Independent Comparative Evaluation of the Pupil Neon - A New Mobile Eye-tracker

Valentin Foucher[1], Alina Krug[1], and Marian Sauter[1]

[1]Ulm University, Insititute of Psychology, General Psychology

Author Note

September 16, 2024: Stage 1 Submission for Peer Community in Registered Reports

Valentin Foucher https://orcid.org/0009-0000-2632-3519

Alina Krug https://orcid.org/0009-0004-7088-1584

Marian Sauter https://orcid.org/0000-0003-3123-8073

Abstract

Due to the rapid adoption of (mobile) eye-tracking devices in both academic and consumer research, it becomes more important that the increasing number of datasets is based on reliable recordings. This study provides an independent evaluation of the Pupil Neon (Pupil Labs GmbH), one of the newest and most affordable mobile eye-trackers, by comparing its performance on a variety of tasks to the EyeLink 1000 Plus (SR Research Ltd.). Using Ehinger et al. (2019)'s test battery, a set of 10 tasks evaluated the accuracy and its decay over time of some of the most common eye-tracking-related parameters: fixations, saccades, smooth pursuit, pupil dilation, microsaccades, blinks, and the influence of head motion on accuracy. Gaze position, eye movements and pupil diameter associated with each task were recorded simultaneously by the two eye-trackers and compared concurrently. The results provide some ideas on what singularities should be expected by the newer Pupil Neon for the recording of specific eye movements or the performance in various kinds of tasks.

Keywords: eye tracking, mobile eye-tracker, Pupil Neon, Eyelink 1000 Plus, performance evaluation

Independent Comparative Evaluation of the Pupil Neon - A New Mobile Eye-tracker

Introduction

The saying "One look is worth a thousand words" highlights the significant role of eye movements in understanding how individuals perceive and interpret their world. This concept has been extensively applied in fields such as psychology and human-computer interaction (Duchowski, 2007; Majaranta & Bulling, 2014). Over the past decades, eye-trackers, once confined to a small group of researchers, have become widely available to a broader audience (Duchowski, 2018; Gunawardena et al., 2022), including applied researchers (Ahlström et al., 2021) and practitioners in marketing and gaming (Mancini et al., 2022). The increase in reliability, coupled with less invasive devices and more affordable prices, has democratized the use of eye-trackers to study human behavior. However, the expanding range of eye-tracking applications makes it crucial to understand the performance of current eye-trackers and how their capabilities and limitations make them suitable for different types of experimental protocols (Titz et al., 2018). This study aims to evaluate the performance of a recently released mobile eye-tracker, the Pupil Neon from Pupil Labs, by examining some of the most common eye-tracking-related parameters: fixations, saccades, smooth pursuit, pupil dilation, microsaccades, blinks, and the influence of head motion on accuracy (Duchowski, 2018). By conducting this independent comparative evaluation, we seek to provide researchers with information on the strengths and weaknesses of the Pupil Neon, facilitating its effective use in diverse research contexts.

Stationary and mobile eye-trackers

Two types of eye-tracking devices are usually distinguished: stationary (or desk/screen-mounted) eye-trackers, and mobile (or head-mounted) eye-trackers (Pentus et al., 2020).

Stationary eye-trackers are ideal for two-dimensional stimuli presented via screen-based tasks, making them traditionally popular in basic research where a controlled experimental setup is feasible (Holmqvist et al., 2011). These eye-trackers often have high accuracy and precision, potentially reaching up to 0.3 degrees under optimal conditions (Ehinger et al., 2019). However, achieving such performance comes at the cost of restricting participants in their head and body movements, lowering ecological validity (Holmqvist et al., 2011). Such setups often require a fixed sitting position or even head fixation via chinrest, limiting natural behaviour. Additionally, the highly controlled environment of lab experiments may not accurately represent real-life conditions, prompting the eye-tracking scientific community to seek tools that enable monitoring in real-world settings (Gunawardena et al., 2022; Takahashi et al., 2018).

Conversely, mobile head-mounted eye-trackers allow much more freedom in head and body movements by tracking directly from sensors located on the participant's head (e.g. glasses), making them a prior candidate for in-the-wild studies and applied research where it is necessary to move in an environment (Bulling & Gellersen, 2010). Notably, this refers to the contemporary mobile eye-trackers and not the first scleral coil eye-tracking devices that were directly mounted to the participant's eye (Huey, 1900). However, this freedom introduces challenges in tracking gaze accurately, resulting in noisier data and lower precision, typically around 0.9 to 1.8 degrees of visual angle (Baumann & Dierkes, 2023; MacInnes et al., 2018). Mobile eye-trackers also face technical issues such as device heating, which can affect user experience, limited battery life leading to restricted data collection duration, and the need for a stable wireless connection (Gunawardena et al., 2022). Despite these challenges, technological advancements are continuously improving the performance of mobile eye-trackers, necessitating regular updates on their capabilities. In the present study, we aim to assess the performance of one of the most recent mobile eye-tracking devices on the market.

