Behav Res
DOI 10.3758/s13428-017-0954-y

Comparing eye trackers by correlating their eye-metric data

Johannes Titz[1] · **Agnes Scholz**[2] · **Peter Sedlmeier**[1]

Abstract Up to now, the potential of eye tracking in science as well as in everyday life has not been fully realized because of the high acquisition cost of trackers. Recently, manufacturers have introduced low-cost devices, preparing the way for wider use of this underutilized technology. As soon as scientists show independently of the manufacturers that low-cost devices are accurate enough for application and research, the real advent of eye trackers will have arrived. To facilitate this development, we propose a simple approach for comparing two eye trackers by adopting a method that psychologists have been practicing in diagnostics for decades: correlating constructs to show reliability and validity. In a laboratory study, we ran the newer, low-cost EyeTribe eye tracker and an established SensoMotoric Instruments eye tracker at the same time, positioning one above the other. This design allowed us to directly correlate the eye-tracking metrics of the two devices over time. The experiment was embedded in a research project on memory where 26 participants viewed pictures or words and had to make cognitive judgments afterwards. The outputs of both trackers, that is, the pupil size and point of regard, were highly correlated, as estimated in a mixed effects model. Furthermore, calibration quality explained a substantial amount of individual differences for gaze, but not pupil size. Since data quality is not compromised, we conclude that low-cost eye trackers, in many cases, may be reliable alternatives to established devices.

Keywords Eye tracking · Pupil size · Low cost

Mass production of originally expensive technologies can revolutionize society. Imagine you are living in the 19th century and suddenly find that hitherto expensive books are now affordable to everyone. In Germany, this happened in 1867 when Reclam published a softcover edition of Goethe's *Faust* for only two silver groats (Johann & Junker, 1970, p. 239). This was the first book in their "Universal-Bibliothek" series, which rapidly became a popular education source and remains so today. Every student owns at least a couple of these books that are part of the literary canon.

Mass-production technologies can also have a profound effect on the scientific world. Between 1830 and 1850, sciences such as modern cell biology, cellular pathology, and normal histology sprouted up under the advent of the microscope. European countries led this development, with one plausible reason being that cheap but well-made devices were available in Germany, France, and Austria (Bradbury, 1967, p. 204).

We believe the time is ripe for another technology to fall into this category of sudden affordability and wide use in society and science: eye tracking. Over the last 15 years, the number of journal articles with the keyword "eye tracking" has increased exponentially (Fig. 1). Furthermore, the method has many practical applications outside science: In virtual reality, eye tracking allows "foveated rendering", which reduces the graphics processing unit load and power consumption (e.g., Guenter, Finch, Drucker, Tan, & Snyder, 2012; Pai et al., 2016). Eye tracking can be used in cars to test if the driver is dozing off (e.g., Nguyen, Chew, & Demidenko, 2015; Scholz, Franke, Platten, & Attig, in press;

✉ Johannes Titz
johannes.titz@gmail.com

[1] Department of Psychology, Research Methods and Evaluation in Psychology, Chemnitz University of Technology, Wilhelm-Raabestr. 43, 09111 Chemnitz, Germany

[2] Department of Psychology, Cognitive Decision Psychology, University of Zurich, Zurich, Switzerland

Published online: 06 September 2017

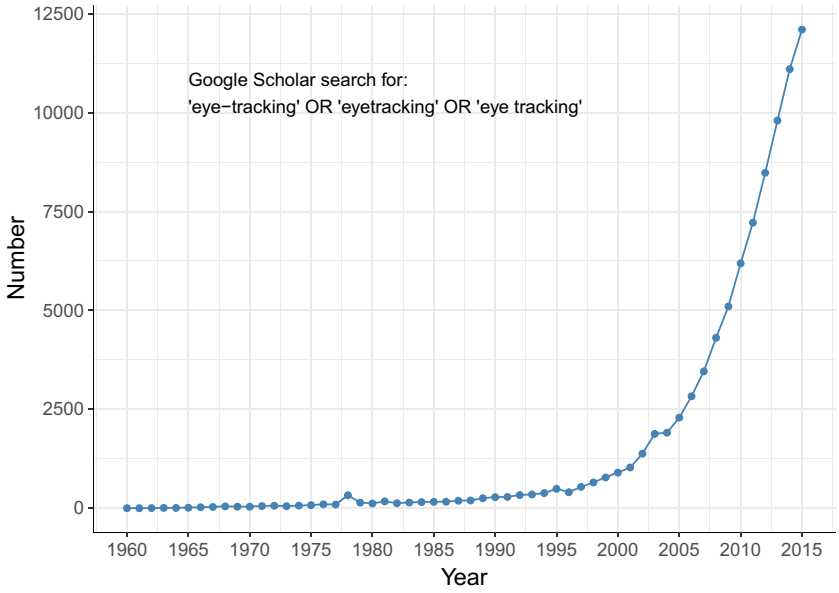

Fig. 1 Number of articles in Google Scholar for eye tracking as a function of year (*not cumulative*)

Zhang, Cheng, & Lin, 2012) or as support for brain–computer interfaces for people with disabilities (e.g., Lee, Woo, Kim, Whang, & Park, 2010; Lim, Lee, Hwang, Kim, & Im, 2015).

The goal of affordable eye trackers is within reach. Recently, do-it-yourself systems with webcams have appeared (Burton, Albert, & Flynn, 2014; Petridis, Giannakopoulos, & Spyropoulos, 2013; Xu et al., 2015), free (libre) software flourishes (e.g., Dalmaijer, Mathôt, & Van der Stigchel, 2013; Lejarraga, Schulte-Mecklenbeck, & Smedema, 2016; Peirce, 2009), and manufacturers have started to offer low-cost, compact devices such as the EyeTribe (The Eye Tribe ApS, Copenhagen) and the Tobii EyeX (Tobii AB, Danderyd).

One problem with mass production is quality control. For books this can be ignored, since the words of Goethe have as much meaning on cheap paper as on vellum, but in the case of microscopes and eye trackers, quality is at stake. In a laboratory study, we evaluated a simple method for testing the accuracy of eye trackers. We compared the new, low-cost EyeTribe device[1] to the well-established[2] SMI-RED 120-Hz device (SensoMotoric Instruments, Teltow, Berlin) in a straightforward way: We ran them simultaneously while conducting a psychological experiment and correlated the data. Here we report a statistic familiar to every psychologist: R^2. We show that the EyeTribe eye tracker is a good instrument, suitable for psychological research as well as applications in everyday life. In the following we first discuss why devices should be compared, how devices can be compared in general, what has already been done with the EyeTribe eye tracker, and how our approach differs. Then we will motivate our experimental approach.

