Studying Eye Movements As A Basis For Measuring Cognitive Load

Johannes Zagermann
HCI Group
University of Konstanz
johannes.zagermann@uni.kn

Ulrike Pfeil
HCI Group
University of Konstanz
ulrike.pfeil@uni.kn

Harald Reiterer
HCI Group
University of Konstanz
harald.reiterer@uni.kn

CHI'18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada
ACM 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3188628

Abstract
Users' cognitive load while interacting with a system is a valuable metric for evaluations in HCI. We encourage the analysis of eye movements as an unobtrusive and widely available way to measure cognitive load. In this paper, we report initial findings from a user study with 26 participants working on three visual search tasks that represent different levels of difficulty. Also, we linearly increased the cognitive demand while solving the tasks. This allowed us to analyze the reaction of individual eye movements to different levels of task difficulty. Our results show how pupil dilation, blink rate, and the number of fixations and saccades per second individually react to changes in cognitive activity. We discuss how these measurements could be combined in future work to allow for a comprehensive investigation of cognitive load in interactive settings.

Author Keywords
cognitive load; eye tracking; evaluation

ACM Classification Keywords
H.5.2 [User Interfaces]: Evaluation/methodology

Introduction
Traditional measurements of usability (effectiveness, efficiency, and satisfaction) allow to assess the quality of system support for task performance [4]. Additionally, the

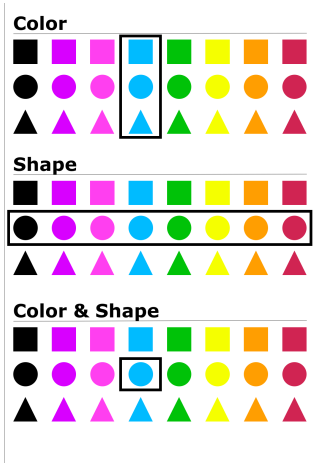

Figure 1: Target elements for each task are encircled – other elements are distractors. In **color**, participants had to find a randomly assigned blue element (disregarding different shapes); in **shape** they had to find a randomly assigned circle (disregarding different colors). Both tasks resemble a preattentive visual search task, yet without pop-out effect due to a high variety of distractors [9]. In **color & shape** they had to conduct a conjunction search by finding a blue circle [9].

cognitive load that systems and tasks place on users can serve as a valuable metric for evaluation. Cognitive load describes the amount of mental effort of the working memory when interacting with a system or solving a task [8]. It can be split into the load imposed by a task (*intrinsic*), the load to understand and process content (*germane*), and the way information is presented (*extraneous*) [8]. The later is valuable for practitioners, as a smooth system interaction can lower the extraneous load to free cognitive capacities.

Standardized measurements to assess the load imposed by a system are post-hoc questionnaires like the NASA TLX [3]. It relies on subjective ratings of participants, but does not provide in situ information or assess unconscious processes. To address these limitations, eye tracking is suggested to measure users' cognitive processes while interacting with a system in an objective and unobtrusive way [6]. To this end, several studies have investigated the relation of eye-movements and cognitive processes: Findings reveal a relation of increased cognitive load with an increment in pupil dilation [6], a higher frequency of fixations and saccades [2], as well as a lower blink rate [1]. Most of these studies either focused on individual eye movements or investigated these measurements in a very artificial context neglecting influences of real-world activities [10]. In our work, we aim to measure cognitive load based on multiple eye movements in interactive settings to detect cognitive changes during task activities. Based on our results, we discuss the potential of cross-validation, where weaknesses of one measurement might be compensated for by other measurements' strengths [5].

Research Goal

Our overall goal is to allow for the classification and thus evaluation of cognitively demanding activities using unobtrusive eye tracking devices. To that end, we need to better understand how different eye movements are affected by cognitive load. As an initial step, we tackle this by analyzing different eye movements individually while linearly increasing the cognitive demand on different levels of task difficulty. In future work, results from this analysis will be the basis of complementary cross-validation of the measurements towards assessing cognitive load.

User Study

This section provides an overview of the employed methods, including the tasks, apparatus, participants and the procedure that we followed in our user study.

