Measuring Cognitive Load for Map Tasks through Pupil Diameter

Peter Kiefer[1], Ioannis Giannopoulos[1],
Andrew Duchowski[2], and Martin Raubal[1]

[1] Institute of Cartography and Geoinformation, ETH Zürich
Stefano-Franscini-Platz 5, CH-8093 Zürich, Switzerland
{pekiefer,igiannopoulos,mraubal}@ethz.ch

[2] School of Computing, Clemson University
100 McAdams Hall, Clemson, S.C., USA
duchowski@clemson.edu

Abstract. In this paper we use pupil diameter as an indicator for measuring cognitive load for six different tasks on common web maps. Two eye tracking data sets were collected for different basemaps (37 participants and 1,328 trials in total). We found significant differences in mean pupil diameter between tasks, indicating low cognitive load for *free exploration*, medium cognitive load for *search*, *polygon comparison*, *line following*, and high cognitive load for *route planning* and *focused search*. Pupil diameter also changed over time within trials which can be interpreted as an increase in cognitive load for *search* and *focused search*, and a decrease for *line following*. Such results can be used for the adaptation of maps and geovisualizations based on their users' cognitive load.

1 Introduction

The cognitive load users must cope with has been identified as a major criterion for the design of geographic visualizations and geographic human-computer interfaces [5, 9, 7]. This has become even more relevant in the age of mobile computing where geographic information is presented on constrained interfaces and under stressful, distractive and multi-tasking conditions [11, 14].

The notion of cognitive load was introduced by Cognitive Load Theory (CLT) as a means of describing how the mental effort of learners is influenced by the design of learning material [30, 31, 23]. The geovisualization community has considered CLT mainly with respect to *extraneous cognitive load*, which denotes the cognitive load determined by the complexity of the information presentation, e.g., by the design of the map [5], or the number and type of animations [12]. It has been argued that high cognitive load may lead to less efficient and less effective map reading [20] and spatial orientation [28], as well as decreased spatial learning [21]. Recently, the cognitive load of experts and novices during a visual search task on a map was compared [25, 24] and interpreted as differences in *germane cognitive load*, which 'reflects the effort that contributes to the construction of schemas' [32] in permanent memory.

This paper takes a different view: for two map designs and one level of expertise, it investigates the *intrinsic cognitive load* [31] of six different tasks people typically perform on maps. We hypothesize that certain tasks (e.g., route planning) are more demanding for the working memory than others (e.g., comparing the area of polygons), thus inducing a higher cognitive load. This hypothesis is investigated through a user study on two different basemaps (Google Maps™ and OpenStreetMap; 1,328 trials in total, taken from 37 participants).

We use the pupil diameter while performing these tasks as a measure for cognitive load, which has repeatedly been shown to be a reliable indicator [15, 13, 19, 2]. Significant within-subject differences in the mean pupil diameter between tasks were found which we interpret as evidence for differences in the intrinsic cognitive load these tasks evoke. More precisely, we conclude on low cognitive load for the task *free exploration*, medium cognitive load for *search*, *polygon comparison*, *line following*, and high cognitive load for *route planning* and *focused search*. A further analysis reveals changes of pupil diameter over time within trials, suggesting an increase in cognitive load for *search* and *focused search*, and a decrease for *line following*.

Our paper is the first using pupillometry for the analysis of eye tracking data recorded during map interaction, thus aiming to contribute to *"fundamental empirical research and state-of-the-art evaluation methods within [...] geographic information visualization and cognition"* [8]. Further, since pupil diameter can be measured in real-time, our results have the potential to be used for adaptive maps [27] that change based on the user's current cognitive load.

We proceed as follows: section 2 provides background on cognitive load and how it can be measured with eye tracking. Section 3 introduces our method, including experimental design and stimulus selection for the eye tracking experiments. We report and discuss results in sections 4 and 5, before concluding the paper in section 6.

2 Related Work

2.1 Cognitive Load

Cognitive load was introduced in the 1980s as a theory of learning [30], targeted at an improvement of learning material. The theory suggests *'that total cognitive load is an amalgam of at least two quite separate factors: extraneous cognitive load which is artificial because it is imposed by instructional methods and intrinsic cognitive load over which instructors have no control'* [31, p.307].

Sweller identifies element interactivity as the main reason for intrinsic cognitive load [31] (we return to this in section 5). Extraneous cognitive load, on the other hand, is determined by the presentation of the material and can be influenced by the instructor. For instance, the design of a map can be either supportive or impedient for task solving [5]. The higher the intrinsic and/or the extraneous load, the less capacity remains in working memory for germane cognitive load – a third type of cognitive load which occurs during schema acquisition and automation [32, 26].

