Contents lists available at ScienceDirect

Applied Ergonomics

journal homepage: www.elsevier.com/locate/apergo

An analysis of the suitability of a low-cost eye tracker for assessing the cognitive load of drivers

Tomaž Čegovnik*, Kristina Stojmenova, Grega Jakus, Jaka Sodnik

Faculty of Electrical Engineering, University of Ljubljana, Trzaska 25, Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Eye tracking
Pupillometry
Eye Tribe
Driver
Driving simulator
Cognitive load

ABSTRACT

This paper presents a driving simulator study in which we investigated whether the Eye Tribe eye tracker (ET) is capable of assessing changes in the cognitive load of drivers through oculography and pupillometry. In the study, participants were asked to drive a simulated vehicle and simultaneously perform a set of secondary tasks with different cognitive complexity levels. We measured changes in eye properties, such as the pupil size, blink rate and fixation time. We also performed a measurement with a Detection Response Task (DRT) to validate the results and to prove a steady increase of cognitive load with increasing secondary task difficulty. The results showed that the ET precisely recognizes an increasing pupil diameter with increasing secondary task difficulty. In addition, the ET shows increasing blink rates, decreasing fixation time and narrowing of the attention field with increasing secondary task difficulty. The results were validated with the DRT method and the secondary task performance. We conclude that the Eye Tribe ET is a suitable device for assessing a driver's cognitive load.

1. Introduction

Operating a vehicle is a cognitively demanding and responsible task, where the driver's primary task is to focus on the road, the surrounding traffic, to obey traffic rules, and to safely operate a vehicle. Today's vehicles provide an increasing amount of other features such as advanced communication and infotainment systems, and other luxury facilities, which can divert the driver's focus and attention. There are also other non-driving related sources of distraction, such as simple mobile phone conversations or interactions with other passengers. Driver distraction has been shown to be one of the major causes of vehicle accidents (Klauer et al., 2014; Kahn et al., 2015). It is therefore reasonable to research a driver's cognitive load and the available methods for assessing it, in order to find solutions to decrease driver's distraction and, consequently, increase on-road safety.

Cognitive load has been assessed in various ways, mainly with the use of self-evaluating questionnaires, monitoring of psychophysiological activities, and the evaluation of driving and in-vehicle related task performance. Although the self-evaluation questionnaires and self-ratings may appear questionable, studies have demonstrated that people are capable of giving a numerical indication of their perceived load (Paas et al., 2003). Most subjective measures are multidimensional in that they assess groups of associated variables, such as mental effort, fatigue, and frustration, which are all highly correlated (Paas et al., 2003). A commonly used test isthe Rating Scale Mental Effort questionnaire (Zijlstra and Van Doorn, 1985) which is a scale for measuring perceived effort. Nasa Task Load questionnaire (NASA-TLX) (Hart and Staveland, 1988) on the other hand, focuses on perceived mental, physical and temporal demands, user's performance, effort and frustration. An adapted version of NASA TLX adjusted for driving for accessing specifically drivers' efforts – Driving Activity Load Index (DALI) (Pauzié, 2008). For perceived effort the Subjectively experienced effort scale (Eilers et al., 1986) has been used, and other one-dimensional scales (Likert 7 and 9 point scales), which mostly concentrate on cognitive load demands (Paas et al., 2003).

Cognitive load can be assessed indirectly also by measuring the performance of various secondary tasks (performed in addition to driving). The most typical example is the use of a Detection Response Task (DRT), which can be performed simultaneously with any kind of visual-manual or pure cognitive secondary task (ISO, 2016). The DRT is specified by ISO Standard 17488 as a standard for assessing the attentional effects of a driver's cognitive load. In this method, the driver's task is to respond to a stimulus by pressing a button attached to the driver's index finger. The stimuli can be visual or tactile. Visual stimulus can be presented remotely (remote DRT) or can be head-mounted to the driver's head, so it is always in the driver's visual field (head-mounted DRT). For visually demanding secondary tasks, tactile stimulus can be used, in the form of a vibration presented through a small tactor placed on the driver's collarbone. Changes in cognitive load can be assessed from the driver's response time – the longer the response time, the

[*] Corresponding author. Faculty of Electrical Engineering, University of Ljubljana, Trzaska 25, 1000 Ljubljana, Slovenia.
E-mail addresses: tomaz.cegovnik@fe.uni-lj.si (T. Čegovnik), kristina.stojmenova@fe.uni-lj.si (K. Stojmenova), grega.jakus@fe.uni-lj.si (G. Jakus), jaka.sodnik@fe.uni-lj.si (J. Sodnik).

http://dx.doi.org/10.1016/j.apergo.2017.10.011
Received 29 March 2017; Received in revised form 7 September 2017; Accepted 17 October 2017
Available online 31 October 2017

higher the cognitive load – and miss rate (non-detected stimuli for longer than 2500 ms).

On the other hand, changes in cognitive load can also be detected through different psychophysiological measures of drivers. The greatest advantage of these types of measurements is that they allow the continuous collection of data and do not require any kind of response from the driver (self-report or performing an additional task), leaving the driver's attentional and cognitive resources focused only on the observed tasks. Research studies report on monitoring ocular activities (Fakuda et al., 2005; Palinko et al., 2010a,b; Korbach et al., 2017), cardiac functions (Humphrey and Kramer, 1994; Mehler et al., 2011a; Ferreira et al., 2014; Mackersie and Calderon-Moultrie, 2016), electrodermal activity (Shi et al., 2007; Haapalainen et al., 2010; Ferreira et al., 2014) and electrical brain activity (Borghini et al., 2014), among others.

Research indicates that the pupil dilation is a very precise indicator of cognitive activity (Seeber, 2013; Chen et al., 2016, Marquart et al., 2015). Furthermore, the pupil dilation will increase when performing more difficult mental tasks compared to easier tasks. It indicates that the pupil response can be correlated to changes in cognitive load. Therefore, pupillometry has been used for assessing changes in cognitive load in various research fields, including driving and driver's cognitive load. Pupil dilation and other properties of the eye and eye movements can be measured with various eye tracking (ET) devices. The main obstacle to a broader use of ET devices used to be a relatively high cost as well as high complexity of use. However, with a significant drop in prices and availability of some low-cost and simple-to-use ET devices, pupillometry has become more accessible to everybody.

2. Related work

2.1. Observing pupillometry for driver's cognitive load assessment

In a study on assessing driver's cognitive load, Palinko et al. concluded that using a remote tracking device might be reliable for assessing driver's cognitive load, especially in simulators (Palinko et al., 2010a,b). Their results showed high correlation between pupillometric data and driving performance data. Significant differences in pupil sizes for increased cognitive load of drivers was found also in other studies, suggesting that the pupillary response could be an important indicator for estimating cognitive load of drivers (Heeman et al., 2013; Gable et al., 2013).

However, assessing driver's cognitive load by observing eye activity can be challenging, because of the driving environment's characteristics, both on the road and in simulated environments. One specific challenge lies in the changing lightning conditions and the trackers' mounting position. For example, the intensity of illumination of the experimental environment influences the results collected with remote eye-tracking devices. Furthermore, pupils are also very sensitive to light and their diameter decreases with increased luminosity (Palinko and Kun, 2011). Palinko et al., decoupled the effects of light and cognitive load and assessed their impact on the driver's pupil size. Their results showed that it is possible to dissociate those two effect on pupil diameter.