Why should devices be compared?

Imagine you are an eye-tracking researcher in a standard laboratory with one high-quality but expensive eye-tracking device. To conduct an experiment, you have to test participants one by one, needing much more time compared to your colleague who runs a standard cognitive experiment simultaneously with many participants. To increase efficiency, you have to find an eye tracker that is affordable, so that you can equip an entire seminar of about 30 people without breaking your budget. This is feasible with new, low-cost eye trackers such as the EyeTribe or the EyeX. The only problem is that you do not know how accurate the device is or whether its results will differ from those of an already-established device.

[1]Note that at the time of the study, the EyeTribe was the most affordable eye tracker worldwide. In December 2016 The Eye Tribe company stopped development of their products and was acquired by Oculus VR (e.g., Constine, 2016). The results presented here are still relevant to any scientist or practitioner who uses the EyeTribe or a comparable low-cost device. Furthermore, the method we discuss can be generalized to any case where one wants to compare two different devices and simultaneous data acquisition is possible.

[2]Reputation of devices is difficult to quantify. We see the SMI-RED 120-Hz as representative for the whole SMI-RED series, as spatial resolution and gaze position accuracy are identical and only the sampling frequency differs. For some high-quality studies using the RED system, see a list provided in Ooms et al. (2015, p. 5). Some recent high-quality studies also used specifically the SMI-RED 120-Hz: (Eldar & Niv, 2015; Nordmeyer & Frank, 2014; Scholz et al., 2015).

This problem can be reduced to the general question of reproducibility, which is a hot topic at the moment, especially in psychology (Open Science Collaboration, 2015). The reasons why a scientist might not be able to reproduce a colleague's findings are manifold, for instance, sampling error or differences between experimental setups. In eye-tracking research, one obvious difference in the setup might be that a colleague has used a different device to measure pupil size or point of regard. Since eye tracking has not yet reached the standardization of, say, electroencephalography (e.g., Bagić, Knowlton, Rose, & Ebersole, 2011; Beniczky et al., 2013) or functional magnetic resonance imaging (fMRI; e.g., Poldrack et al., 2008), it is even more difficult to reproduce findings (but see COGAIN, 2011; EMRA, 2013). For instance, many algorithms exist for eye detection (Hansen & Ji, 2010), and the pupil can be modeled as either a circle (e.g., Petridis et al., 2013) or an ellipse (e.g., Lin, Klette, Klette, Craig, & Dean, 2003). If one has not built one's own eye tracker from scratch, these algorithms remain proprietary and will provide only a raw pupil size measure. Thus, it is not possible to know a priori how similar the outputs will be between different devices. Although manufacturers provide quality measures, these are usually a bit too optimistic and have to be corrected downward (Nyström, Andersson, Holmqvist, & van de Weijer, 2013).

This is why researchers are interested in testing algorithms and evaluating devices themselves. This interest might also come from the tradition of thoroughness when it comes to psychologists developing diagnostic material (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Because of this rigor, communication is highly standardized and intelligible. Every psychologist can easily find out how reliable a specific test (e.g., an intelligence test) is and whether it is suitable for a specific research question. For eye-tracking devices, this is not necessarily true, although evaluating an eye tracker is not much different from evaluating a diagnostic test.

How can devices be compared?

Imagine the following example: If two different eye-tracking devices measure pupil size, the results of both should correlate with changes in screen luminance, and if both eye trackers measure gaze position, the results should correlate with actual gaze position. One can realize these tests by changing the luminance of the screen and asking participants to look at different locations, while measuring pupil size and gaze.

Two work groups have shown that the EyeTribe performs well in such a test (Dalmaijer, 2014; Ooms et al., 2015). Regardless of whether the EyeTribe or a high-quality device (EyeLink 1000, SR Research, Ottawa; hereafter, EyeLink) is used, pupil size adapts in the same way when the luminance of the screen changes (Dalmaijer, 2014). But psychologists are usually interested not in large pupil size changes due to luminance but in small changes due to cognitive or emotional stimuli. Furthermore, a skeptic will point out the small sample size of five participants and that a forehead rest was used only for the EyeLink. With accuracies of <1 degree of visual angle and reasonable precision values (Dalmaijer, 2014), the EyeTribe is only slightly worse in measuring the point of regard compared to its more expensive competitors, although problems can occur at screen borders (Ooms et al., 2015). Only for studying saccades is the EyeTribe unsatisfactory, since it produces low-quality trajectory and velocity data (Dalmaijer, 2014). This is plausible, as with a frequency of 60 Hz, the device is simply too slow to acquire very accurate saccadic data (see Anderson, Nyström, & Holmqvist, 2010).

Overall, the results are promising, but what is lacking in both studies is a simple correspondence value between the measurements of the devices. If you want to convince your funding agency to buy 30 low-cost eye trackers instead of one reputable but more expensive device, you still have to rely at least partly on a qualitative argument.

A simpler evaluation is to conduct a standard experiment while running both eye trackers simultaneously and then correlate the measures of the devices themselves. The major advantage is perfect standardization, whereas during two separate administrations slightly different light conditions or mood states of participants will induce error variance, leading to underestimations of correspondence. Two problems arise when trying to run devices simultaneously: First, the eye trackers have to be positioned one above the other, potentially leading to different calibration accuracies. Second, the amount of infrared light radiated by an eye tracker is optimized to that specific eye tracker. Adding a second infrared source by positioning another eye tracker nearby may break the algorithms for eye detection and calibration. Dalmaijer (2014) reported such calibration problems with the EyeLink, which operates on the assumption of only a single source of infrared light.

Yet Popelka et al. (2016) did not have problems employing such a design, probably because they used a different device (SMI-RED-250) from Dalmaijer. They showed that the EyeTribe is good enough for cartographic research, although fixations at the border region (in this case especially the bottom region) were shifted upward. Exploiting the parallel design, they reported a correlation of the aggregated number of fixations. They could have expanded this statistic by an overall correspondence measure of x and y coordinates as well as pupil size. In psychology, pupil size is a popular metric (e.g., Laeng, Sirois, & Gredeback, 2012), but it is irrelevant for cartographic research.