Evaluating eye-tracker performances

Evaluating the performances of data recording devices is essential for any research field, as it allows the assessment of data quality and reliability. Understanding the capabilities and limitations of eye-trackers is essential in order to optimizing their utilization. While several studies have examined data quality from field eye-tracking experiments in various experimental contexts (Funke et al., 2016; Hooge et al., 2023; MacInnes et al., 2018; Niehorster et al., 2020) or using artificial eyes (Wang et al., 2017), the complexity and diversity of human eye movements should also be considered when measuring an eye-tracker's performances (Holmqvist et al., 2012). Estimating an eye-tracker's performance is challenging, as comparisons to a theoretical true value are not possible (Ehinger et al., 2019). When asking participants to fixate on a visual stimulus for calibration, the actual eye fixation point is not steady due to miniature, unconscious eye movements like drift and microsaccades, which can corrupt the recorded fixation baseline (Rolfs, 2009). To address this lack of a truth reference, earlier studies used two eye-trackers simultaneously to evaluate and compare their performances across a variety of tasks (Drewes et al., 2011; Ehinger et al., 2019; Titz et al., 2018): a reference and a target eye-tracker to be evaluated. Building on the study conducted by Ehinger et al. (2019), the current study uses the Eyelink 1000 (SR Research Ltd., 2022) as a reference eye-tracker due to its high precision and accuracy. It is considered one of the best video-based eye-trackers available (Holmqvist, 2017; Kaduk et al., 2023). Comparing a mobile eye-tracker to a stationary one in terms of gaze accuracy and precision may appear to be of limited value, given that these two types of eye-trackers often serve different purposes. The goal of such comparisons is not to favour one type of device over another, but rather to highlight the distinctive characteristics exhibited by each device when recording specific types of eye movements. Various types of eye movements, including changes in pupil size provide diverse information about visual and cognitive processing (Martinez-Conde et al., 2004; Rayner, 2009; Rayner, 1998). For example, fixations are essential for detailed visual processing and information acquisition, allowing the eyes to remain steady and to absorb information from a specific area of the visual field (Henderson, 2003). Saccades are rapid eye movements that reposition the fovea to a new location of interest and are critical for visual attention and scene perception (Rayner, 1998). Microsaccades however are tiny, involuntary eye movements that help in the fine-tuning of visual fixation and are linked to covert attention (Martinez-Conde et al., 2004; Martinez-Conde et al., 2013). Relative to saccades, smooth pursuits are characterized by slow eye movements to maintain a moving object on the fovea and are associated with tracking moving stimuli (Krauzlis, 2004). Eye blinks can indicate cognitive load and fatigue (Schleicher et al., 2008) and changes in pupil size are indicative of arousal and cognitive effort (Beatty & Lucero-Wagoner, 2000). Each type of eye-based measure has specific tracking requirements: the accuracy of fixations and saccades is impaired by head movements, particularly in free-viewing or or extreme head movement conditions (Einhäuser et al., 2007). Pupillometry also demands minimal head movement, a fixed stimulus position and steady brightness conditions (Mathôt & Vilotijević, 2023). The analysis of smooth pursuits however requires smooth stimuli velocity and a high temporal resolution to distinguish from saccades and microsaccades (Holmqvist et al., 2011); blink frequency is influenced by fatigue and experiment duration (Schleicher et al., 2008). To adequately evaluate an eye-tracker's performance, it is essential to consider more than just the accuracy and precision typically reported by manufacturers. To date, publicly available data are limited, and independent evaluations are even scarcer. To address this, Ehinger et al. (2019) developed a comprehensive evaluation paradigm, assessing fixation and saccade accuracy in grid and free-viewing tasks, accuracy decay over time, smooth pursuit, pupil dilation, microsaccades, blinks, and the influence of head motion. At the time of their evaluation, mobile eye-trackers such as the Pupil Core (Pupil Labs GmbH) predominantly recorded the eyes with infrared video-based methods and detected the pupil using common computer vision algorithms to track gaze. Instead of 'simple' computer vision approaches based on infrared eye-tracking, the newer Pupil Neon (Pupil Labs GmbH) uses a proprietary deep learning approach. It has the advantage that it is supposedly more flexible in terms of environmental context and does not require a calibration procedure. However, it has the disadvantage inherent to all deep learning approaches: we do not really know how it works and thus do not know whether it captures all types of eye movements equally well. Thus, this independent evaluation will benefit researchers intending to use the Pupil Neon by demonstrating the advantages and limitations of such eye-tracking technology before employing it in their studies.