Tasks and Apparatus

We employed a visual search task (similar to [7]) in three variants, representing different levels of difficulty: **color**, **shape**, and **color & shape** (see Figure 1). In line with [7], we decided for stimuli that involve perception as well as higher-level cognitive processes. Yet, they also represent a basic set of elements excluding the need for specific domain knowledge. For each task, participants were asked to find one specific target element in a field of distractors by clicking on it as soon as they found it using a mouse. We defined an invisible grid covering the entire screen space and set the number of elements (width and height: 20 pixels) to 155, representing 10% of the screen space in round 1. We linearly increased the number of distractors (and by this the difficulty) by 10% for each round until the display was completely covered (1550 elements) in round 10. In order to gather sufficient eye tracking data, each round had to be passed three times – each with different randomized positions of elements – resulting in 30 trials per task and participant. We decided to limit the duration of each visual search activity to 10 seconds to assure comparability of tasks. After this, participants were asked to place the cursor in the middle of the screen to avoid lucky target hits. Addi-

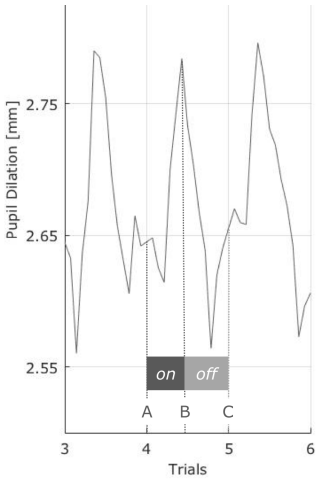

Figure 2: Exemplary pupil dilation during trials. ***A*** represents a trial's start, ***B*** correlates with the target hit, and ***C*** indicates a trial's end.

tionally, the content of the screen was frozen for a resting period of two seconds – avoiding possible influences on eye movements due to differences in e.g. screen brightness.

We used a Microsoft Perceptive Pixel 55″display with a resolution of 1920×1080 pixels as a wall display. We adjusted the height of the upper edge to 1.60m and positioned a small desk at 1.30m distance from the display at a comfortable height to control a mouse while standing (1.20m). To record eye movements, we employed SMI Eye Tracking Glasses 2, running at 120Hz. Interaction logs were matched with the eye tracking data for analysis.

Participants

34 participants were recruited for the user study. Due to insufficient eye tracking data, we excluded eight participants for analysis. The remaining 26 participants (16 female, 10 male) had a mean age of 24.15 (SD=2.49, aged 20-29). All participants were students from our local campus. They had a mixed background ranging from economics, politics, and computer science. None of the participants were color-blind and all of them had normal or corrected to normal eye-sight using contact lenses. Consequently, they had no problems with the employed size and color of items.

Procedure

At the beginning, each participant was asked to fill out a demographic questionnaire including questions e.g. about eye-sight. Then, each participant received an introduction into the system including the eye tracking glasses. Already wearing the eye tracking glasses, participants had to perform a demo task to get familiar with the procedure. This allowed the eye tracking software to adapt to individual eye features to improve the tracking accuracy of the subsequent 3-point calibration process. After that, participants were asked to start the first assigned task. Afterwards participants were asked to fill out a NASA TLX [3] questionnaire. This sequence was the same for all three tasks (**color**, **shape**, **color & shape**). To avoid learning effects, we counterbalanced the order of tasks by using a random selection of 34 out of 36 possible combinations. Finally, each participant filled out a post-questionnaire concerning ratings of cognitive demand. Each session lasted about an hour in total and participants were compensated for their time.

Analysis and Results

For statistical analyses, we chose dependent *t*-tests (if the assumption of normal distribution was met) or Wilcoxon signed-rank tests for comparisons of ***on*** and ***off*** phases (see Figure 2) within a task (e.g. **color**) and repeated-measures ANOVA for comparisons across tasks (e.g. **color** vs. **shape**).

Subjective Ratings

We used a NASA TLX questionnaire [3] to investigate participants' task load when working with each of the three tasks. A repeated-measures ANOVA revealed statistically significant differences for the overall NASA TLX score ($F(2, 50) = 83.82$, $p < .05$) with mean scores of 23.40 (**color**, SD=12.18), 32.44 (**shape**, SD=14.96), and 48.14 (**color & shape**, SD=14.63). A post-hoc analysis revealed a statistically significant higher task load for **color & shape** compared to **color** ($p < .016$) and to **shape** ($p < .016$). In addition, **shape** resulted in a statistically significantly higher task load compared to **color** ($p < .016$).