Furthermore, Popelka et al. (2016) used a method of analysis that seems suboptimal. They reported that data loss and calibration quality varied between participants (e.g., because one participant had eyeglasses), but this individuality is not considered when the metrics are compared because data are averaged across participants.

In sum, three recent studies provide evidence that low-cost eye trackers such as the EyeTribe can produce comparable results to well-established devices such as the SMI-RED-250 or the EyeLink 1000. All three studies looked at point-of-regard measures, whereas only one study investigated pupil size. In this study, pupil-size variations were triggered by changing luminance and not by psychologically relevant variables such as arousal. Only one of the studies employed a parallel setup, but it did not compare pupil size. All studies lack a simple measure of correspondence and a proper statistical model that controls for differences between subjects. To avoid these shortcomings, we designed an experiment that focused on psychological pupil-size effects to test the EyeTribe quality. We provide a simple measure of correspondence (R^2) between the EyeTribe and the SMI-RED 120-Hz, while statistically controlling for calibration quality by using a mixed effects model.

Design considerations

To trigger small changes in pupil size, we showed participants stimuli that varied in their arousal. The pupil acts like the aperture of a camera; it dilates or constricts to optimize the amount of light reaching the sensor (Loewenfeld & Lowenstein, 1999), which in this case is the retina. Unlike the technical analogue, the pupil also reacts to emotional stimuli and cognitive demands (for an overview see Laeng et al., 2012). For instance, when participants watched erotic material, the pupil dilated more strongly compared to a control group watching neutral material (Aboyoun & Dabbs, 1998; Hamel, 1974; Hess & Polt, 1960; Peavler & McLaughlin, 1967; Rieger & Savin-Williams, 2012). When participants had to mentally calculate the product of two numbers, the pupil dilated more strongly when the task was more difficult (Ahern & Beatty, 1977; Klingner, Tversky, & Hanrahan, 2011), and when participants had to react to an incongruent trial in the Stroop task, the pupil dilated more strongly than for a congruent or neutral trial (Laeng, Ørbo, Holmlund, & Miozzo, 2011).

Based on these findings, we chose images and words with different arousal ratings from standardized databases to induce different intensities of information processing. Bradley, Miccoli, Escrig, and Lang (2008) used a similar paradigm and showed that the pupil dilates more strongly for high-arousal images. We believe this to be a good standard paradigm to test the validity of new eye trackers because it allows one to study pupillometry effects but does not neglect measures of gaze behavior that are more commonly used in eye tracking. Participants were able to observe the presented stimuli freely, so there should be some natural variation of gaze behavior. We expected to find high correlations between the two devices in pupil size and the x and y coordinates in all experimental conditions. In addition, pupil size was expected to be higher for high-arousal stimuli.

Method

Participants, apparatus, and software

We recruited 26 participants (81% female, mean age = 22.5 years, SD = 3.7 years) from the Chemnitz University of Technology to participate in our study. As the low-cost eye tracker, we selected the EyeTribe, the cheapest device known to us at the time of the study in 2015 ($99). As a comparison device, we chose the established SMI-RED 120-Hz (hereafter, SMI). Following the procedure of Popelka et al. (2016), we positioned the EyeTribe above the SMI to obtain data simultaneously and used a chin-and-forehead rest to reduce head movement to a minimum (Fig. 2).

The experiment was programmed in PsychoPy (Peirce, 2009) with the PyTribe package (Dalmaijer, 2014) for communication with the EyeTribe and ioHub for communication with the SMI. We performed preprocessing and statistical analyses of the data in R (R Core Team, 2016), mostly with self-written functions and the saccades package (von der Malsburg, 2015) for fixation detection. Note that we explicitly did not use SMI's software to analyze data because this software is proprietary and thus we do not know what exact algorithms are used. To make an objective

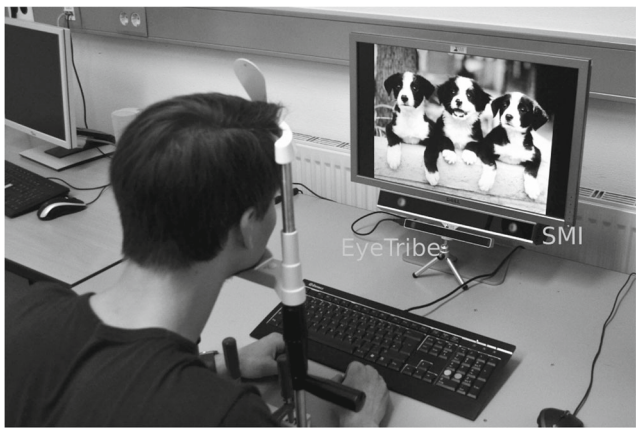

Fig. 2 Setup of experiment with the SMI-RED 120-Hz eye tracker positioned above the EyeTribe eye tracker. Note that the images presented on the screen were actually in gray scale

comparison, the same calculations should be performed on raw data; otherwise correspondence might be underestimated.

Procedure and design

The experimenter calibrated each tracker separately with its internal calibration procedure. Before the experiment started, the calibration quality was assessed automatically with a self-programmed nine-point validation. The actual memory experiment was a mixed design, where participants saw either 64 different pictures or 64 different words (between subjects). We set the interstimulus interval to 0 to reduce large pupil variations due to luminance changes caused by a blank screen (e.g., Bradley et al., 2008).[3] Half of the stimuli were high and the other half low arousal (within subjects), which was the most important manipulation for the evaluation of the pupil-size metric of the two eye trackers. Half of the stimuli were presented once and half two times (within subjects, total: 96 stimuli) for a duration of 5, 6, 7, or 8 s (within subjects), which was mainly important for a memory experiment, to be reported separately. After the presentation phase participants made recognition and duration judgments of targets and distractors, which again are not important to show the validity of the eye trackers. Tracking of the eyes started with the first instruction slide, but the analyses were performed only on the data that were recorded between the onset of the first stimulus and the offset of the last stimulus.

