conference on Artificial Intelligence*. https://doi.org/10.1145/2702613.2702627

Jimin Pi and Bertram E. Shi. 2017. Probabilistic adjustment of dwell time for eye typing. In *Proceedings - HSI 2017: 10th International Conference on Human System Interactions*. https://doi.org/10.1109/HSI.2017.8005041

Ondrej Polacek, Adam J. Sporka, and Brandon Butler. 2013. Improving the methodology of text entry experiments. In *4th IEEE International Conference on Cognitive Infocommunications, CogInfoCom 2013*. https://doi.org/10.1109/CogInfoCom.2013.6719232

Ondrej Polacek, Adam J. Sporka, and Pavel Slavik. 2017. Text input for motor-impaired people. *Universal Access in the Information Society* (2017). https://doi.org/10.1007/s10209-015-0433-0

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Kari-Jouko Räihä and Saila Ovaska. 2012. An exploratory study of eye typing fundamentals. In *Proceedings - CHI 2012: CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/2207676.2208711

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* (2005). https://doi.org/10.1016/j.ergon.2005.04.005

Dario D. Salvucci. 2000. An Interactive Model–based Environment for Eye–movement Protocol Analysis and Visualization. *Proceedings - ETRA 2000: ACM Symposium on Eye Tracking Research & Applications* (2000). https://doi.org/10.1145/355017.355026

Korok Sengupta, Jun Sun, Raphael Menges, Chandan Kumar, and Steffen Staab. 2017. Analyzing the Impact of Cognitive Load in Evaluating Gaze-Based Typing. In *Proceedings - IEEE Symposium on Computer-Based Medical Systems*. https://doi.org/10.1109/CBMS.2017.134 arXiv:1706.02637

John Sweller. 2011. Cognitive Load Theory. *Psychology of Learning and Motivation - Advances in Research and Theory* (2011). https://doi.org/10.1016/B978-0-12-387691-1.00002-8 arXiv:arXiv:1011.1669v3

Yuu Tanaka and Kiyoshi Yamaoka. 1993. Blink activity and task difficulty. *Perceptual and motor skills* (1993). https://doi.org/10.2466/pms.1993.77.1.55

Outi Tuisku, Veikko Surakka, Ville Rantanen, Toni Vanhala, and Jukka Lekkala. 2013. Text Entry by Gazing and Smiling. *Advances in Human-Computer Interaction* (2013). https://doi.org/10.1155/2013/218084

Sabine Webel, Uli Bockholt, Timo Engelke, Nirit Gavish, Manuel Olbrich, and Carsten Preusche. 2013. An augmented reality training platform for assembly and maintenance skills. *Robotics and Autonomous Systems* (2013). https://doi.org/10.1016/j.robot.2012.09.013