Bunch and Lloyd distinguish subjective from objective ways of measuring cognitive load. While the first are based on interviews or questionnaires, the latter can be achieved by measuring the performance of participants in a secondary (parallel) task [5]. What they omit are ways of measuring cognitive load using physiological sensors, such as eye tracking, galvanic skin response or electroencephalogram [10]. The first of these – eye tracking – has been applied to map tasks in recent work [25, 24]: experts and novices were found to have different average fixation durations and frequency which has been attributed to differences in germane cognitive load. In this paper, we focus on intrinsic cognitive load of different tasks and utilize a different eye tracking measure: the pupil diameter.

2.2 Cognitive Load and Pupil Diameter

We are not the first using eye tracking methodology in GIScience and Cartography. Much progress has been made in topics such as map interpretation, map interaction, spatial decision-making, and wayfinding (see [17], section 2 for a comprehensive overview). In this paper, instead of analyzing *where* on a stimulus someone is looking, we focus on pupil diameter - a novel approach in GIScience and Cartography.

It has long been recognized that a relationship exists between cognitive load and pupil diameter [4]. Although Hess and Polt [13] demonstrated correlation between pupil dilation and problem difficulty, i.e., pupil size increases with problem difficulty, their early study was limited by several factors, including state of the technology available at the time. Their study observations were based on camera recordings of five participants' eyes, with a 16-mm Arriflex camera taking image samples at 2 frames per second. Given multiplication problems of different complexity to solve, the pupils of each participant typically showed a gradual increase in diameter, reaching a maximum dimension immediately before a response was given, then reverting to the previous control size.

Pupil dilation, in response to a given assignment meant to elicit mental activity, is referred to as Task-Evoked Pupillary Dilation (TEPD) or Task-Evoked Pupillary Response (TEPR) [1, 3]. Using a television pupillometer sampling at 20 Hz, Ahern and Beatty [1] measured pupil diameter in a slightly updated replication of Hess and Polt's mental arithmetic experiment. In all correct responses to the assigned multiplication, pupillary responses showed a common pattern of dilation followed by a slight constriction after presentation of the multiplicand. A larger dilation was evoked by the multiplier; this increase in pupillary dilation was maintained during the problem-solving period. More difficult problems evoked larger pupillary dilations, reconfirming the relationship between problem difficulty and task-evoked activation.

Here, we test the *dilation reflex*, i.e., the relationship of pupil dilation to varying task demands, in the context of mentally processing geographic information. Instead of analog or digital cameras, we evaluate the utility of the pupil diameter as produced by a head-mounted eye tracker. Klingner et al. [18] review past uses of eye trackers for measuring TEPR. Confirming that an eye tracker can be used to measure cognitive load via measurement of pupil diameter, they suggest

measurement following a 2 sec delay after stimulus onset. While they advocate detailed timing and evaluation of short-term pupillary response, we adopt what Klingner et al. refer to as a coarse measurement of the time-aggregated style of data processing, i.e., an aggregated measurement of pupil diameter over a long period of time. Such coarse measurements have been successfully applied in previous studies, such as Hyönä et al.'s experiment on language tasks of different complexity [15, Experiment 1].

Marshall analyzes pupil diameter [22] suggesting that the dilation reflex undergoes oscillatory changes during different levels of cognitive load. They claim the measurement is reliable across hardware platforms and sampling rates [2]. Their approach relies on a sophisticated multiscale (wavelet) analysis of the pupil diameter frequency, e.g., effectively measuring pupillary hippus, or *pupil unrest* [29]. However, according to Beatty and Lucero-Wagoner [4], in addition to reflexive control of pupillary size, the tiny, cognitively related, fluctuations in pupillary diameter are visually insignificant and appear to serve no functional purpose whatsoever. Whether characterization of pupil unrest is a reliable measure of cognitive load appears debatable.

Here we intend to evaluate pupil diameter in state space instead of frequency space, a more straightforward and accessible method albeit potentially more susceptible to confounds stemming from the light reflex, or the pupil's response to light levels.

3 Method

Data collection was performed in two separate studies following the same design, setup, and procedure but differing in the basemap used: Google Maps™ for the first study (**GMaps**), OpenStreetMap for the second (**OSM**). Both studies took place in 2013. The **GMaps** dataset has previously been used for a paper on activity recognition [16].

We are not studying map design here, i.e., we will not compare cognitive load of **GMaps** vs. **OSM**. The rationale for using two datasets is rather to get an indication on whether results generalize over at least two map designs.

3.1 Experimental Design

The study followed a within-subject design with one independent variable (*task*) and one dependent variable (*mean pupil diameter*, measured in millimeters). Six test conditions were considered for *task* (see also [16]):

- **T1** *free exploration*: exploring the map at free will. ("You have 20 seconds for exploring the map. You can look at whatever you want.")
- **T2** *search*: searching for a point of interest ("On the following map, please search for X", where X is given by its label.)
- **T3** *route planning*: planning the shortest route between two cities ("Do you see X and Y? Please, plan the shortest route from X to Y.")

Fig. 1. Hardware setup for the two studies.