For the ratings of cognitive demand, we asked participants to rank the tasks (1 for low, 2 for medium, and 3 for high). The mean scores are 1.04 (**color**), 2.12 (**shape**), and 2.85 (**color & shape**). 25 participants rated **color** with the lowest demand, 21 rated **shape** with a medium demand, and 22 rated **color & shape** as task with the highest demand.

Based on the results of the NASA TLX and the subjective

	Color	Shape	Color & Shape	
R1	2.01	2.37	2.19	on
	2.24	2.22	2.09	off
R2	1.81	2.06	2.20	on
	1.99	1.91	2.10	off
R3	1.85	2.00	2.59	on
	2.02	1.87	1.76	off
R4	1.94	2.25	2.54	on
	1.98	1.74	1.74	off
R5	2.03	2.37	2.72	on
	2.15	1.81	1.28	off
R6	2.06	2.26	2.87	on
	2.04	1.74	1.27	off
R7	2.12	2.20	2.79	on
	1.98	1.74	1.39	off
R8	2.13	2.25	2.89	on
	2.04	1.57	0.96	off
R9	2.13	2.35	2.98	on
	1.90	1.70	1.04	off
R10	2.23	2.16	2.92	on
	1.95	1.79	0.70	off

Figure 3: Mean number of fixations per second. Each round is represented by values for ***on*** and ***off*** phases. Standard deviations are hidden to increase readability. Green background color represents statistically significant differences ($p < .05$).

ratings, we conclude that the tasks pose three distinct levels of cognitive demand on participants, with **color** being the easiest, followed by **shape**, and **color & shape**.

Pupil Dilation

Analyzing participants' pupil dilation during the trials, we identified a peak that correlates with the task activity (see detail in Figure 2). We consider the phase between a trial's start (Figure 2-A) and the target hit (Figure 2-B) as ***on*** (as the participant is actively searching for the target element) and the time span between the target hit (Figure 2-B) and the start of the next trial (Figure 2-C) as ***off***. We follow this approach for all other metrics – additionally, this approach (including constraints given by the task) keeps the number of elements and the luminance constant, which minimizes the influence on eye movements. As the ***on/off*** approach counterbalances values for pupil dilation, we do not consider statistical calculations for it.

Fixations

Figure 3 shows the comparisons of the number of fixations per second between ***on*** and ***off*** phases for all ten rounds in the three tasks. Focusing on statistically significant differences concerning the number of fixations per second across the three tasks, we identify following differences: In **color**, there are no statistically significant differences in the number of fixations per second between **on** and **off** phases. As **color** was identified as easiest task, we conclude that the number of fixations is not reliable to detect differences in cognitive load for easy tasks. However, in **shape** and **color & shape** the number of fixations per second is statistically significantly higher in ***on*** than in ***off*** phases, starting from R3 (**color & shape**) and R4 (**shape**) onward.

We identified an increase for the differences between ***on*** and ***off*** phases in **color & shape**: The number of fixations per second in ***on*** phases increases per round while it decreases for ***off*** phases. We conclude that more difficult tasks lead to larger differences in the number of fixations per second. However, this tendency is not visible in **color** or **shape**. Thus, to measure changes in cognitive load, tasks have to be sufficiently demanding, as small changes (e.g. in easy tasks) cannot be detected. In substantially demanding tasks, we might not only distinguish different states of cognitive load using the number of fixations per second, but also describe the extent of such differences.

To identify task-specific differences regarding cognitive load, we applied a repeated-measures ANOVA to compare the mean number of fixations per second in ***on*** phases for tasks **color** ($M = 2.03$, $SD=0.20$), **shape** ($M = 2.23$, $SD=0.55$), and **color & shape** ($M = 2.67$, $SD=0.41$). This difference was statistically significantly different ($F(2, 50) = 20.95$, $p < .05$). A post-hoc analysis revealed statistically significant more fixations per second in **color & shape** compared to **color** ($p < .016$) and compared to **shape** ($p < .016$).