Material

We selected images from the International Affective Picture System (IAPS; Lang, Bradley, & Cuthbert, 1999) and words from the Leipzig Affective Norms for German (LANG; Kanske & Kotz, 2010, 2011). The mean rating for the selected images was 3.12 for low and 5.91 for high arousal and for the selected words 2.24 and 6.01 for low and high arousal, respectively (scales were both from 1 to 9). Examples for every stimulus condition are a fork as a low-arousal image, a nature scene with a cheetah as a high-arousal image, "Monat" (month) as a low-arousal word, and "Angriff" (attack) as a high-arousal word.

The stimuli were presented on a Dell P2210 22-in. monitor with a resolution of 1680 × 1050 pixels (42.78 × 28.07 degrees of visual angle). Participants sat at a distance of 60 cm. Images were at a resolution of 800 × 600 pixels (26.90 × 20.78 degrees of visual angle). For words, the visual angle differed because number and type of letters varied, with a maximum of 30.75 and 6.68 degrees horizontally and vertically, respectively.

To reduce different lightness responses of different participants to colors, we transformed all images to gray scale. Furthermore, we adjusted the luminance of every image to the average luminance of all images with the help of the EBImage package in R. To validate that the monitor emitted, the same amount of light independent of the arousal condition, we measured illuminance with a luxmeter (iClever, LU1 LX1010BS) at about the position where the participants' eyes would be located. As expected, only minimal, nonsystematic differences were found.[4] The windows were blacked out, leaving the monitor and ceiling light as the only sources of illumination.

Preprocessing

Timing synchronization of the eye trackers was problematic, as in all experiments with several devices. The EyeTribe was running at 60 Hz, and the SMI at 120 Hz. Even if the devices had been running at the same frequency, time stamps would not be identical. To solve this problem, we linearly interpolated both devices to a frequency of exactly 60 Hz (the SMI was downsampled). This is necessary to correlate the two time series and calculate average values across stimuli and participants for specific time points.

Both eye trackers have a state variable that displays whether the device is currently recognizing and tracking the eye. For the main analysis, only data where both trackers had valid states were analyzed. This approach already subsumes the removal of blinks in a satisfactory way for our study. Since we wanted to restrict our analyses to fixations, we used an algorithm (Engbert & Kliegl, 2003) to detect and exclude saccades with the R saccades package (von der Malsburg, 2015). Preprocessing was part of the correspondence analysis of the devices and relative frequencies of removed data are reported in the Results section.

Pupil size is given in arbitrary units, which is not a problem for a simple correlation analysis. To see how pupil size develops during the presentation of a stimulus, we standardized pupil size by subtracting a baseline average of 400 ms before stimulus onset (cf. Klingner et al., 2011) and dividing by that baseline average, giving us a percentage increase or decrease in comparison to the baseline period. This made it easier to compare the devices and also to compare our results to established findings in the literature (e.g., the

[3] Thus, no fixation cross was used before each presentation.

[4] High-arousal images had a mean illuminance of 86.06 lux, low arousal images of 85.91 lux [the difference is not significant at an α-level of 5%, $t(62) = 1.417$, $p = 0.162$]. For high-arousal words, the mean illuminance was 88.38 lux, for low-arousal words 88.56 lux [the difference is not significant at an α-level of 5%, $t(62) = -0.643$, $p = 0.523$]. Note that we were not interested in comparing the illuminance of images with words, as these are fundamentally different stimuli.

assumption that psychologically relevant stimuli can evoke a maximal dilation of about 20%, Laeng et al., 2012).

Calibration quality was estimated by first calculating the distance of the nine calibration points and the corresponding points of regard in pixels, then averaging them separately for the horizontal and vertical dimensions (cf. Dalmaijer, 2014). Although for an overall measure of accuracy it makes sense to take the absolute difference before averaging, we refrained from this because we needed to know in which direction a possible shift would go in order to use this information as a predictor for the correspondence between the two devices.

Statistical analysis

Since we are dealing with a partial within-subject design, the appropriate choice of analysis is a mixed model, which accounts for individual variation and different sample sizes (e.g., Hox, Moerbeek, & van de Schoot, 2010). The analysis was performed with the nlme package in R (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2016), and R^2 was categorized into two types: marginal and conditional (Nakagawa & Schielzeth, 2013). Marginal R^2 represents the variance explained by fixed factors (individual differences are not included), whereas conditional R^2 is interpreted as variance explained by both fixed and random factors (individual differences are included). We specified the model with fixed slopes and random intercepts. Theoretically, the slope should not vary between participants. For instance, a change of 1 pixel in the SMI should lead to a change of 1 pixel in the EyeTribe, but the intercept will likely vary because calibration quality differs between participants.[5]

Results

In the following, we first report whether the two eye trackers correspond on fundamental measures of technical tracking state (recognizing eyes and tracking gaze) and eye-tracking state (fixation or saccade). Then we look at the correlation between the devices for pupil size, as a measure of convergent validity, and also explore how the pupil reacts to stimuli that vary in their arousal, as a measure of criterion validity. Finally, we analyze the correspondence concerning the point of regard.

The SMI tracks more accurately with 93.32% of all states being valid (recognizing eyes and tracking gaze), compared to 91.47% for the EyeTribe. Although with such a large sample size, the standard errors are microscopic (0.041% and 0.047%, respectively), the absolute difference is still negligible for practical applications. In 88.52% of all cases both devices showed a valid state and in 3.72% both showed an invalid state, yielding an agreement of 92.24%. We assume that the nonoverlapping states can be attributed to random technical dropouts of the devices, which are statistically independent. Of the overlapping valid eye-tracking data (92.24% from above), 88.24% were classified as a fixation for both devices and 4.16% as a saccade for both devices, making a total of 92.40% of correspondence. Thus, we found a substantial overlap of the two devices for technical tracking state (valid/invalid) as well as eye-tracking state (fixation, saccade).

The following analyses were restricted to valid states that were classified as fixations to reduce error variance. This is a realistic scenario for preprocessing: removing states where the device is not tracking and analyzing only fixations where the eye can be measured more accurately because it is stable.[6]

If the pupil size that one device is measuring is known, one can accurately predict what the other device will measure (Fig. 3a, Table 1, Models 1 and 2). If it is additionally known who the participant is, one can predict pupil size even more accurately, since about 9% of the variance can be accounted for by differences between participants. This variance cannot be explained by calibration quality, since R^2 does not rise much for the second model, which controls for how accurately the x and y gaze coordinates were estimated after calibration.