T4 *focused search*: searching for the 3 closest points of interest of a certain type on a 'you are here'-map ("Do you see your position (the blue dot)? Please, search for the three closest Z", where Z is an object type.)

T5 *line following*: counting intersections while following a road with one's gaze ("Do you see X? Please, follow X from North to South and count the number of intersections", where X is a road name and cardinal directions were systematically varied)

T6 *polygon comparion*: comparing the area of two lakes ("Do you see X and Y? Please compare the areas of these two lakes and name the bigger one.")

3.2 Participants

Participants for each of the two experiments were recruited through a university mailing list. All were university students or already holding a university degree. None of them used maps in their profession (i.e., no cartographer, geographer, land planner etc.); therefore they can all be regarded as having the same level of expertise. A monetary compensation of 15 CHF (Swiss Francs) was offered.

GMaps: 19 participants took part; 2 were excluded from further analyses due to calibration errors. From the remaining 17 participants, 10 were female. The average age was 28 years (SD: 8.7). **OSM**: 20 participants (11 female) took part and none was excluded. The average age was 23.8 years (SD: 7.4).

3.3 Apparatus

Data were recorded using the SMI (v1.8) head-mounted eye tracking glasses (30 Hz)[3] and transmitted via a USB cable to a laptop. A chin rest was placed at a distance of 65cm to the stimulus in order to guarantee that the viewer would look at the monitor along an axis perpendicular to the monitor plane. We used

[3] http://www.smivision.com/en.html

two 24" widescreen LED monitors (1920x1200 pixels, Samsung *S24A850DW*). One monitor was used to display the stimulus, the other one for controlling the experiment (see Figure 1). The experiment was controlled through our own software framework which chooses and presents a random set of stimuli, including instructions and previews, plus an (optional) re-calibration screen (refer to sections 3.4 and 3.5). Shutters and constant ceiling lights ensured the same lighting conditions in the room over all trials.

3.4 Procedure

Participants were introduced to the experiment. They were told they would have to solve simple tasks on maps. The eye tracker was mounted, and the participant was asked to rest her head on the chin rest. A three-point calibration was performed. Each participant had 36 trials on different stimuli (refer to section 3.5), presented in randomized order, where no two successive trials were from the same task. Each trial consisted of three phases:

1. *Instruction phase*: the participant was presented a textual description of the task (in German) and could ask questions.
2. *Preview phase*: either a preview showing small parts of the stimulus (**T3**, **T4**, **T5**, **T6**), or a black dot in the center (**T1**, **T2**) was shown. The goal of this phase was to clearly separate the task to be analyzed from an orientation activity beforehand. For instance, start and destination points for the route planning tasks were shown here. At the end of the preview phase the participant was asked to fixate a certain point in order to provide equal start conditions for all participants.
3. *Task phase*: the stimulus was shown, and the eye movements recording was started. The recording was either ended as soon as the participant indicated with a move of her hand that she had solved the task, or after a maximum of 20 seconds.

The experimenter checked the calibration after each trial. In case the calibration had been lost, the previous trial was considered 'not valid' (excluded from later analyses) and a re-calibration was performed.

3.5 Stimuli

Since we are not investigating map design here we chose stimuli from standard web maps as used by people in their daily routines[4]. Two different web maps were used as sources for the stimuli: Google Maps™[5] for the **GMaps** study, and OpenStreetMap[6] for the **OSM** study.

In order to ensure that participants see the exact same map extents, stimuli were static images (screenshots) without the possiblity of panning or zooming.

[4] Studies on standard web maps have become quite common recently, e.g. [6]
[5] Before the 2013 redesign (classic style); not available online any more (6 May 2016).
[6] http://www.openstreetmap.org/

(a) **T1**: *free exploration* (b) **T2**: *search*

(c) **T3**: *route planning* (d) **T4**: *focused search*

(e) **T5**: *line following* (f) **T6**: *polygon comparison*

Fig. 2. Example stimuli (**GMaps** dataset). Zoom levels: 12 for (a,c,f), 18 for (b,d,e).

Participants were supposed to be unfamiliar with the geographic area shown in the stimulus, but familiar with the language and cultural context to allow for reasonable search tasks. Since all participants were from Switzerland and native German speakers, we chose map extents from Germany and Austria. With a brief interview after the experiment we asserted they were indeed unfamiliar with the areas they had seen during the trials.

It is not possible to identify *the* representive instance of a certain task type, which implies that complexity within a task type generally varies (we return to this issue in section 5). Stimuli were chosen in a way that all task instances for one task type were of a similar difficulty level. More specifically, easy tasks were avoided to ensure a certain task duration which would allow us to collect a sufficient amount of data. Selection criteria are detailed in the following. One researcher selected stimuli following these criteria and discussed the selection with a second researcher.