2.2. Using a low-cost eye-tracker for measuring eye activity

Measuring the exact size of the pupil can be difficult due to the varying distances between the eyes and position of the tracker. Changes of the pupil size, on the other hand, can be tracked easily and therefore they suffice to detect different levels of cognitive activity. Studies have shown that even low cost eye-tracking devices can provide reliable results. An evaluation of the low cost Eye Tribe's (THE EYE TRIBE, 2016) suitability and usability for scientific studies as a low-cost and affordable device, showed that it’s accuracy and precision seem to be very good under predefined optimal circumstances (Dalmaijer, 2014). He got this conclusions by comparing it to a professional eye-tracking device – the EyeLink 1000 – and validated its precision and accuracy, pupillometry and fixation. Another study showed that the accuracy and precision of the Eye Tribe is comparable to other similar devices (Ooms et al., 2015). As its accuracy and precision were compared with another well-established and comparable eye-tracker, the SMI RED 250. Eye Tribe's advantage against the SMI RED 250 is that it is easy to transport and set up very quickly, but it has to be used correctly because there are many factors that can affect the recordings.

2.3. Our research contribution

The aim of this study is to assess the suitability of the Eye Tribe – as a low cost ET – for detecting and measuring changes in the cognitive load of drivers. While the majority of the referenced studies focused primarily or solely on the pupillometry as the main indicator of changes in cognitive load, we decided to observe also other ocular activities that could also indicate these changes (i.e. blink rate, fixation time and eye movements). We systematically induced different levels of cognitive load by a delayed digit recall task and performed a reference measurement of changes in cognitive load with the standardized method – Detection Response Task.

The two unique contributions of this study are therefore:

1. The use of the low-cost ET for the holistic assessment of eye activities in order to detect changes in cognitive load
2. Direct comparison of the ET-based method for detecting changes in cognitive load and the standardized DRT method

3. Experiment design

Our study was designed as a within-subject (repeated measures) experiment carried out in a driving simulator. The participants were asked to follow the leading vehicle at a constant safety distance and simultaneously perform a secondary cognitively demanding task. We measured the pupil size, blink rate, eye movement and fixation rate with the ET and also performed a reference DRT measurement.

3.1. Driving simulator

The driving simulation was running on a high performance gaming computer, consisting of an i7-6700K CPU and GeForce GTX 980Ti graphics card. The visual output was a triple-screen combination of three Samsung 48″ curved TV displays using the Nvidia Surround technology, placed at an angle of 70°. The speakers of the central display were used for the audio output. The driver's seat was placed in front of the central screen at a driver's head distance of 140 cm. The steering wheel, consisting of a Fanatec wheel base V2 and a Fanatec Porsche 911 wheel, a -set of three Fanatec ClubSport pedals and an H-pattern Fanatec ClubSport shifter with locked reverse and 7th gear were mounted on the same platform as the driver's seat (see Fig. 1).

The software used was OKTAL SCANeR Driver Training Studio (provided by NERVteh) (NERVTEH, 2016), which provides a highly realistic driving environment including realistic car physics and appropriate feedback. A special scenario was created for this experiment, containing a standard two-lane two-way country road without intersections, surrounding objects and altitude changes. There was an autonomous leading car on the road periodically changing its speed, and random traffic on the opposite lane.

3.2. Tasks

3.2.1. Driving performance

The primary task was to drive safely and follow the leading car at a steady safety distance. The leading car was changing its speed every 20 s to a random velocity between 50 km/h and 90 km/h. The driving

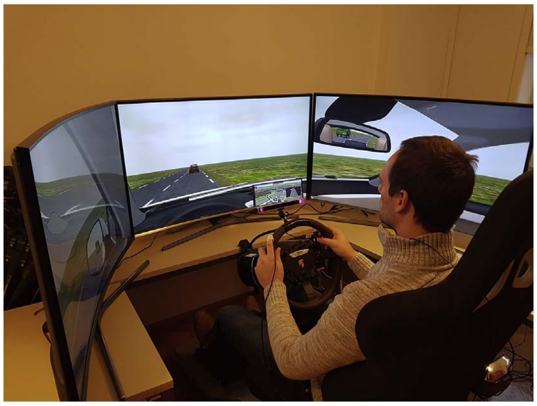

Fig. 1. Driving simulator set-up.

performance parameters were recorded throughout the experiment. These parameters are: distance to the followed vehicle, lane deviation and speed deviation.

3.2.2. DRT

In this study the tactile DRT was used in all phases of the experiment. The stimulus was presented through a vibrator attached to the participant's collar bone randomly every 2–5 s in order to avoid anticipation effect. Participants were asked to respond as quickly as possible by pressing a button attached to their index finger on the left hand against the wheel (see Fig. 2 and Fig. 3). The DRT was controlled via an Arduino board, which logged all the data to a file. All responses collected between 100 ms and 2500 ms were considered valid while other responses out of this time interval were considered as early or late. These were considered as missed responses. The ratio between correctly detected stimuli and all presented stimuli is called a hit rate. Response times and hit rates were taken as valid values for data processing and reference indicators of induced cognitive load (ISO/DIS 17488).

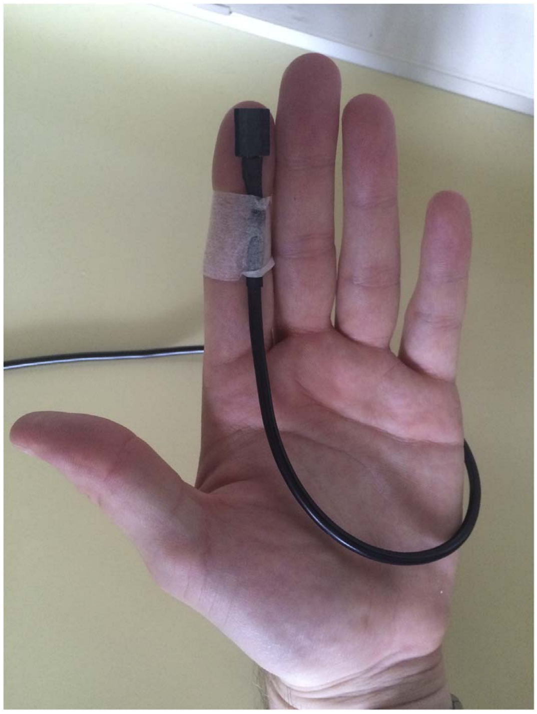

Fig. 2. Button attached to the index finger.

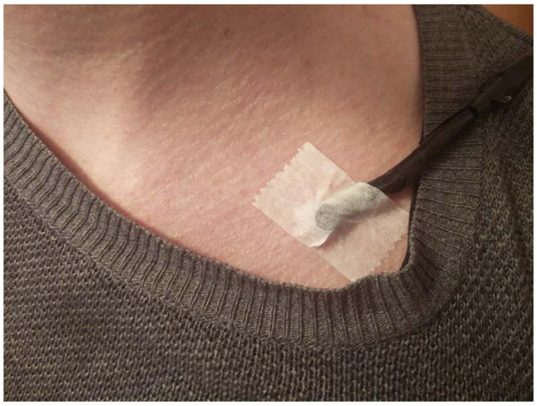

Fig. 3. Vibrator attached to the participant's collar bone.

3.2.3. Secondary *n*-back task

The secondary task was the Delayed Digit Recall *n*-back task, which is a modified form of the *n*-back task, also used in the draft ISO standard (Mehler et al., 2011a,b). The task requires the driver to repeat aloud a number dictated by the experimenter. In the 0-back test, the participant has to repeat the last dictated number. In the 1-back task, the participant has to recall and repeat the penultimate number, in the 2-back task, the number presented before the penultimate number, etc. In a similar pattern, the participant had to recall and repeat the number presented n time slots ago. This procedure forces the participant to store long sequences of numbers in a short-term memory resulting in highly increased cognitive load. Each number sequence was generated randomly for every separate phase of the session. All instructions and responses were given vocally. Sequences of numbers were generated with a Java program using the number generator in the Random class.