Saccades

Figure 4 shows the mean number of saccades per second for ***on*** and ***off*** phases for all ten rounds in the three tasks. The analysis of the number of saccades per second reveals similar tendencies as the analysis of fixations, yet more sensitive with regard to changes in cognitive activity for easier tasks. Figure 4 shows that for **color**, the number of saccades is statistically significantly higher in ***on*** phases compared to ***off*** phases for rounds eight, nine, and ten. Thus, even for the easiest task (**color**), the number of saccades can be used to identify differences if the number of distractors is high enough. In tasks **shape** and **color & shape** the number of saccades is statistically significantly higher in ***on*** phases compared to ***off*** phases for all runs. Again, the larger difference between **on** and **off** phases in advanced rounds of **color & shape** indicate a possibility to measure

	Color	Shape	Color & Shape	
R1	1.76	2.25	1.98	on
	1.81	1.83	1.64	off
R2	1.61	1.93	2.05	on
	1.55	1.58	1.75	off
R3	1.68	1.83	2.46	on
	1.58	1.55	1.76	off
R4	1.78	2.13	2.39	on
	1.62	1.44	1.50	off
R5	1.81	2.23	2.56	on
	1.71	1.45	1.06	off
R6	1.82	2.11	2.77	on
	1.61	1.48	1.13	off
R7	1.93	2.05	2.68	on
	1.82	1.44	1.13	off
R8	1.92	2.11	2.78	on
	1.62	1.34	0.82	off
R9	1.94	2.20	2.85	on
	1.51	1.36	0.79	off
R10	2.03	2.03	2.80	on
	1.56	1.46	0.63	off

Figure 4: Mean number of saccades per second. Each round is represented by values for ***on*** and ***off*** phases. Standard deviations are hidden to increase readability. Green background color represents statistically significant differences ($p < .05$).

the extent of the in- or decrease of cognitive activities.

We applied a repeated-measures ANOVA to compare the mean number of saccades per second in ***on*** phases for **color** (M = 1.83, SD=0.26), **shape** (M = 2.09, SD=0.62), and **color & shape** (M = 2.53, SD=0.48). This difference was statistically significant ($F(2, 50) = 19.36$, $p < .05$). A post-hoc analysis revealed statistically significantly more saccades per second in **color & shape** compared to **color** ($p < .016$) and compared to **shape** ($p < .016$).

Blinks

Figure 5 shows the mean number of blinks per second for ***on*** and ***off*** phases for all ten rounds in the three tasks. We expected a lower blink rate being related to a higher cognitive load [2]. Our results are in line with previous work (cf. Figure 5). We observed statistically significantly less blinks per second for ***on*** phases compared to the corresponding ***off*** phases for **shape** and **color & shape**. This tendency is also visible in **color**, statistically significant differences could be identified for round 6 and round 8 onward.

We applied a repeated-measures ANOVA to compare the mean number of blinks per second in ***on*** phases for the tasks **color** (M = 0.14, SD=0.12), **shape** (M = 0.10, SD=0.10), and **color & shape** (M = 0.12, SD=0.10). This difference was statistically significantly different ($F(2, 50) = 6.31$, $p < .05$). However, a post-hoc analysis revealed no statistically significant differences after applying Bonferroni correction.

Discussion and Future Work

The goal of our study was to measure cognitive load in an interactive setting based on the analysis of multiple eye movements as a basis for future cross-validation of measurements. As our results show, all four measurements (pupil dilation, fixations, saccades, and blink rate) react to changes in cognitive activity, yet in slightly different ways.

Our analysis revealed the pupil to adapt very quickly to changes in cognitive demand, allowing us to differentiate cognitively active phases (***on***) from passive ones (***off***). This novel technique allows to measure cognitive load beyond tightly controlled conditions. However, due to the fluctuation, aggregating pupil dilation data over time was difficult. Also, the pupil is more sensitive to changes in light than to cognitive demand [6]. We accounted for that by keeping light conditions constant, however this is difficult to realize in real-world settings.

Regarding the number of fixations and saccades, we conclude that a higher number of these eye movements is related to an increased level of cognitive load. We found that both measurements are applicable to detect significant changes in cognitive demand. However, the measurement of saccades per second also detects changes on easier tasks, whereas the number of fixations reacts to changes in cognitive demand on a higher level. Yet, missing statistically significant differences might not necessarily indicate that an eye movement is less applicable, it might also reflect a lack of cognitive load. Additionally, we found a tendency, that the size of the difference between ***on*** and ***off*** phases relates to the level of cognitive load, as the differences seem to be larger for more difficult tasks. Envisioning the application of these measurements in real-world settings, we need to note that fixations and saccades are very dependant on the visualization, the interaction, and the nature of the task [10]. Relying on a single metric bears the danger to measure how people visually interact with the system, rather than their cognitive engagement. Investigating fixations and saccades for different types of tasks that involve an even stronger focus on interaction (e.g. sorting, moving, or rotating objects) in future work will help us to validate our findings and analyze to what extend these measurements can be used to investigate cognitive load task-independently.