In total, each participant was shown 36 out of 40 stimuli (see Figure 2 (a-f) for examples from the **GMaps** study). Each stimulus was used only for one task type and only shown once to a participant to avoid a learning effect:

- **T1** *free exploration*: 6 stimuli (3 urban, 3 rural). Criterion: similar density of point and line features across the whole stimulus.
- **T2** *search*: 9 stimuli (urban). Criteria: the stimulus must contain at least 30 labeled points. Instances with the type as the specific point of interest to look for must be present across the whole stimulus No large part of the map must be covered by empty polygons that would allow for limiting the search space, such as an ocean.
- **T3** *route planning*: 6 out of 8 stimuli (rural). Criteria: start and destination must be located at the edges of the stimulus. One stimulus for each pair of opposite cardinal directions (e.g., start in the North-East, destination in South-West). The highest road priority present between start and destination must allow for several (at least 5) possible route options of similar length (i.e., no clear short route on a highway or similar).
- **T4** *focused search*: 5 stimuli (urban). Criteria: as for **T2**. The distance between the third closest point to the 'you are here'-dot and the fourth closest should be similar. One stimulus with dot in the map center, and one for each of North/South/East/West.
- **T5** *line following*: 6 out of 8 stimuli (urban). Criteria: the road to follow must traverse the whole stimulus, starting and ending at opposite edges. One stimulus for each pair of opposite cardinal directions. There must be at least 10 intersections along the road.
- **T6** *polygon comparion*: 4 stimuli (rural). Criteria: the two lakes to compare must be located on opposite edges of the map. They should have similar size. One stimulus for each pair of opposite cardinal directions.

Stimulus selection criteria were the same for both studies (**GMaps** and **OSM**), therefore 80 stimuli were used in total.

The luminance was measured at the distance of the participant's eyes to the stimulus (accumulated local luminance) for each map stimulus. The results showed that the luminance was constant throughout the whole experiment, with a constant *lux* value of *270* (measured with testo 540, ISO 9001:2008). This ensures changes in pupil diameter are not caused by different color, hue, or contrast profile of the individual stimuli.

4 Results

As described in section 3.4, some trials were considered 'invalid' due to calibration issues. The number of valid trials (out of 1,404 recorded) used for the analysis was 1,328 (**T1**: 222; **T2**: 332; **T3**: 220; **T4**: 185; **T5**: 221; **T6**: 148). The average trial duration was 15.27 seconds (SD=5.55 seconds).

The eye tracker recorded for each gaze (at 30 Hz) the pupil diameter in millimeters which will be used as the basis for the following analyses.

4.1 Differences in Mean Pupil Diameter Between Tasks

The mean pupil diameter was calculated for every single task (aggregated trials) performed by each participant and was used as input for within-subjects analyses [15]. A Friedman test revealed that there were statistically significant differences between the measured mean pupil diameter for the six map tasks, $\chi^2(5) = 89.649$, $p < .001$. Post-hoc analyses with the Wilcoxon signed-rank test were performed, revealing statistically significant differences between several map tasks (see Table 1 (a)). Median (IQR) pupil diameters for tasks **T1** to **T6** were *2.53*, *2.63*, *2.68*, *2.66*, *2.64* and *2.61*, respectively. Minimum and maximum pupil diameters for tasks **T1** to **T6** were (*1.85*, *3.09*), (*1.96*, *3.47*), (*1.97*, *3.59*), (*2.03*, *3.65*), (*2.00*, *3.55*), (*1.89*, *3.44*) (all in millimeters).

Figure 3(a) illustrates an ordering between the tasks, based on the above results. An example is illustrated in Figure 3(b), showing the results obtained from a single user.

Analyses were also performed on the two different map services separately. A Friedman test revealed that there was a statistically significant difference between the measured mean pupil diameter for the six map tasks in each of the two map cases, GMaps and OSM, $\chi^2(5) = 46.681$, $p < .001$ and $\chi^2(5) = 63.629$, $p < .001$, respectively.

Post-hoc analyses with the Wilcoxon signed-rank test revealed statistically significant differences between several map tasks (see Table 1 (b) for **GMaps** and Table 1 (c) for **OSM**). Median (IQR) pupil diameters for task **T1** to **T6** for **GMaps** were *2.64*, *2.78*, *2.90*, *2.83*, *2.81* and *2.95*, respectively. Median (IQR) pupil diameters for task **T1** to **T6** for **OSM** were *2.36*, *2.49*, *2.54*, *2.53*, *2.48* and *2.45*, respectively.

4.2 Change in Pupil Diameter Within Trials

To evaluate the change in pupil diameter within each task, we follow to a certain extent Klingner et al. [18]. That is, Klingner et al. compute the change in pupil diameter (in mm), presumably with respect to a baseline signal. It is not clear, however, how large a temporal window was used over which the baseline was measured. They note that stimulus onset (spoken multiplicand) occurred 5 seconds after measurement began. From the data reported, it appears that the baseline measurement occurred over the first 2 seconds. Klingner et al. note that a smoothing filter was used to smooth the pupil diameter data.