Our participants performed the secondary tasks at four difficulty levels (from 0-back to 3-back). Each level was timed to 2 min and every next number in the sequence was presented after a random period of time, approximately every 5–6 s. The *n*-back performance rate – the percentage of correct answers – was considered as the *n*-back performance indicator.

3.3. Eye tracker

The Eye Tribe (THE EYE TRIBE, 2016) is an inexpensive (99 USD) ET whose tracking principle is non-invasive, image-based eye tracking technique: pupil with corneal reflection. According to the specifications, the ET's sampling rate can be set to 30 Hz or 60 Hz, the accuracy is 0.5–1°, the operating range is 45–75 cm in front of the tracker. The tracking area of the device is 40 cm × 30 cm at a 65 cm distance (when sampling at 30 Hz).

The device basically functions as a USB 3.0 camera device that sends and receives data via USB 3.0 to a server running on a computer. The binocular gaze and pupil data can be accessed through the API using a variety of programming languages including Java, C+ + and C#.

The ET was placed on a flexible arm right over the steering wheel – in the driver's visual field under the central screen. The position was corrected for each participant so that the tracker-to-eye distance was always approximately 60 cm. The correct position and camera direction was set with the help of the EyeTribe UI software. We used the tracker's Java SDK and built a special logging software for this study so that we could synchronize the data of all devices for a given user and experimental phase. The tracker was connected to the computer running also the driving simulation software. Since the simulation software loaded primarily the GPU resources on the computer, no lag or variation of the ET's sampling rate was observed. The processing speed was not compromised at any time and – we observed that the processes running on the computer have never used more than 70% of the available computing resources.

Before starting the experiment, the tracker was calibrated with the Eye Tribe UI software. The tracker's calibration procedure requires the user to look at nine different points on the screen. This procedure generates the calibration matrix which is used for transforming the recorded eye data to the pixels of the gaze on screen.

We recorded data on the pupil size, blink rate, eye movement speed, fixation time and attention maps.

Steady illumination conditions were provided in the driving simulator room. The room was completely dimmed - there were no external light sources and no light sources in the room except the light emitted by the 3 TV screens used for the driving simulator. The average illuminance, measured at the driver's head position in the direction screen-to-eyes, was 30 lx. No measurable illumination was observed at the eye tracker position in the direction entering the eye tracker camera. The illuminance was measured with an illumination sensor on the smart phone Samsung Galaxy S7 Edge.

3.3.1. Pupil size

The ET measures the pupil size separately for the left and the right eye. The unit of pupil size data is not the exact metric diameter of the pupil, but the raw pupil size on the camera element without any corrections regarding the ET-to-eye distance. This is sufficient because we are interested in the relative pupil size between the normal size and the size while the subject is experiencing different levels of cognitive load. The pupil size is expected to increase with higher cognitive load.

3.3.2. Blink rate

Blink rate is the number of eye blinks that were observed in a time period. A blink rate of 0.2 s^{-1}, for example, means one blink every 5 s. The blink count was measured by counting the number of samples where the data contained pupil size values of zero, which was interpreted as the eye was closed.

3.3.3. Eye movement speed

Eye movement is defined as a sum of the Euclidean distances between consecutive eye gaze points on the screen, divided by the corresponding time period.

3.3.4. Fixation time

A fixation is an act of focusing the visual gaze on a specific location. A fixation is observed when the eye gaze is fixed at a specific point or radius for at least a given threshold time period. In our case, a fixation is detected when three consecutive eye gaze positions are recorded in a 30 px radius on screen. The relative fixation time is the ratio between the sum of all fixations within the selected time period and the elapsed time. Fixations were calculated using the Ogama tool (Ogama), which is open software for gaze analysis.

3.3.5. Attention map

The attention map is calculated as aggregated Gaussian distributions of all fixation points, which are weighted with their duration. Typically, this distribution is presented graphically as a color or a heat map overlaid on the original image of the screen.

3.4. Participants

22 participants (4 female), aged from 22 to 61 years, participated in the experiment. The average age was 32.9 years. All of them had a valid driving license and normal or corrected-to-normal vision.

4. Experiment procedure

Before the experiment, a short description of the purpose of the experiment was presented to each participant. At the beginning, participants filled out a survey about their demographic data, information about the driving license, driving experience and their medical condition related to sight. After the survey, participants took a 5-minute free ride in the driving simulator to get acquainted with the simulator.

Table 1
Experimental phases.

Phase number	N-back secondary task	Duration
1	No task (reference)	90 s
2	0-back	120 s
3	No task (recovery)	90 s
4	1-back	120 s
5	No task (recovery)	90 s
6	2-back	120 s
7	No task (recovery)	90 s
8	3-back	120 s
9	No task (recovery)	90 s

Next, the DRT equipment was attached and tested, and the Eye Tribe tracker was fixed into the appropriate position and calibrated.

The experiment was split into nine consequential phases. In all phases, participants had to respond to tactile DRT stimuli and follow the leading vehicle in a driving simulation. In phases 2, 4, 6 and 8 they additionally had to perform an n–back task with different complexity.

Phase 1 was the initial phase where participants had to just drive and perform the DRT. We consider this phase as a reference phase or reference measurement as no additional cognitive load was induced in this phase. The other odd phases are recovery phases to avoid delayed physiological responses of the eye. The phase order is presented in Table 1.

In each phase, ET data and DRT response time + miss rate was recorded. All participants were instructed to drive as safely as possible at all times and to consider driving as their high priority task. The driving simulation software logged all data related to the driving safety – safety distance to the vehicle being followed, speed, acceleration and lane deviation.

Each session was also video-taped to enable research and investigation of any potential anomalies in the recorded data. The average session time for each participant was approximately 45 min.

5. Results

5.1. Eye tracker data

5.1.1. Pupillometry – mean pupil size

Fig. 4 shows mean pupil sizes for all secondary task difficulty levels and the reference measurement. Considering the non-homogeneity of variance across test groups and the repeated measures experiment design, a Friedman test was used to detect differences across the test groups. The results of the Friedman test revealed a statistically significant difference ($p < 0.05$) in the mean pupil size depending on the level of difficulty, $\chi^2(4) = 36.615, p < 0.001$. A post-hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a corrected significance level of $p = 0.01$. The Wilcoxon test revealed significant differences among the majority of test groups. Individual comparisons are shown in Table 2.

To investigate if the task difficulty represents a significant factor in predicting the pupil size, we used a linear mixed model. Task difficulty was considered a fixed factor, and subjects were considered random factors. For the purpose of the linear mixed model analysis, pupil size data were log-transformed to meet the assumption of normality. The results of the analysis are shown in Table 3. The analysis confirmed our hypothesis that the pupil size is affected by difficulty ($p < 0.05$).