Although this evidence is compelling, a skeptic might argue that pupil size did not vary much during the experiment and high correlations are unsurprising. This is similar to the critique that even when two intelligence tests correlate highly, they may still not measure intelligence. To counteract this argument, we looked at how the pupil develops during the presentation of stimuli that differ in their arousal. Recall that from our theoretical argument we expected that high-arousal stimuli would provoke a stronger pupil dilation than low-arousal stimuli, but even if the manipulation had failed, we could still test whether the two devices show a similar pattern in pupil-size development. In fact, the pupil develops in a characteristic way when a stimulus is presented (Fig. 3b): a fast and strong constriction, followed by a dilation (consistent with Bradley et al., 2008; Naber, Frässle, Rutishauser, & Einhäuser, 2013). For high-arousal images, the dilation is stronger than for low-arousal images. For words, the manipulation might have been too weak to

[5] More complex models with random slopes could be specified, but then more parameters would have to be estimated, which has to be justified.

[6] We have also performed analyses where we only removed invalid states. On average we found slightly smaller effect sizes, and the calibration quality was predictive of only the y coordinate (in the original analysis it was predictive of the x coordinate as well). The main conclusions are not affected by this alternative analysis because the correspondence between the two devices is still large.

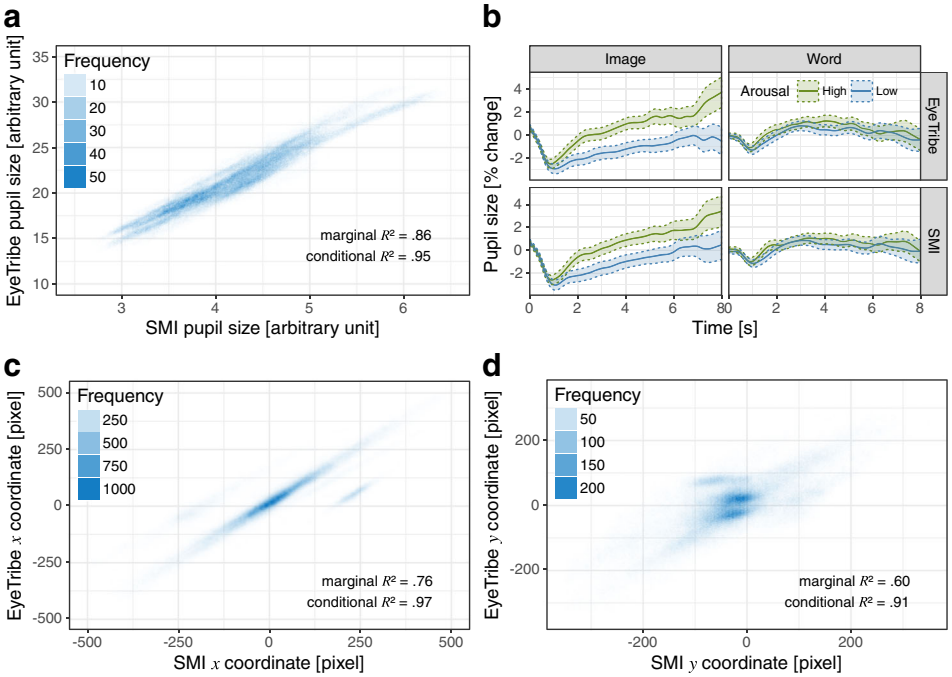

Fig. 3 **a** EyeTribe pupil size as a function of SMI pupil size. **b** Developing of pupil size over time as a function of eye tracker, arousal, and stimulus type; *shaded areas* represent 1 SE. Pupil size is given as percentage change from baseline (400 ms average before stimulus onset). **c** EyeTribe gaze *x* coordinate as a function of SMI gaze *x* coordinate. **d** EyeTribe gaze *y* coordinate as a function of SMI gaze *y* coordinate

Table 1 Mixed effect models for the correspondence in eye-tracking measures between EyeTribe and SMI

	Dependent variable					
	SMI pupil size		SMI *x* coordinate		SMI *y* coordinate	
Independent Variable	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)
ET pupil size	0.17	0.17				
	(0.16, 0.19)	(0.16, 0.19)				
ET *x*			0.98	0.98		
			(0.95, 1.02)	(0.95, 1.02)		
ET *y*					−0.89	−0.89
					(−0.97, −0.80)	(−0.97, −0.80)
ET *x* accuracy		0.001		−0.51		
		(−0.001, 0.003)		(−1.23, 0.20)		
SMI *x* accuracy		−0.0004		0.46		
		(−0.001, 0.001)		(0.24, 0.67)		
ET *y* accuracy		0.0002				0.38
		(−0.003, 0.003)				(−0.25, 1.01)
SMI *y* accuracy		0.0003				0.96
		(−0.002, 0.002)				(0.64, 1.27)
Constant	0.46	0.45	−23.70	−5.47	−31.23	−8.69
	(0.15, 0.77)	(0.14, 0.75)	(−50.26, 2.86)	(−28.33, 17.39)	(−54.39, −8.08)	(−25.69, 8.32)
R^2_{marginal}	.86	.86	.76	.84	.60	.74
$R^2_{\text{conditional}}$	.95	.95	.97	.97	.91	.90

Note. The effects are unstandardized regression slopes from a mixed random intercepts model. The *values in parentheses* are 95% confidence intervals for the corresponding effect. *ET* EyeTribe; *SMI* SMI-RED 120-Hz; accuracy is the average distance between the calibration point and the actual point of regard for fixations in pixels that was established in a calibration procedure. $n_{\text{level1}} = 525, 638$, $n_{\text{level2}} = 26$

see differences between the arousal conditions. Overall, one can conclude that it does not matter which device one uses for pupillometry, since they produce almost identical results.

Compared to the pupil size, the *x* coordinates correspond slightly less between the devices (Fig. 3c, Table 1, Model 3) and the *y* coordinates show the smallest correspondence (Fig. 3d, Table 1, Model 5). Still they overlap substantially, and when we include the participant-specific component, we can explain about as much variance as for pupil size. The graphs in Fig. 3 show why the participant-specific component is more important for the coordinates: For some participants, the regression is translated, probably resulting from a distorted calibration. When we include a simple measure of calibration quality, R^2_{marginal} rises substantially (Table 1, Models 4 and 6); in absolute values the gain is 8 and 14% of variance for the *x* and *y* coordinates, respectively. Although this finding appears trivial, in other studies calibration quality has not been included in the statistical models, which can potentially lead to different results.