We follow Klingner et al. [18] by computing our within-trial pupil change with respect to a baseline signal, captured over a variable-length temporal window (0.5, 1.0, 1.5, and 2.0 seconds). Prior to our computation, following Klingner et al., we also apply a Butterworth filter to smooth the raw pupil diameter data (see Fig. 5). We use a 2^{nd} degree Butterworth filter set to 1/30 half-cycles per sample (the point at which the gain drops to $1/\sqrt{2}$ of the passband). Smoothing of the pupil diameter effectively denoises the signal by removing the high frequency component, attributable to high frequency pupil diameter oscillation known as *pupil unrest* or *hippus* [29].

(a) All trials (both datasets combined).

	T1		**T2**		**T3**		**T4**		**T5**		**T6**	
	Z	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*
T1	-		*-5.288*	<.001	*-5.303*	<.001	*-5.137*	<.001	*-5.273*	<.001	*-4.956*	<.001
T2	-		-		*-3.432*	<.001	*-4.247*	<.001	-		-	
T3	-		-		-		-		-		-	
T4	-		-		-		-		-		-	
T5	-		-		*-2.663*	<.01	*-3.251*	<.001	-		-	
T6	-		-		*-2.663*	<.01	*-2.467*	<.05	-		-	

(b) **GMaps** dataset.

	T1		**T2**		**T3**		**T4**		**T5**		**T6**	
	Z	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*
T1	-		*-3.621*	<.001	*-3.621*	<.001	*-3.621*	<.001	*-3.621*	<.001	*-3.574*	*<.001*
T2	-		-		-		*-3.574*	<.001	-		*-2.817*	<.01
T3	-		-		-		-		-		-	
T4	-		-		-		-		-		-	
T5	-		-		-		*-2.627*	<.01	-		*-2.533*	<.05
T6	-		-		-		-		-		-	

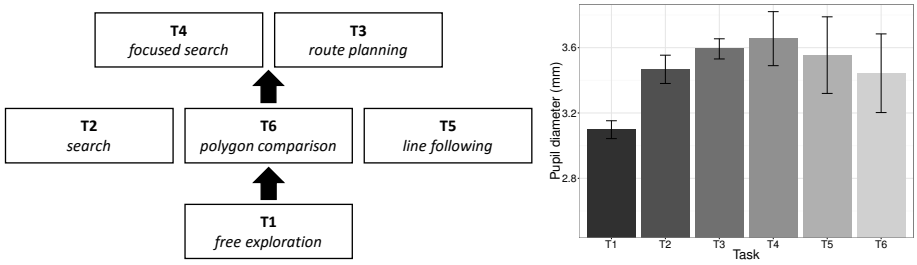
(c) **OSM** dataset.

	T1		**T2**		**T3**		**T4**		**T5**		**T6**	
	Z	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*	*Z*	*p*
T1	-		*-3.883*	<.001	*-3.920*	<.001	*-3.659*	<.001	*-3.845*	<.001	*-3.211*	<.001
T2	-		-		*-2.912*	<.005	*-2.539*	<.05	-		-	
T3	-		-		-		-		-		-	
T4	-		-		-		-		-		-	
T5	-		-		*-2.576*	<.05	*-2.240*	<.05	-		-	
T6	-		*-3.845*	<.001	*-3.920*	<.001	*-3.509*	<.001	*-4.247*	<.001	-	

Table 1. Differences in avg. pupil diameter within participants between tasks. Read the tables as follows: avg. pupil diameter for *task in line* is significantly smaller than for *task in column*.

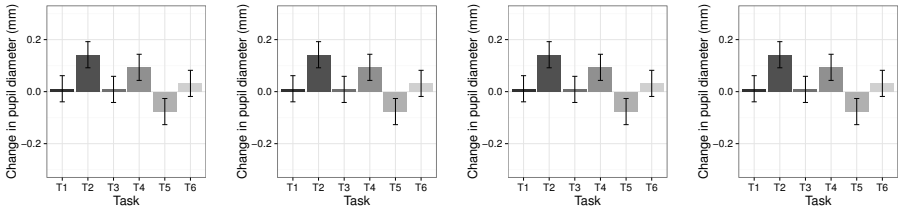
For each of the temporal windows, we used a univariate type-III repeated-measures ANOVA assuming a 2×6 mixed design where the independent variables were map type (between-subjects at two levels: **GMaps**, **OSM**) and task (within-subjects at 6 levels; see section 3.1). The dependent variable was mean pupil change computed as the mean of the pupil diameter difference from the mean diameter over the baseline time window, averaged over 20 seconds.

For a 0.5 second baseline (see Figure 4(a)), the effect of task was significant ($F(5, 175) = 14.64, p < 0.01$) but the map type was not ($F(1, 35) = 1.38, p = 0.25$,n.s.). The mean pupil difference was smallest during task **T5** (M $= -0.07$), and differed significantly from each of the tasks **T2**, **T4**, and **T6** ($p < 0.01$). Significant differences between mean pupil difference (at the $p < 0.01$ level) were also observed between tasks **T1** and **T2**, **T2** and **T3**, and **T2** and **T6**. Similar results were observed at larger baseline windows of 1.0–2.0 seconds (see Figures 4(b)– 4(d)).