5.1.2. Blink rate

Fig. 5 shows blink rates for all task difficulties. As can be seen, the ratio increases with task difficulty indicating more eye blinking when

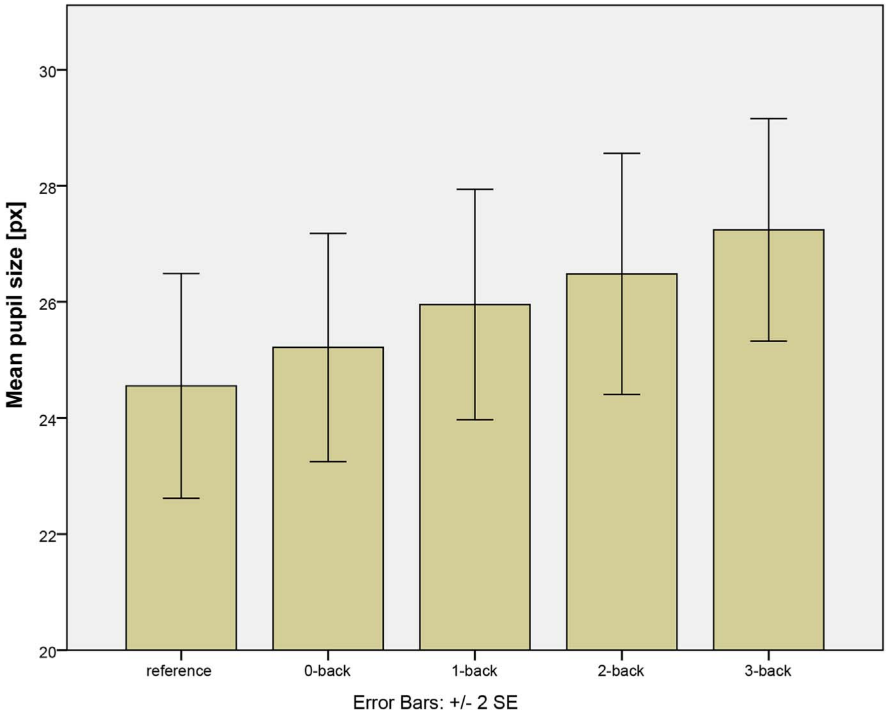

Fig. 4. Mean pupil size for both eyes and all levels of difficulty, and the reference measurements.

subjects performed increasingly cognitively demanding tasks.

The results of the Friedman test confirmed a statistically significant difference ($p < 0.05$) in the mean blink rate depending on the level of difficulty, $\chi^2(4) = 41.846$, $p < 0.001$. A post-hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a corrected significance level of $p = 0.01$. As shown in Table 4, the Wilcoxon test revealed significant differences among all test groups with the exceptions of the comparisons of the 0-back to the reference group and 2-back to 1-back group.

To confirm that the task difficulty represents a significant factor in predicting the pupil size, we also used a linear fixed model. Task difficulty was considered a fixed factor, and subjects were considered random factors. The analysis confirmed our hypothesis that the blink rate is affected by difficulty ($p < 0.05$) (see Table 5).

5.1.3. *Eye movement speed*

Mean eye movements speeds are shown in Fig. 6. The results of the Friedman test revealed a statistically significant difference ($p < 0.05$) in the mean eye movement speed depending on the level of difficulty, $\chi^2(4) = 10.031$, $p = 0.040$. However, a post-hoc analysis with Wilcoxon signed-rank tests with a Bonferroni correction applied (significance level set at $p = 0.01$) indicated no significant differences between different levels of task difficulties, including comparisons to the reference group.

5.1.4. *Fixation time*

Fig. 7 shows the mean relative fixation time for all levels of difficulty. The results of the Friedman test revealed a statistically significant difference ($p < 0.05$) in the relative fixation time depending on the level of difficulty, $\chi^2(4) = 21.477, p < 0.001$. A post-hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a corrected significance level of $p = 0.01$. As shown in Table 6, the Wilcoxon test revealed significant differences when comparing 1-back, 2-back and 3-back groups to the reference group, and 3-back to the 0-back group. In this case, the analysis using a linear mixed model did not show a linear relation between the fixation time and task difficulty.

Table 3
Estimates of fixed effects for mean pupil size.

Parameter	Estimate	Std. Error	df	t	p
Intercept	1.431921	0.016359	13.884	87.531	0.000
Reference vs. 3-back	-0.043662	0.004845	21.246	−9.011	0.000
0-back vs. 3-back	-0.034277	0.004830	17.733	−7.097	0.000
1-back vs. 3-back	-0.021654	0.004852	9.025	−4.463	0.002
2-back vs. 3-back	-0.013107	0.005093	14.350	−2.574	0.022

−2 Log Likelihood = −326.805; AIC = −302.805.

5.1.5. *Attention maps*

Fig. 8 shows one example of attention maps for one participant through all phases. Areas with a red overlay are areas with high gaze density (i.e. high focus), while areas with no coloring are areas with no attention and focus at all (OGAMA).

However, the analysis of attention maps is rather subjective and only provides an illustration of someone's visual attention span. As we can see in Fig. 8, the visual attention of our test subject is focused primarily on the road directly in front of the car and on the leading vehicle (red overlay). We can see some attention was also paid to the back mirror (violet overlay). The attention maps of phases with the *n*-back tasks (2nd, 4th, 6th and 8th from top to bottom) are noticeably smaller than the attention maps in other phases.

Table 2
The results of the Wilcoxon signed-rank tests for all levels of difficulty of the n-back task.

	0-back vs. ref.	1-back vs. ref.	2-back vs. ref.	3-back vs. ref.	1- vs. 0-back	2- vs. 0-back	3- vs. 0-back	2- vs. 1-back	3- vs. 1-back	3- vs. 2-back
Z	−2.411	−2.970	−3.180	−3.180	−2.132	−2.691	−3.180	−1.642	−2.970	−1.852
p	0.016	0.003	0.001	0.001	0.033	0.007	0.001	0.101	0.003	0.064

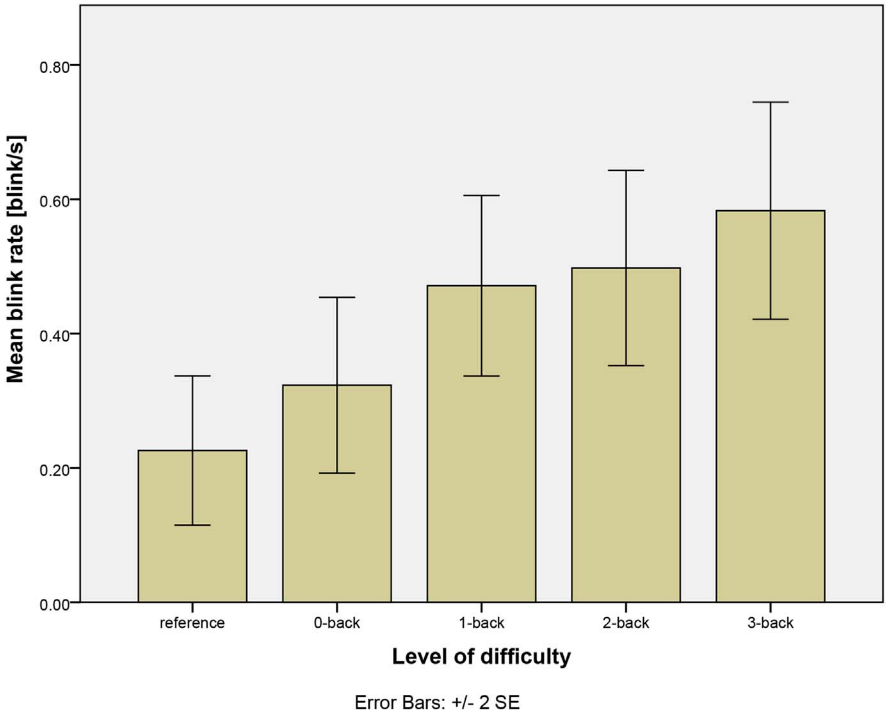

Fig. 5. Mean blink rate for all levels of difficulty, and the reference measurements.