Discussion

We tested whether a low-cost eye tracker can be used in psychological research without sacrificing accuracy to a noticeable degree. Specifically, we compared the new, affordable EyeTribe eye tracker with a more expensive established device (SMI-RED 120-Hz). We were motivated by employing the simplest method one could think of to compare two devices that would still be useful. Thus, we copied what psychologists have been practicing in psychological diagnostics for decades: correlating constructs to show reliability and validity. We ran two eye trackers simultaneously and correlated the eye-tracking metrics to produce a simple correspondence measure between the devices.

For pupil size, we found straightforward results: The correlation between the two devices is high and the development of the pupil size during a stimulus presentation shows a characteristic pattern that is analogous to previous research. We conclude, in agreement with other researchers (Dalmaijer, 2014), that as long as scientists reduce head movement of participants (e.g., via a chin rest and/or forehead rest), they can rely on the EyeTribe as much as on a more expensive device for pupillometry research. In general, we suggest using a mixed model for analyses to control for individual differences between participants in pupil-size effects. In the simple case of a correlation between two devices, about 9% of the variance is caused by differences between participants and this variance cannot be explained by calibration quality.

For the *x* and *y* coordinates, the results are more varied: The correlation between the two devices is only high when we include the variance between the participants. A part of this variance reduces to calibration quality. Researchers interested in gaze behavior could include calibration quality in the regression model. This way, the results will be more accurate and easier to generalize to research with other eye-tracking devices. The *y* coordinates show the lowest correspondence between the two devices, probably because the trackers were placed one above the other, always measuring slightly different positions, but it could also be a problem of the EyeTribe device itself, since Dalmaijer (2014) reported the worst accuracy for the *y* coordinate. In our study, the 95% confidence interval for the fixed effect of the *y* coordinate also excludes 1.0 (the expected correct value), which is not the case for the *x* coordinate. Furthermore, Popelka et al. (2016) noted that for the bottom region, the gaze position of the EyeTribe is shifted upward, so one would expect less correspondence between the devices for the *y* position. Even though this sounds problematic, as long as interindividual differences are included in the model, the correspondence is still substantial (a correlation of over 0.9).

We can generalize that under similar conditions (e.g., preprocessing, material) psychologists can use the EyeTribe for pupillometry and expect that their results will only deviate marginally from those obtained with an SMI-RED 120-Hz. To be more specific, if one exchanges the SMI for an EyeTribe, one can expect to find a correlation between studies of about 0.97 (the root of $R^2 = .95$) for pupillometry if the same participants take part in identical experiments and behave in exactly the same way. This value should be regarded as a theoretical upper limit because there will always be additional sources of error variance. Still, the communication of this result is straightforward because it is a relationship in context and not just a single measure of accuracy or precision. Even a novice in eye tracking will be able to interpret this result and decide whether using a low-cost device is justifiable in certain situations.

For point of regard, we have found slightly smaller correlations between the devices that are still large enough to be judged accurate for psychological science. Still, our study was not exclusively aimed at comparing point of regard, so participants, most of the time, looked at the center of the screen. One reason for this is that in contrast to the pupillometry, we had no manipulation that focused on point of regard. A new study could employ a design that manipulates gaze allocation in order to test the correspondence between the devices on a wider range of *x* and *y* coordinates.

Point of regard between the two devices is somewhat more highly correlated when calibration quality at the participant level is controlled for. Besides being relevant with regards to content, this demonstrates the flexible applicability of mixed models. Most of the time, eye-tracking data will have a within-subject factor to reduce the cost of testing many participants one at a time. The appropriate model to analyze the data will almost always be a two-level mixed

model that incorporates interindividual differences. Further, z transformations that are common in pupillometric studies (e.g., Naber et al., 2013; Smallwood et al., 2011) will become obsolete. Instead, scientists will focus on explaining differences between participants in eye-tracking effects, enriching theoretical models.

One conclusion from our study is that cheaper eye trackers may yield results that are equivalent to those from expensive devices, but when presenting the complete cost/benefit analysis, one should not forget about other costs besides that of the device itself, the most important being time. Spending more time setting up a cheaper device will be inevitable for psychologists who have no experience with programming, because cheaper devices are targeted at developers. Furthermore, no sophisticated software analysis packages are enclosed, so inexperienced users will also spend more time making sense of the raw data. Since scientists are well paid, the additional time might quickly exceed the cost of the more expensive device. Thus, if a psychologist wants a perfectly working all-in-one solution, it may be reasonable to buy a more expensive device that also includes support from the company. In the long term, this argument might carry less weight, because free (libre) software for eye-tracking experiments and analysis already exists (Dalmaijer et al., 2013; Lejarraga et al., 2016; Peirce, 2009) and will likely become more accessible to nonprogrammers.

The EyeTribe in our experiment was running at 60 Hz, so our results can potentially be generalized to other devices with a similar frequency, such as Tobii's EyeX that runs at 70 Hz or the second EyeTribe version that has a frequency between 30 and 75 Hz, but this needs to be tested. Furthermore, the method itself can be generalized to any case where one wants to compare two different devices and simultaneous data acquisition is possible (e.g., heart rate monitoring, galvanic skin response).

Finally, we would like to cover the possibilities opening up with low-cost eye trackers that produce high-quality data. As mentioned before, eye trackers can function as sleep detectors in cars, and with the advent of affordable accurate devices they could be included as standard in every vehicle. This would make driving much safer, as sleepiness is clearly associated with accidents (e.g., Lyznicki, Doege, & Davis, 1998; Pedn et al., 2004). If everyone could afford an eye tracker, these "personal devices" could enhance interaction with computers. To give a simple example: Once a user has left his or her workspace, a tracker could register that it no longer detects the user's eyes and put the monitor in power-saving mode. With personal eye trackers it would also be possible to conduct mass online studies, not only for scientific purposes but also for applications such as usability testing. The availability of inexpensive devices means researchers could afford more of them. Universities could buy dozens of eye trackers that could be used in seminars on empirical experimentation, replicating classic experiments. Recently, Lejarraga et al. (2016) made this potential a reality. They developed a framework that makes it possible to simultaneously track the eyes of several participants, such that one participant could see information based on the eye metrics of others. This opens the potential to conduct experiments one would not have thought of before. We believe there is plenty of room for more innovation to come.