(a) Ranking between tasks. (b) Example result for one user.

Fig. 3. Figure 3(a) ranks the tasks from significantly smaller (bottom) to significantly bigger (top) mean pupil diameter based on the results illustrated in Table 1(a). Figure 3(b) exemplifies the results obtained from a single user.



(a) $t_{\text{baseline}} = 0.5$ (b) $t_{\text{baseline}} = 1.0$ (c) $t_{\text{baseline}} = 1.5$ (d) $t_{\text{baseline}} = 2.0$

Fig. 4. Change in Pupil Diameter (CPD) with different baseline windows.

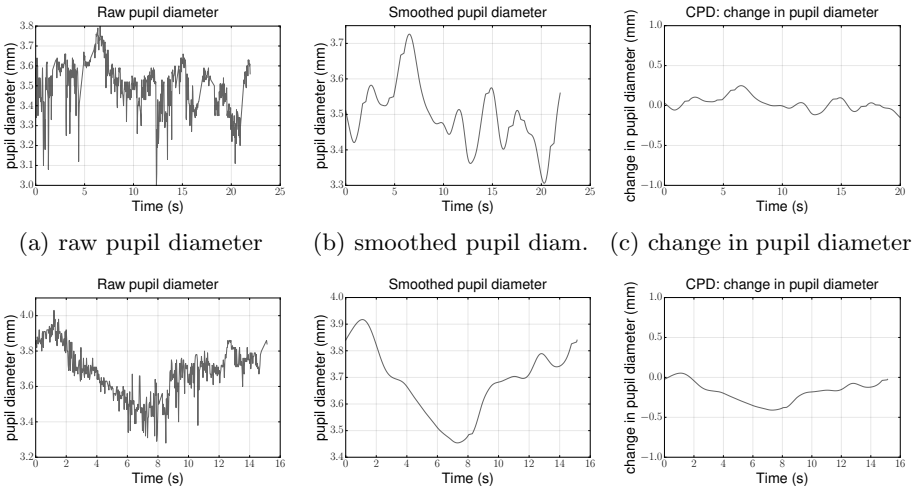
5 Discussion

Although pupil diameter is a well-known indicator for cognitive load [15, 13, 19, 2], it is also influenced by other factors, most importantly luminance (which was controlled for by the study setup) and fatigue. A potential effect of fatigue would apply to all tasks which were shown in a randomized order, therefore it is safe to assume the observed effect has been caused by differences in cognitive load.

Figure 3(a) summarizes our main results: we hypothesized differences in cognitive load between 6 tasks, and we indeed were able to group them into 3 classes of significantly different mean pupil diameter, suggesting differences in cognitive load. Setting up more detailed hypotheses from the beginning would have been speculation since, to our knowledge, no complete and heuristically proven cognitive model for these 6 map tasks exists yet.

Still, based on Sweller's idea of intrinsic cognitive load being influenced by the interactivity[7] of elements relevant for the task [31], our results make sense: it is no surprise that *free exploration* (**T1**) has the lowest cognitive load since nothing needs to be kept in working memory. *Polygon comparison* (**T6**), with

[7] A potential definition of element interactivity here would be the number of elements whose relation needs to be kept in working memory to solve a task successfully without having to keep the relations to or between any other elements in memory.



(a) raw pupil diameter (b) smoothed pupil diam. (c) change in pupil diameter

(d) raw pupil diameter (e) smoothed pupil diam. (f) change in pupil diameter

Fig. 5. Plots of representative pupil diameter and CPD ($T_{\text{baseline}} = 0.5$s) for user performing task **T1** (a,b,c) and **T5** (d,e,f) on Google Maps$^{\text{TM}}$.

medium cognitive load, can be solved by regarding the interaction of two map elements (the two lakes). During *line following* (**T5**), the participant at any moment needs to keep in working memory the road, the previous and current intersection, and a counter. *Search* (**T2**), another task with medium cognitive load, can be solved by keeping in memory all point objects that have been looked at already and their positions. *Focused search* (**T4**) is similar to **T2**, but with the additional requirement to estimate distances to the blue dot. Finally, a high cognitive load for *route planning* (**T3**) is reasonable since it requires a large number of map elements and their interaction to be considered.

The temporal within-trial analyses (section 4.2) added further insights: they indicate that the cognitive load of tasks **T1**, **T3**, and **T6** remained on the level it was at the start of the task. For instance, *free exploration* does neither have higher or lower cognitive load in later phases of the task than at the beginning (see Figure 5 (a,b,c) for an example). Cognitive load of the two search tasks (**T2**, **T4**) seems to increase, which is plausible since the number of visited points that needs to be kept in working memory increases as well. The decrease of cognitive load for **T5** (refer to Figure 5 (d,e,f)) is more difficult to interpret: peripheral vision might play a role here. The next intersection(s) relevant for the counting is/are most likely already perceived in the periphery in later phases of the task, which is not true when the stimulus 'pops up' at the start of the task.