5.2. *Detection Response Task*

5.2.1. *Response times*

Fig. 9 shows the mean response times of the participants to the tactile stimuli for all levels of the secondary *n*-back task and the reference measurement.

The results of the Friedman test revealed a statistically significant difference ($p < 0.05$) in the mean response time depending on the level of difficulty, $\chi^2(4) = 33.400, p < 0.001$. A post-hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p = 0.01$. As shown in Table 7, the Wilcoxon test revealed significant differences when comparing groups to the reference and 0-back groups; the comparison of the 2-back and 0-back groups showed a difference ($p = 0.015$) slightly above the corrected significance level. No significant differences were found when comparing the 1-, 2- and 3-back tests to each other ($p > 0.01$).

5.2.2. *Hit rate*

Fig. 10 shows the mean hit rate of the tactile DRT for all secondary task difficulty levels and the reference measurements. The results of the Friedman test revealed a statistically significant difference ($p < 0.05$) in the mean DRT hit rate depending on the level of difficulty, $\chi^2(4) = 28.940, p < 0.001$. A post-hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance level set at $p = 0.01$. In general, the Wilcoxon test revealed significant differences when comparing groups to the reference and 0-back groups (Table 8). The exceptions are comparisons of the 0-back and 2-back groups to the reference. In the first case, the difference in the DRT fail rate is not significant ($p > 0.01$), and in the second case ($p = 0.021$) it is relatively close to the corrected significance level. No significant differences were found when comparing the 1-, 2- and 3-back back test groups to each other.

Table 5
Estimates of fixed effects for blink rate.

Parameter	Estimate	Std. error	df	t	p
Intercept	0.583014	0.068873	15.870	8.465	0.000
Reference vs. 3-back	-0.357002	0.054496	17.703	−6.551	0.000
0-back vs. 3-back	-0.259840	0.041369	21.307	−6.281	0.000
1-back vs. 3-back	-0.111732	0.024038	13.319	−4.648	0.000
2-back vs. 3-back	-0.085346	0.026736	8.595	−3.192	0.012

−2 Log Likelihood = −89.374; AIC = −65.374.

5.2.3. *N-back task performance rate*

Fig. 11 shows the *n*-back test performance rate for all secondary task difficulty levels and the reference measurement.

The results of the Friedman test revealed a statistically significant difference ($p < 0.05$) in the mean *n*-back performance rate depending on the level of difficulty, $\chi^2(3) = 27.059$, $p < 0.001$. A post-hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied, resulting in a corrected significance level of $p = 0.0125$. As shown in Table 9, the Wilcoxon test revealed significant differences when comparing the 3-back to all other groups. When comparing the 1-back group and the 2-back group to the 0-back group, the significance of the differences in the mean *n*-back performance rates showed to be relatively close to a corrected significance level ($p = 0.027, p = 0.018$). No significant differences in the mean *n*-back task performance rates were found between the 1-back and 2-back groups.

Table 4
The results of the Wilcoxon signed-rank tests for all levels of difficulty of the n-back task.

	0-back vs. ref.	1-back vs. ref.	2-back vs. ref.	3-back vs. ref.	1- vs. 0-back	2- vs. 0-back	3- vs. 0-back	2- vs. 1-back	3- vs. 1-back	3- vs. 2-back
Z	−1.992	−3.180	−3.180	−3.180	−3.180	−2.900	−3.180	−1.503	−2.830	−2.970
p	0.046	0.001	0.001	0.001	0.001	0.004	0.001	0.133	0.005	0.003

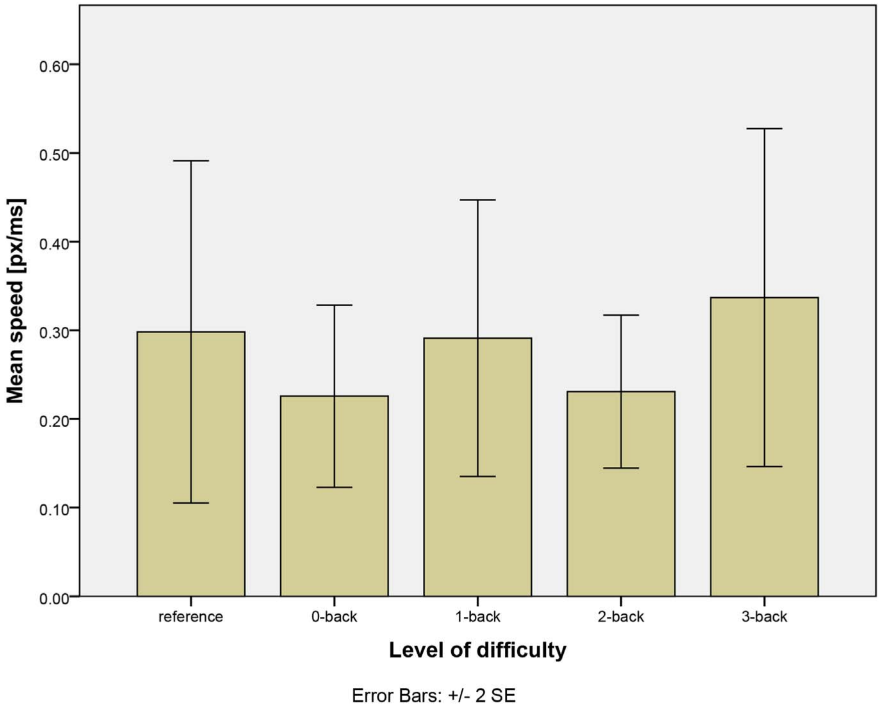

Fig. 6. Mean eye movement speed.

5.3. *Driving performance*

The evaluation of the driving performance parameters, which are: distance to the followed vehicle, lane deviation and speed deviation, revealed no statistically significant differences comparing all nine phases. This indicates that the participants performed the driving task with a high priority and drove safely and at a steady safety distance during the whole time of the experiment.

6. Discussion and conclusion

In this study we evaluated the suitability of a low-cost ET for detecting and measuring changes in the cognitive load of drivers by measuring various properties of the eye-pupil size, blink rate, fixation time and eye movements. In addition, we also performed the DRT measurement as a standardized reference method for assessing cognitive load in order to increase the validity of our results.

The measured pupil size shows a clear correlation with the *n*-back task difficulty by indicating the increase in the pupil diameter with the increase of the cognitive load level. The latter is clearly in line with

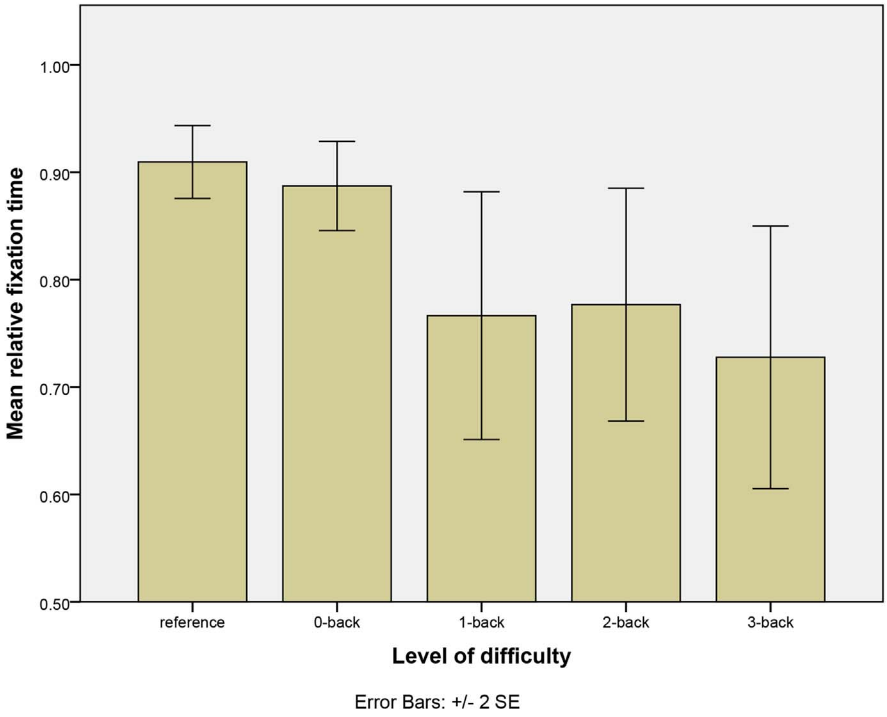

Fig. 7. Mean relative fixation time for all levels of difficulty.