We have shown that the EyeTribe tracker as a representative of low-cost eye trackers does not have to hide behind its established siblings. It is compact, affordable, and accurate. It makes a perfect "paperback edition" and may contribute to a wider use of eye tracking in science as well as to the ever-growing level of technology (e.g., virtual reality) in society.

Acknowledgements Johannes Titz gratefully acknowledges the support of the German National Academic Foundation. Agnes Scholz gratefully acknowledges the support of the Swiss National Science Foundation (grant PP00P1_157432). Furthermore, the authors would like to thank Edwin Dalmaijer and Michael Schulte-Mecklenbeck for comments on an earlier version of the manuscript. Data and code to reproduce Table 1 and Figure 3 are available on the gitlab server of the Chemnitz University of Technology (Link: https://gitlab.hrz.tu-chemnitz.de/titz--tu-chemnitz.de/eyetracker-comparison).

References

Aboyoun, D. C., & Dabbs, J. M. (1998). The Hess pupil dilation findings: Sex or novelty?. *Social Behavior and Personality: An International Journal*, *26*(4), 415–419. https://doi.org/10.2224/sbp.1998.26.4.415.

Ahern, S., & Beatty, J. (1977). Pupillary responses during information processing vary with scholastic aptitude test scores. *Science*, *205*(4412), 1289–1292.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Anderson, R., Nyström, M., & Holmqvist, K. (2010). Sampling frequency and eye-tracking measures: How speed affects durations, latencies, and more. *Journal of Eye Movement Research*, *3*(3), 1–12. Retrieved from http://portal.acm.org/citation.cfm?doid=1344471.1344500.

Bagić, A. I., Knowlton, R. C., Rose, D. F., & Ebersole, J. S. (2011). American clinical magnetoencephalography society clinical practice guideline 1. *Journal of Clinical Neurophysiology*, *28*(4), 1. https://doi.org/10.1097/WNP.0b013e3182272fed.

Beniczky, S., Aurlien, H., Brogger, J. C., Fuglsang-Frederiksen, A., Martins-Da-Silva, A., Trinka, E., & Wolf, P. (2013). Standardized computer-based organized reporting of EEG: SCORE. *Epilepsia*, *54*(6), 1112–1124. https://doi.org/10.1111/epi.12135.

Bradbury, S. (1967). *The evolution of the microscope*. Oxford: Pergamon Press.

Bradley, M. M., Miccoli, L. M., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and automatic activation. *Psychophysiology*, *45*(4), 602. https://doi.org/10.1111/j.1469-8986.2008.00654.x.

Burton, L., Albert, W., & Flynn, M. (2014). A comparison of the performance of webcam vs. infrared eye tracking technology. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 1437–1441. https://doi.org/10.1177/1541931214581300.

COGAIN (2011). COGAIN - Communication by Gaze Interaction. Retrieved from http://wiki.cogain.org/index.php/Main_Page.

Constine, J. (2016). Oculus acquires eye-tracking startup The Eye Tribe. TechCrunch. Retrieved from https://techcrunch.com/2016/12/28/the-eye-tribe-oculus/.

Dalmaijer, E. S. (2014). Is the low-cost EyeTribe eye tracker any good for research *PeerJ*, *606901*, 1–35. https://doi.org/10.7287/peerj.preprints.585v1.

Dalmaijer, E. S., Mathôt, S., & Van der Stigchel, S. (2013). PyGaze: An open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior Research Methods*, *46*(4), 1–16. https://doi.org/10.3758/s13428-013-0422-2.

Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications*, *6*, 1–9. https://doi.org/10.1038/ncomms7149.

EMRA (2013). About EMRA. Retrieved from http://www.eye-movements.org/about.

Engbert, R., & Kliegl, R. (2003). Microsaccades uncover the orientation of covert attention. *Vision Research*, *43*(9), 1035–1045. https://doi.org/10.1016/S0042-6989(03)00084-1.

Guenter, B., Finch, M., Drucker, S., Tan, D., & Snyder, J. (2012). Foveated 3d graphics. *ACM Transactions on Graphics (TOG)*, *31*(6), 164.

Hamel, R. F. (1974). Female subjective and pupillary reaction to nude male and female figures. *The Journal of Psychology*, *87*(2), 171–175. https://doi.org/10.1080/00223980.1974.9915687.

Hansen, D. W., & Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE transactions on pattern analysis and machine intelligence*, *32*(3), 478–500. https://doi.org/10.1109/TPAMI.2009.30.

Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science*, *132*(3), 3–4.

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2010). Multilevel analysis: Techniques and applications. Routledge.

Johann, E., & Junker, J. (1970). *Illustrierte deutsche Kulturgeschichte der letzten hundert Jahre*. München: Nymphenburger Verlagshandlung.

Kanske, P., & Kotz, S. A. (2010). Leipzig Affective Norms for German: A reliability study. *Behavior Research Methods*, *42*(4), 987–991. https://doi.org/10.3758/BRM.42.4.987.

Kanske, P., & Kotz, S. A. (2011). Cross-modal validation of the leipzig affective norms for German (LANG). *Behavior Research Methods*, *43*(2), 409–413. https://doi.org/10.3758/s13428-010-0048-6.

Klingner, J., Tversky, B., & Hanrahan, P. (2011). Effects of visual and verbal presentation on cognitive load in vigilance, memory, and arithmetic tasks. *Psychophysiology*, *48*(3), 323–332. https://doi.org/10.1111/j.1469-8986.2010.01069.x.

Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary stroop effects. *Cognitive Processing*, *12*(1), 13–21. https://doi.org/10.1007/s10339-010-0370-z.

Laeng, B., Sirois, S., & Gredeback, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). International affective picture system (IAPS): Instruction manual and affective ratings. The center for research in psychophysiology, University of Florida.

Lee, E. C., Woo, J. C., Kim, J. H., Whang, M., & Park, K. R. (2010). A brain–computer interface method combined with eye tracking for 3D interaction. *Journal of Neuroscience Methods*, *190*(2), 289–298. https://doi.org/10.1016/j.jneumeth.2010.05.008.