Is it possible that we are observing changes in extraneous or germane, instead of intrinsic cognitive load [31, 26]? Germane cognitive load would occur if the participants learned schemata for the tasks. Our tasks are common, so it is unlikely participants created new schemata for, say, *route planning*. Ex-

traneous cognitive load would be an issue if the design of the basemap were specifically supportive or obstructive for some tasks. We approach this question by comparing the overall results (Table 1(a)) with the basemap-specific results (Tables 1(b,c)). The differences between **OSM** and the overall results are small; instead of being in the same 'medium cognitive load' class, **T6** in **OSM** causes significantly less cognitive load than **T2** and **T5** (which makes sense w.r.t. the number of interacting elements). In **GMaps**, on the other hand, there are larger differences: **T6** is now on the same (high) level as **T4**, while **T3** is only significantly higher than **T1**, but not than any other task. This result might be interpreted as Google Maps^{TM} being more supportive for *route planning* (**T3**), but less supportive for *polygon comparison* (**T6**) than OpenStreetMap.

As described in section 3.5, the stimuli were selected by two human raters following a set of criteria with the aim of identifying *'common'* cases for each task (neither too easy nor too difficult). The results are thus generalizable to tasks that are close to the introduced selection criteria, but probably do not apply to *all* potential instances of a task type, such as route planning with start and destination being directly connected by one street segment. Also, performing the same task on a different scale (e.g., route planning in a city) might lead to different results. Concerning generalizability over maps, the presented ranking (see Figure 3(a)) is based on the two tested popular map services. We do not claim that the presented results will hold independent of any map service.

Though we controlled for familiarity with the geographic areas, we did not control for familiarity with the map design. It can be assumed that participants were more familiar with **GMaps** than with **OSM**, potentially leading to lower cognitive load for **GMaps**. A comparison of cognitive load between map types, however, was not the aim of this study.

We did not include a short delay before task onset (unlike, e.g., Klingner et al. [18]). Instead, task onset began as soon as the stimulus appeared, and our analyses relied on either coarse (aggregated) pupil diameter (section 4.1, similar to [15]) or within-trial changes (section 4.2). Determination of the baseline temporal window for the latter is difficult. In our case it appears that cognitive demand begins fairly quickly. This gives credence to the use of a short temporal window. On the other hand, the longer the temporal baseline window, the less change in pupil diameter, on average, can be expected.

6 Conclusion

This paper is the first using pupil diameter as a measure for cognitive load while solving map tasks. We applied this measure to two datasets collected through studies on different web maps. We were able to group 6 map tasks into 3 classes of significantly different mean pupil diameter which we interpreted as differences in cognitive load: low (*free exploration*), medium (*search*, *polygon comparison*, *line following*), and high (*focused search*, *route planning*).

These results may motivate pupillometry to be used for future studies on cognitive load in GIScience research, such as during wayfinding [17]. It would

further be interesting to investigate the correlation between the number of interacting elements on a map and cognitive load more systematically (refer to section 5, [31]). Future gaze-contingent map interfaces may use our method to recognize cognitive load in real-time and adapt accordingly.

Acknowledgement

Supported by the Swiss National Science Foundation (grant no. 200021_162886).

References

1. Ahern, S., Beatty, J.: Pupillary Responses During Information Processing Vary with Scholastic Aptitude Test Scores. Science 205(4412), 1289–1292 (1979)
2. Bartels, M., Marshall, S.P.: Measuring Cognitive Workload Across Different Eye Tracking Hardware Platforms. In: ETRA '12: Proceedings of the 2008 Symposium on Eye Tracking Research & Applications. ACM, Santa Barbara, CA (2012)
3. Beatty, J.: Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. Psychological Bulletin 91(2), 276–292 (1982)
4. Beatty, J., Lucero-Wagoner, B.: The Pupillary System. In: Cacioppo, J.T., Tassinary, L.G., Bernston, G.G. (eds.) Handbook of Psychophysiology, pp. 142–162. Cambridge University Press, 2nd edn. (2000)
5. Bunch, R.L., Lloyd, R.E.: The cognitive load of geographic information. The Professional Geographer 58(2), 209–220 (2006)
6. Coltekin, A., Lokka, I.E., Boér, A.: The utilization of publicly available map types by non-experts - a choice experiment. In: Proceedings of the 27th International Cartographic Conference (ICC2015), Rio de Janeiro, Brazil. pp. 23–28 (2015)
7. Fabrikant, S.I., Goldsberry, K.: Thematic relevance and perceptual salience of dynamic geovisualization displays. In: Proceedings, 22th ICA/ACI International Cartographic Conference, Coruna (2005)
8. Fabrikant, S.I., Lobben, A.: Introduction: Cognitive issues in geographic information visualization. Cartographica: The International Journal for Geographic Information and Geovisualization 44(3), 139–143 (2009)
9. Giannopoulos, I., Kiefer, P., Raubal, M., Richter, K.F., Thrash, T.: Wayfinding decision situations: A conceptual model and evaluation. In: Proceedings of the Eighth International Conference on Geographic Information Science (GIScience 2014), pp. 221–234. Springer International Publishing (2014)
10. Haapalainen, E., Kim, S., Forlizzi, J.F., Dey, A.K.: Psycho-physiological measures for assessing cognitive load. In: Proceedings of the 12th ACM international conference on Ubiquitous computing. pp. 301–310. ACM (2010)
11. Harrison, R., Flood, D., Duce, D.: Usability of mobile applications: literature review and rationale for a new usability model. Journal of Interaction Science 1(1), 1–16 (2013)
12. Harrower, M.: The cognitive limits of animated maps. Cartographica: The International Journal for Geographic Information and Geovisualization 42(4), 349–357 (2007)
13. Hess, E.H., Polt, J.M.: Pupil Size in Relation to Mental Activity during Simple Problem-Solving. Science 143(3611), 1190–1192 (March 1964)