Our study

Due to the rapidly increasing use of (mobile) eye-tracking devices in both academic and consumer research, it becomes more important that the increasing number of datasets is based on reliable recordings. Given the use case for mobile eye-tracking devices in certain research and consumer settings, a major factor influencing widespread adoption is a device's ease of use (Davis et al., 1989). This is our reason for choosing to evaluate the Pupil Neon over other mobile eye-tracking devices. To our knowledge, it is the only device that requires no calibration, significantly simplifying setup and reducing the time required for participants to begin tasks. Moreover, the Pupil Neon is one of the more affordable options available, with costs starting at €5,950 as of July 2024, making it accessible to a broader range of researchers and institutions. Recent manufacturer evaluations indicate that despite not having a calibration procedure, it performs comparably well with an accuracy of around 1.3° (Baumann & Dierkes, 2023). However, it employs a proprietary deep-learning algorithm for calibration-free classification of eye movements, which complicates performance evaluation based solely on available data and code. This study aims to provide an independent evaluation of the Pupil Neon's performance across various eye-based tasks. Following Ehinger et al. (2019) procedure, participants will perform a set of tasks while being tracked simultaneously by both the Pupil Neon and the EyeLink 1000. These tasks include fixations on a large grid to assess spatial accuracy, smooth pursuit tasks, free viewing tasks to evaluate eye movements and gaze trajectories, microsaccades tasks, blink tasks, pupil dilation tasks, fixations on a small grid to evaluate the decay of accuracy over time, head yaw movements, head roll movements, and fixations on a small grid after head movements to assess the decay of accuracy. The results will provide insights into the specific characteristics and performance of the Pupil Neon in recording various eye movements and performing different tasks. These findings will help identify tasks where the Pupil Neon excels and highlight tasks that might be less advisable to conduct with this device due to differing eye movement requirements.

Methods

The methodology employed in this study is largely consistent with that described by Ehinger et al. (2019).

Participants

We recruited *[tbd]* participants from Ulm University, with an average age of *[tbd]* years (range *[tbd]* -*[tbd]* years); *[tbd]* were female, *[tbd]* were left-handed, and *[tbd]* had a left-dominant eye. The inclusion criteria were: no use of glasses or hard contact lenses, no drug use, no history of photosensitive migraines or epilepsy, and at least 5 hours of sleep the night before the experiment. Written consent was obtained from all participants, and the study was declared exempt from ethical approval by the ethics committee of Ulm University (letter from 06.06.2024). Participants received compensation of either €12 or one course-credit per hour. *[tbd]* participants were excluded from the analysis since they exceeded the predetermined calibration accuracy limits of the EyeLink 1000.

Experimental setup and recording devices

The experimental setup and recording devices are largely similar to those employed by Ehinger et al. (2019), except for the use of the Pupil Neon glasses instead of the Pupil Core glasses. The description of the experimental setup and recording devices is adapted from Ehinger et al. (2019). The experiment took place in a light and soundproof laboratory at Ulm University. The lights were left on during the experimental procedure to ensure constant lighting conditions throughout the experiment. The original experimental code was written by Ehinger et al. (2019) in MATLAB (2016). In the present study, the code was adapted and programmed in MATLAB (2021) on a computer with Windows 10 OS using the Psychophysics Toolbox 3 (Brainard & Vision, 1997; Kleiner et al., 2007; Pelli, 1997), EyeLink Toolbox (Cornelissen et al., 2002), and custom scripts based on the ZMQ protocol for communication with the Pupil Neon. Stimuli were presented on an ASUS ROG SWIFT PG279QM screen (27 inch, 2560 × 1440 pix) running at 100 Hz. Stimuli were presented on a constant gray background, except for the pupil dilation task, in which different backgrounds were used to stimulate pupil dilation and constriction. The participants were seated at a distance of 60 cm from the screen, at which the display subtends *[tbd]*° x *[tbd]*° of visual angle. Two Logitech Multimedia Speakers Z200 emitted a 300 Hz sound for the auditory stimuli.