Table 6
The results of the Wilcoxon signed-rank tests for all levels of difficulty of the n-back task.

	0-back vs. ref.	1-back vs. ref.	2-back vs. ref.	3-back vs. ref.	1- vs. 0-back	2- vs. 0-back	3- vs. 0-back	2- vs. 1-back	3- vs. 1-back	3- vs. 2-back
Z	-0.594	−3.110	−2.691	−3.110	−2.341	−2.132	−2.621	-0.804	−1.503	−1.922
p	0.552	0.002	0.007	0.002	0.019	0.033	0.009	0.422	0.133	0.055

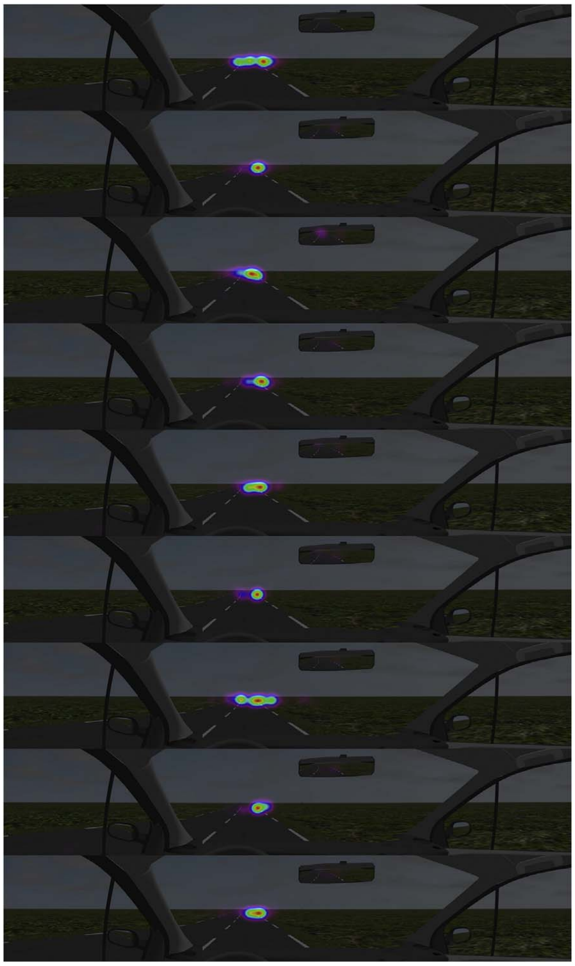

Fig. 8. Attention maps for all phases of one participant (top – 1st phase, bottom – 9th phase).

several previously reported studies (Chen and Epps, 2014). The blink rate also significantly increases with increasing *n*-back task difficulty. Similar findings were also reported in a previous study of Recarte et al. (2008). Both eye properties could therefore be used as clear indicators of increased mental activity.

The results of the measured eye movement speed, on the other hand, did not show any significant correlation with changes in the *n*-back task difficulty. These results might be affected by a rather low sampling rate of the Eye Tribe (e.g. it is limited to a max. of 60 Hz). We assume that ETs with a higher sampling rate could record eye movements more accurately and reveal some significant changes. Other studies have also reported on such issues (Ooms et al., 2015).

The mean fixation time clearly decreases in the phases with *n*-back tasks, however, these changes were not found to be significant. Therefore, we can only assume that a decrease of the fixation frequency can reveal the presence of increased cognitive load but cannot differentiate between different intensities of cognitive load.

The attention maps show different sizes of gaze distributions among different levels of task difficulty. There are clear differences between the cognitively demanding phases (i.e. we can clearly see the horizontal gaze narrowing) and the reference/recovery phases. This correlates with the results of Reimer (2009), who also found significantly smaller gaze distribution during the cognitively demanding tasks. We can confirm that in cognitively demanding tasks the so-called tunnel vision occurs.

DRT response times, hit rates and secondary *n*-back task performance rate are results of the reference measurement, which also clearly indicates an increase in cognitive load with increasing *n*-back task difficulty. However, the actual estimation of cognitive load is rather complex and should consider all three indicators simultaneously. During low to moderate cognitive load (0-back and 1-back), response times increase systematically while hit rates and *n*-back task performance rate decrease accordingly. During high and very high cognitive load (2-back and 3-back), response times remain at the approximate same level while hit rates and *n*-back performance rate further decrease. For example, there is a significant decrease in the *n*-back task performance rate at the 3-back phase.

These results indicate the presence of cognitive overload when, due to the limited cognitive resources, participants were forced to make priorities between different secondary tasks. Based on the results of the *n*-back task fail rate, it seems that the secondary tasks (1-back, 2-back and 3-back) were prioritized to the DRT stimuli.

We included the DRT measurement primarily because it is a standardized reference measurement for the assessment of cognitive load. However, the requirement to combine three different results is an important drawback of this method. Additionally, the method itself is a secondary task that requires user's action and certain amount of attention. As a consequence, it induces additional cognitive load that cannot be distinguished from the cognitive load induced by other observed secondary tasks. Therefore, psychophysiological measurements such as pupillometry and gaze activity represent much simpler and even more accurate methods for assessing cognitive load. These methods proved to be able also to detect and distinguish different levels of load rapidly and accurately. However, we would like to stress out particularly the importance of observing the pupilometry and gaze together. Relying solely on pupilometry might be tricky since the pupil size reacts also to changes in lightning conditions, stress and different substances.

The Eye Tribe proves to be a suitable device for measuring cognitive load in driving simulator experiments. The study of Ooms et al. (2015), on the other hand, says that the main pitfall is the need to use the tracker 'correctly', which means correct set-up, calibration and data processing. The device requires only physical calibration – the ET needs to be pointed at the eyes and the eyes need to be at the center of the frame. This way the measured pupil size does not change when the eyes and the head move. However, there is no need to run the advanced software calibration procedure, which defines the transformation matrix between pupil position and gaze direction. The latter is required only to measure the exact gaze position. Finally, data processing with the Eye Tribe is also fairly easy since the manufacturer provides an SDK for Java, C+ + and C# and enables easy integration and post-processing of ET data. We did not encounter any problems in measuring the pupil size variation of participants wearing prescription glasses. A frame of typical glasses does not block or hinder the eye tracker's field of view and does not affect the measurement of pupil size.