14. Hirtle, S.C., Raubal, M.: Many to many mobile maps. In: Raubal, M., Mark, D., Frank, A. (eds.) Cognitive and Linguistic Aspects of Geographic Space - New Perspectives on Geographic Information Research, pp. 141–157. Springer, Berlin, Heidelberg (2013)
15. Hyönä, J., Tommola, J., Alaja, A.M.: Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. The Quarterly Journal of Experimental Psychology 48(3), 598–612 (1995)
16. Kiefer, P., Giannopoulos, I., Raubal, M.: Using eye movements to recognize activities on cartographic maps. In: Proceedings of the 21st SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 488–491. ACM, New York, NY, USA (2013)
17. Kiefer, P., Giannopoulos, I., Raubal, M.: Where am I? Investigating map matching during self-localization with mobile eye tracking in an urban environment. Transactions in GIS 18(5), 660–686 (2014)
18. Klingner, J., Kumar, R., Hanrahan, P.: Measuring the task-evoked pupillary response with a remote eye tracker. In: Proceedings of the 2008 symposium on Eye tracking research & applications. pp. 69–72. ACM (2008)
19. Kruger, J.L., Hefer, E., Matthew, G.: 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. pp. 62–66. ACM (2013)
20. Lloyd, R.E., Bunch, R.L.: Explaining map-reading performance efficiency: gender, memory, and geographic information. Cartography and Geographic Information Science 35(3), 171–202 (2008)
21. Lloyd, R.E., Bunch, R.L.: Learning geographic information from a map and text: learning environment and individual differences. Cartographica: The International Journal for Geographic Information and Geovisualization 45(3), 169–184 (2010)
22. Marshall, S.P.: Method and Apparatus for Eye Tracking Monitoring Pupil Dilation to Evaluate Cognitive Activity. US Patent No. 6,090,051 (18 July 2000)
23. Mayer, R.E., Moreno, R.: Nine ways to reduce cognitive load in multimedia learning. Educational psychologist 38(1), 43–52 (2003)
24. Ooms, K., De Maeyer, P., Fack, V.: Study of the attentive behavior of novice and expert map users using eye tracking. Cartography and Geographic Information Science 41(1), 37–54 (2014)
25. Ooms, K., De Maeyer, P., Fack, V., Van Assche, E., Witlox, F.: Interpreting maps through the eyes of expert and novice users. International Journal of Geographical Information Science 26(10), 1773–1788 (2012)
26. Paas, F., Renkl, A., Sweller, J.: Cognitive load theory and instructional design: Recent developments. Educational psychologist 38(1), 1–4 (2003)
27. Reichenbacher, T.: Adaptive concepts for a mobile cartography. Journal of Geographical Sciences 11(1), 43–53 (2001)
28. Rossano, M.J., Moak, J.: Spatial representations acquired from computer models: Cognitive load, orientation specificity and the acquisition of survey knowledge. British Journal of Psychology 89(3), 481–497 (1998)
29. Stark, L., Campbell, F.W., Atwood, J.: Pupil Unrest: An Example of Noise in a Biological Servomechanism. Nature 182(4639), 857–858 (1958)
30. Sweller, J.: Cognitive load during problem solving: Effects on learning. Cognitive science 12(2), 257–285 (1988)
31. Sweller, J.: Cognitive load theory, learning difficulty, and instructional design. Learning and instruction 4(4), 295–312 (1994)
32. Sweller, J., Van Merrienboer, J.J., Paas, F.G.: Cognitive architecture and instructional design. Educational psychology review 10(3), 251–296 (1998)