	Color	Shape	Color & Shape	
R1	0.15	0.07	0.16	on
	0.38	0.37	0.40	off
R2	0.16	0.09	0.16	on
	0.38	0.38	0.38	off
R3	0.15	0.10	0.11	on
	0.40	0.33	0.39	off
R4	0.11	0.09	0.14	on
	0.39	0.30	0.46	off
R5	0.14	0.09	0.12	on
	0.48	0.30	0.30	off
R6	0.14	0.11	0.11	on
	0.44	0.51	0.37	off
R7	0.14	0.12	0.12	on
	0.43	0.29	0.35	off
R8	0.15	0.12	0.10	on
	0.43	0.30	0.21	off
R9	0.12	0.11	0.12	on
	0.44	0.30	0.24	off
R10	0.13	0.12	0.10	on
	0.43	0.39	0.17	off

Figure 5: Mean number of blinks per second. Each round is represented by values for ***on*** and ***off*** phases. Standard deviations are hidden to increase readability. Green background color represents statistically significant differences ($p < .05$).

Regarding the measurement of blink rates, our findings reveal significantly lower blink rates for ***on*** than for ***off*** phases for rather difficult tasks, showing their applicability to detect short term changes of cognitive load. As with fixations, this measurement does not seem to detect changes in cognitive load on a rather low level. Also, the analysis of blink rates did not reflect the different levels of difficulty between the three tasks. We conclude that – similarly to the analysis of the pupil dilation – the blink rate is particularly applicable for the identification of cognitive changes during task activity rather than for aggregations and comparisons of data over a longer period of time.

Discussing strengths and weaknesses, we conclude that eye measurements contribute to the detection of cognitive load. In the future, we encourage a cross-validation of eye movements to allow for a reliable measurement of cognitive load. We suggest to use the analysis of pupil dilation and blink rate to identify changes in cognitive demand during a task, while the analysis of the frequency of fixations and saccades seems promising to identify the extent of the cognitive load. Overall, our identified relations of eye tracking measurements to cognitive load encourage future work on cognitive load as an objective metric for evaluations in HCI.

Acknowledgements

We thank the German Research Foundation (DFG) for financial support within project C01 of SFB/Transregio 161.

REFERENCES

1. Siyuan Chen, Julien Epps, and Fang Chen. 2013. Automatic and Continuous User Task Analysis via Eye Activity. In *Proc. of IUI '13*. ACM, New York, NY, USA, 57–66.
2. Siyuan Chen, Julien Epps, Natalie Ruiz, and Fang Chen. 2011. Eye Activity As a Measure of Human Mental Effort in HCI. In *Proc. of IUI '11*. ACM, New York, NY, USA, 315–318.
3. Sandra G. Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Human mental workload* 1, 3 (1988), 139–183.
4. Kasper Hornbæk. 2006. Current Practice in Measuring Usability: Challenges to Usability Studies and Research. *Int. J. Hum.-Comput. Stud.* 64, 2 (Feb. 2006), 79–102.
5. Shamsi T. Iqbal, Piotr D. Adamczyk, Xianjun Sam Zheng, and Brian P. Bailey. 2005. Towards an Index of Opportunity: Understanding Changes in Mental Workload During Task Execution. In *Proc. of CHI '05*. ACM, New York, NY, USA, 311–320.
6. Bastian Pfleging, Drea K. Fekety, Albrecht Schmidt, and Andrew L. Kun. 2016. A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In *Proc. of CHI '16*. ACM, New York, NY, USA, 5776–5788.
7. Pernilla Qvarfordt, Jacob T. Biehl, Gene Golovchinsky, and Tony Dunningan. 2010. Understanding the Benefits of Gaze Enhanced Visual Search. In *Proc. of ETRA '10*. ACM, New York, NY, USA, 283–290.
8. John Sweller. 1988. Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science* 12, 2 (1988), 257 – 285.
9. Colin Ware. 2012. *Information visualization: perception for design*. Elsevier.
10. Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. 2016. Measuring Cognitive Load Using Eye Tracking Technology in Visual Computing. In *Proc. of BELIV '16*. ACM, New York, NY, USA, 78–85.