We conclude that the Eye Tribe tracker is very suitable to estimate the current level of cognitive load by measuring the pupil size and the blink rate. Both parameters change systematically with different levels

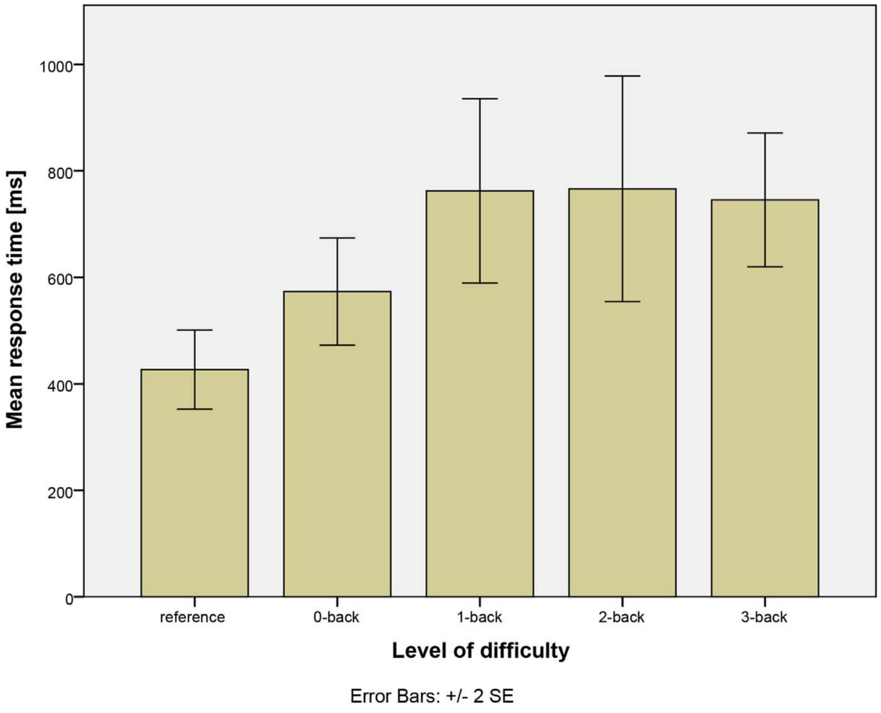

Fig. 9. Mean response times to DRT stimuli for all levels of difficulty of the n-back task.

Table 7
The results of the Wilcoxon signed-rank tests for all levels of difficulty of the n-back task.

	0-back vs. ref.	1-back vs. ref.	2-back vs. ref.	3-back vs. ref.	1- vs. 0-back	2- vs. 0-back	3- vs. 0-back	3- vs. 1-back	2- vs. 1-back	3- vs. 2-back
Z	−3.059	−3.059	−3.059	−3.059	−3.059	−2.432	−2.667	-0.235	-0.784	-0.157
p	0.002	0.002	0.002	0.002	0.002	0.015	0.008	0.814	0.433	0.875

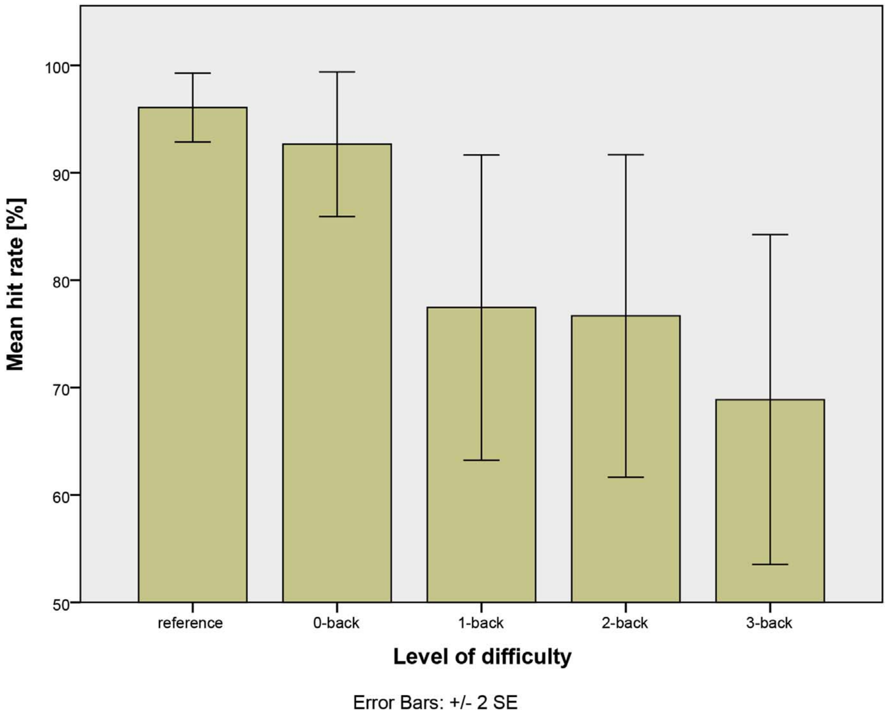

Fig. 10. Mean hit rate of DRT for all levels of difficulty and the reference measurements.

of cognitive load. Information on the current level of a driver's cognitive load could have many applications and advantages. It can, for example, be taken into consideration in designing and improving human-machine interfaces in vehicles that induce low cognitive load and, most importantly, avoid cognitive overload. Secondly, real-time information on cognitive load could be used as part of driving assistance systems to warn the driver about a possible cognitive overload and therefore avoid accidents. It could also be used by context-adaptive interfaces, which

Table 8
The results of the Wilcoxon signed-rank tests for all levels of difficulty of the n-back task.

	0-back vs. ref.	1-back vs. ref.	2-back vs. ref.	3-back vs. ref.	1- vs. 0-back	2- vs. 0-back	3- vs. 0-back	2- vs. 1-back	3- vs. 1-back	3- vs. 2-back
Z	-0.700	−2.805	−2.310	−2.803	−2.805	−2.666	−2.666	-0.459	−1.784	−2.134
p	0.484	0.005	0.021	0.005	0.005	0.008	0.008	0.646	0.074	0.033

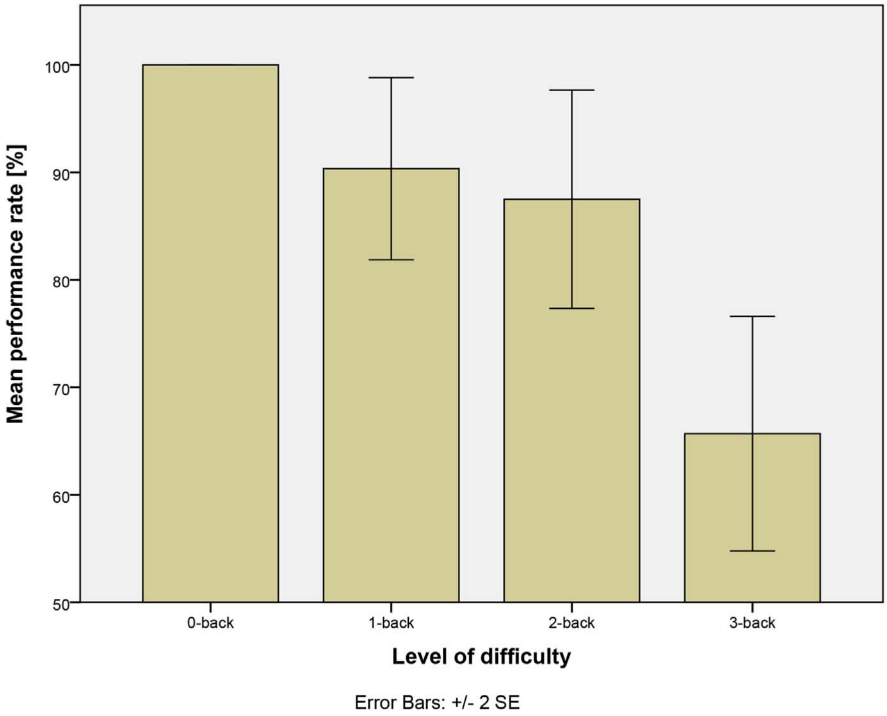

Fig. 11. Mean performance rate of the n-back test for all levels of difficulty and the reference measurements.