Participants' eye movements were simultaneously recorded using one stationary and one mobile eye-tracking device. The desktop-mounted EyeLink 1000 Plus (SR Research Ltd.) recorded monocular movements of the dominant eye at 1000 Hz in head-free mode (Ehinger et al., 2019, cf.). Concurrently, the Pupil Labs Neon glasses (Pupil Labs GmbH.) recorded binocular eye movements. The Pupil Labs Neon glasses include a scene camera (1600 × 1200 pixels at 30 Hz, 132° horizontal and 81° vertical field of view) and two infrared eye cameras (192 × 192 pixels at 200 Hz). These glasses feature real-time neural network technology, providing binocular eye tracking without the need for calibration, and employ deep learning for slippage compensation. Data were captured using the Neon Companion device and pre-processed post-hoc via Pupil Cloud (see Data Analysis section). The glasses also include an inertial measurement unit (IMU) comprising an accelerometer, magnetometer, and gyroscope, along with dual microphones. The experiment used two computers in addition to the Companion device: one for stimuli presentation and one for recording the EyeLink 1000. Experimental messages ("triggers") were sent to the EyeLink 1000 recording computer via the EyeLink Toolbox (Cornelissen et al., 2002), and to the Pupil Labs glasses using zeroMQ packages ("ZeroMQ," 2024). To synchronize the recordings, concurrent trigger signals were sent via Ethernet during experimental events.

Experimental Procedure

The experimental procedure is similar to the one described by Ehinger et al. (2019), from which this subsection is adapted.

Each session began with a brief oral instruction on the experimental procedure and tasks. Then, participants' visual acuity was checked using a calibrated online LogMar chart test with a single test line of five letters. A correct identification of 6/6 was required to proceed with the experiment. Afterwards, Ocular dominance was determined using the "hole-in-card" test with participants' hands and a centered gaze.

The experiment comprised six blocks, each consisting of 10 tasks (see Figure 1), presented in a fixed sequence. Eye-tracker calibration was performed at the beginning of each block. Afterwards, participants completed a grid task (large grid) designed to assess the spatial accuracy of the eye-trackers. Afterwards, participants performed several tasks without head movements comprising smooth pursuit, free viewing, microsaccades, blinks and pupil dilation. Afterwards, the small grid task was performed. Then, participants performed two tasks requiring head movements, namely head yaw and head roll. Half of the participants started with the head yaw task, the other half with the head roll. Task order was balanced between participants. At the end of each block, the small grid task was performed again. Hence, tasks requiring intense fixation (microsaccade and pupil dilation) were interspersed with more relaxing tasks (blinks and free viewing accuracy) to provide participants with periodic breaks. Participants read written instructions prior to each task and saw a green fixation target at the center of the monitor. Further, the experimenter stressed the importance of focusing on the fixation targets before starting the task. Participants initiated each task at their own pace by pressing the space bar. The experimental session lasted approximately *[tbd]* minutes.

Tasks

We used the tasks and code implementation developed by Ehinger et al. (2019), from which the task descriptions are adapted.