Lejarraga, T., Schulte-Mecklenbeck, M., & Smedema, D. (2016). The pyeTribe: Simultaneous eyetracking for economic games. Behavior Research Methods. https://doi.org/10.3758/s13428-016-0819-9.

Lim, J. H., Lee, J. H., Hwang, H. J., Kim, D. H., & Im, C. H. (2015). Development of a hybrid mental spelling system combining SSVEP-based brain–computer interface and webcam-based eye tracking. *Biomedical Signal Processing and Control*, *21*, 99–104. https://doi.org/10.1016/j.bspc.2015.05.012.

Lin, X., Klette, G., Klette, R., Craig, J., & Dean, S. (2003). *Accurately Measuring the Size of the Pupil of the Eye CITR*. New Zealand: University of Auckland. Retrieved from http://sprg.massey.ac.nz/ivcnz/proceedings/ivcnz%7B%5C_%7D40.pdf.

Loewenfeld, I. E., & Lowenstein, O. (1999). *The pupil: Anatomy, physiology, and clinical applications*. Boston: Butterworth-Heinemann.

Lyznicki, J. M., Doege, T. C., & Davis, R. M. (1998). Sleepiness, Driving, and Motor Vehicle Crashes. *JAMA*, *279*(23), 1908–1913. https://doi.org/10.1001/jama.279.23.1908.

Naber, M., Frässle, S., Rutishauser, U., & Einhäuser, W. (2013). Pupil size signals novelty and predicts later retrieval success for declarative memories of natural scenes. *Journal of Vision*, *13*(2), 11. https://doi.org/10.1167/13.2.11.doi.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x.

Nguyen, T. P., Chew, M. T., & Demidenko, S. (2015). Eye tracking system to detect driver drowsiness. In *ICARA 2015 - Proceedings of the 2015 6th International Conference on Automation, Robotics and Applications, (April 2015), 472–477*. https://doi.org/10.1109/ICARA.2015.7081194.

Nordmeyer, A. E., & Frank, M. C. (2014). The role of context in young children's comprehension of negation. *Journal of Memory and Language*, *77*, 25–39. https://doi.org/10.1016/j.jml.2014.08.002.

Nyström, M., Andersson, R., Holmqvist, K., & van de Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, *45*(1), 272–288. https://doi.org/10.3758/s13428-012-0247-4.

Ooms, K., Lapon, L., Dupont, L., & Popelka, S. (2015). Accuracy and precision of fixation locations recorded with the low-cost Eye Tribe tracker in different experimental set-ups. *Journal of Eye Movement Research*, *8*(1), 1–24. https://doi.org/10.16910/jemr.8.1.5.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716. https://doi.org/10.1126/science.aac4716.

Pai, Y. S., Tag, B., Outram, B., Vontin, N., Sugiura, K., & Kunze, K. (2016). GazeSim. In *Acm siggraph 2016 posters on - siggraph '16 (pp. 1–2). New York, New York, USA: ACM Press*. https://doi.org/10.1145/2945078.2945153.

Peavler, W. S., & McLaughlin, J. P. (1967). The question of stimulus content and pupil size. *Psychonomic Science*, *8*(12), 505–506. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000090.

Peden, M., Scurfiled, R., Sleet, D., Mohan, D., Hyder, A., Jarawan, E. (Eds.) (2004). World report on road traffic injury prevention. Geneva: World Health Organization.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 1–8. https://doi.org/10.3389/neuro.11.010.2008.

Petridis, S., Giannakopoulos, T., & Spyropoulos, C. D. (2013). Unobtrusive low cost pupil size measurements using web cameras, 1–6. arXiv:1311.7327. Retrieved from http://arxiv.org/abs/1311.7327.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team (2016). nlme: linear and nonlinear mixed effects models. R package version 3.1-128. Retrieved from http://CRAN.R-project.org/package=nlme.

Poldrack, R. A., Fletcher, P. C., Henson, R. N., Worsley, K. J., Brett, M., & Nichols, T. E. (2008). Guidelines for reporting an fMRI study. *NeuroImage*, *40*(2), 409–414. https://doi.org/10.1016/j.neuroimage.2007.11.048.

Popelka, S., Stachon, Z., Sasinka, C., & Dolezalova, J. (2016). Eyetribe tracker data accuracy evaluation and its interconnection with hypothesis software for cartographic purposes. *Computational Intelligence and Neuroscience*, *2016*, 1–14. https://doi.org/10.1155/2016/9172506.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from https://www.R-project.org/.

Rieger, G., & Savin-Williams, R. C. (2012). The eyes have it: Sex and sexual orientation differences in pupil dilation patterns. *PLoS ONE*, *7*(8), 1–10. https://doi.org/10.1371/journal.pone.0040256.

Scholz, A., Franke, T., Platten, F., & Attig, C. (in press). Eye movements in vehicle control. In Klein, C., & Ettinger, U. (Eds.) *An introduction to the scientific foundations of eye movement research and its applications*. Heidelberg: Springer.

Scholz, A., von Helversen, B., & Rieskamp, J. (2015). Eye movements reveal memory processes during similarity- and rule-based decision-making. *Cognition*, *136*, 228–246. https://doi.org/10.1016/j.cognition.2014.11.019.

Smallwood, J., Brown, K. S., Tipper, C., Giesbrecht, B., Franklin, M. S., Mrazek, M. D., & Schooler, J. W. (2011). Pupillometric evidence for the decoupling of attention from perceptual input during offline thought. *PLoS ONE*, *6*(3), 1–8. https://doi.org/10.1371/journal.pone.0018298.

von der Malsburg, T. (2015). Saccades: Detection of fixations in eye-tracking data. R package version 0.1-1. Retrieved from https://CRAN.R-project.org/package=saccades.

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). Turkergaze: Crowdsourcing saliency with webcam based eye tracking. CoRR, abs/1504.06755. Retrieved from http://arxiv.org/abs/1504.06755.

Zhang, W., Cheng, B., & Lin, Y. (2012). Driver drowsiness recognition based on computer vision technology. *Tsinghua Science and Technology*, *17*(3), 354–362. https://doi.org/10.1109/TST.2012.6216768.