Table 9
The results of the Wilcoxon signed-rank tests for all levels of difficulty of the n-back task.

	1-back vs. 0-back	2-back vs. 0-back	3-back vs. 0-back	2-back vs. 1-back	3-back vs. 1-back	3-back vs. 2-back
Z	−2.214	−2.375	−3.068	-0.140	−2.983	−3.062
p	0.027	0.018	0.002	0.889	0.003	0.002

could present or hide information depending on the current level of a driver's cognitive load.

Acknowledgement

This work has been supported by the Slovenian Research Agency, and by NERVteh who provided the driving simulator software. The authors would like to thank Mr. Primož Žnidar for his help with the organization and execution of the study.

References

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., Babiloni, F., 2014. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. Neurosci. Biobehav. Rev. 44, 58–75. http://dx.doi.org/10.1016/j.neubiorev.2012.10.003.

Chen, S., Epps, J., 2014. Using task-induced pupil diameter and blink rate to infer cognitive load. Hum. Comput. Interact. 29, 390–413. http://dx.doi.org/10.1080/07370024.2014.892428.

Chen, F., Zhou, J., Wang, Y., Yu, K., Arshad, S.Z., Khawaji, A., Conway, D., 2016. Robust Multimodal Cognitive Load Measurement. Springer.

Dalmaijer, E., 2014. Is the Low-cost Eyetribe Eye Tracker Any Good for Research? (Rapport technique). PeerJ PrePrints.

Eilers, K., Nachreiner, F., Hänecke, K., 1986. Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Anstrengung. Zeitschrift für Arbeitswissenschaft 4, 214–224.

Fakuda, K., Stern, J.A., Brown, T.B., Russo, M.B., 2005. Cognition, blinks, eye-movements, and pupillary movements during performance of a running memory task. Aviat. Space Environ. Med. 76, 75–85.

Ferreira, E., Ferreira, D., Kim, S., Siirtola, P., Röning, J., Forlizzi, J.F., Dey, A.K., 2014, December. Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. In: Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), 2014 IEEE Symposium on. IEEE, pp. 39–48.

Gable, T.M., Walker, B.N., Henry, A.G., 2013. Cognitive workload, pupillary response, and driving: custom applications to gather pupillary data. Automot. User Interfaces Interact. Veh. Appl. 37.

Haapalainen, E., Kim, S., Forlizzi, J.F., Dey, A.K., 2010, September. Psycho-physiological measures for assessing cognitive load. In: Proceedings of the 12th ACM International Conference on Ubiquitous Computing. ACM, pp. 301–310.

Heeman, P.A., Meshorer, T., Kun, A.L., Palinko, O., Medenica, Z., 2013, October. Estimating cognitive load using pupil diameter during a spoken dialogue task. In: Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. ACM, pp. 242–245.

Humphrey, D.G., Kramer, A.F., 1994. Toward a psychophysiological assessment of dynamic changes in mental workload. Hum. Factors 36, 3–26.

ISO. (2016). Road vehicles – Transport information and control systems – Detection-Response Task (DRT) for assessing attentional effects of cognitive load in driving. ISO 17488:2016.

Kahn, C.A., Cisneros, V., Lotfipour, S., Imani, G., Chakravarthy, B., 2015. Distracted driving, a major preventable cause of motor vehicle Collisions:"Just Hang Up and drive". West. J. Emerg. Med. 16 (7), 1033.

Klauer, S.G., Guo, F., Simons-Morton, B.G., Ouimet, M.C., Lee, S.E., Dingus, T.A., 2014. Distracted driving and risk of road crashes among novice and experienced drivers. N. Engl. J. Med. 370 (1), 54–59.

Korbach, A., Brünken, R., Park, B., 2017. Differentiating different types of cognitive load: a comparison of different measures. Educ. Psychol. Rev. 1–27.

Mackersie, C.L., Calderon-Moultrie, N., 2016. Autonomic nervous system reactivity during speech repetition tasks: heart rate variability and skin conductance. Ear Hear. 37, 118S–125S.

Marquart, G., Cabrall, C., de Winter, J., 2015. Review of eye-related measures of drivers' mental workload. Procedia Manuf. 3, 2854–2861. http://dx.doi.org/10.1016/j.promfg.2015.07.783.

Mehler, B., Reimer, B., Wang, Y., 2011a. A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload. In: Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design, pp. 590–597.

Mehler, B., Reimer, B., Dusek, J.A., 2011b. MIT AgeLab Delayed Digit Recall Task (n-back). Massachusetts Institute of Technology, Cambridge, MA.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. Adv. Psychol. 52, 139–183.

NERVTEH, 2016. Simulation Technologies. http://nerv-teh.com/, Accessed date: 17 November 2016.

OGAMA (OpenGazeAndMouseAnalyzer): An open source software designed to analyze eye and mouse movements in slideshow study designs (http://www.ogama.net/, http://www.ogama.net/sites/default/files/pdf/OGAMA-BRM2008.pdf).

Ooms, K.G.U., Dupont, L.G.U., Lapon, L.G.U., Popelka, S.P.U.O., 2015. Accuracy and precision of fixation locations recorded with the low-cost Eye Tribe tracker in different experimental set-ups. http://dx.doi.org/10.16910/jemr.8.1.5.

Paas, F., Tuovinen, J.E., Tabbers, H., Van Gerven, P.W., 2003. Cognitive load measurement as a means to advance cognitive load theory. Educ. Psychol. 38 (1), 63–71.

Palinko, O., Kun, A.L., 2011. Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. In: Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design. Public Policy Center, University of Iowa, Iowa City, pp. 329–336.

Palinko, O., Kun, A.L., Shyrokov, A., Heeman, P., 2010a. Estimating cognitive load using remote eye tracking in a driving simulator. In: Proceedings of the 2010 Symposium on Eye-tracking Research & Applications, pp. 141–144.

Palinko, O., Kun, A.L., Shyrokov, A., Heeman, P., 2010b. Estimating cognitive load using remote eye tracking in a driving simulator. In: Proceedings of the 2010 Symposium on Eye-tracking Research & Applications. ACM, pp. 141–144.

Pauzié, A., 2008. Evaluating driver mental workload using the driving activity load index (DALI). In: Procedings European Conference on Human Centred Design for Intelligent Transport Systems, pp. 67–77.

Recarte, M.A., Pérez, E., Conchillo, A., Nunes, L.M., 2008. Mental workload and visual impairment: differences between pupil, blink, and subjective rating. Span. J. Psychol. 11, 374–385.

Reimer, B., 2009. Impact of cognitive task complexity on drivers' visual tunneling. Transp. Res. Rec. J. Transp. Res. Board 2138, 13–19. http://dx.doi.org/10.3141/2138-03.

Seeber, K.G., 2013. Cognitive load in simultaneous interpreting: measures and methods. Target. Int. J. Transl. Stud. 25 (1), 18–32.

Shi, Y., Ruiz, N., Taib, R., Choi, E., Chen, F., 2007. Galvanic Skin Response (GSR) as an Index of Cognitive Load. ACM Press, pp. 2651. http://dx.doi.org/10.1145/1240866.1241057.

THE EYE TRIBE, 2016. http://theeyetribe.com/ (accessed 13.10.2016).

Zijlstra, F.R.H., Van Doorn, L., 1985. Construction of a Scale to Measure Perceived Effort. Department of Philosophy and Social Sciences, Delft University of Technology, Delft, Netherlands (1985).