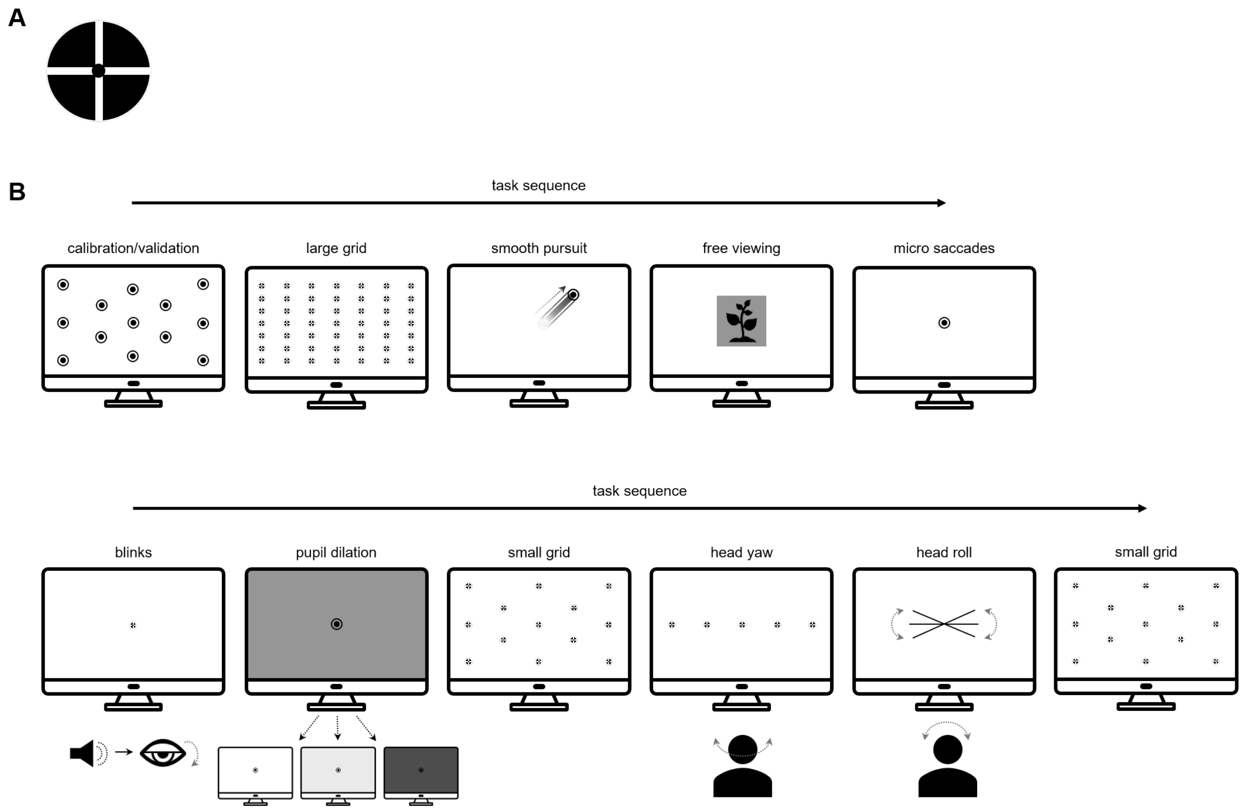

Figure 1

(A) Fixation cross used in the large and small grid tasks, blink task, and the head yaw task. (B) This figure illustrates the task sequence within each experimental block. Adapted from "A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000" by B. V. Ehinger, K. Groß, I. Ibs, & P. König, 2019, PeerJ, 7:e7086 (https://doi.org/10.7717/peerj.7086).

Fixation targets

Throughout the experiment, we used three different fixation targets. For the EyeLink calibration we used the manufacturers calibration targets. For the large and small grid task, blink task, head yaw task, and head roll task a fixation cross was used that has been shown to reduce miniature eye movements (Thaler et al., 2013). It was composed of a 1.5 x 1.5° black disc, superimposed by a white cross (1.5 x 1.5°, linewidth 0.2°) and a smaller black disc (0.2 x 0.2°). The fixation cross is depicted in Figure 1. For the smooth pursuit task, microsaccade task and pupil dilation task, we used a bullseye target (outer circle: black, diameter 0.5°; inner circle: white, diameter 0.25°). For the smooth pursuit task, the bullseye was used due to its aesthetically pleasing diagonal movement. For the microsaccade task, the bullseye was used since minimization of microsaccades was not desired. For the pupil dilation task, the bullseye was used due to its visibility regardless of background illumination.

Eye-tracker calibration

The EyeLink 1000 was calibrated using a 13-point randomized calibration procedure. These 13 calibration points were selected from the large grid used in the accuracy task (see section "Task 1/Task 7/Task 10: Accuracy Task with the Large and the Small Grid"). Calibration points were manually advanced by the experimenter. Following calibration, a 13-point verification process was conducted. The procedure was identical to the initial calibration, yet calibration points were presented within a new randomized sequence. Accuracies were calculated online, and recalibration was performed if necessary until the mean validation accuracies met the manufacturers' recommendations. The EyeLink 1000 required a mean validation accuracy limit of 0.5°, with individual points not exceeding 1° (SR Research Ltd., 2010). If more than 10 calibration attempts failed, despite adjustments to the EyeLink 1000, the recording session was terminated and the participant was excluded from the experiment. The Pupil Labs Neon glasses are calibration-free devices and were not calibrated. However, a personal gaze offset correction was performed for each participant to maximize Neon's accuracy. This gaze offset correction is a linear adjustment applied uniformly across the field of view to the gaze estimation. Thus, it doesn't vary at different eccentricities and will correct for general offsets across the whole visual field. This offset correction was achieved on Pupil Cloud according to the procedure described by Pupil Labs, which consists in fixating a single target at the center of the screen. If the gaze circle from the raw Neon's gaze estimate does not fit the target location, the gaze circle is manually dragged onto the center of the target. The fixation point used for this offset correction was the last central fixation point from the validation procedure.

Task 1/Task 7/Task 10: Accuracy on Large and Small Grids

We used fixation grids to assess the difference between the displayed target location and the estimated gaze point, estimating absolute spatial accuracy and calibration accuracy decay over time. Task 10 is additionally monitoring the influence of head movements on accuracy decay. Two variants were employed: a large grid (7×7) and a small grid (a subset of 13 points). For the grid tasks, fixation cross targets were used. For the large grid, participants fixated on targets at 49 crossing points, equally spaced from -7.7 to 7.7° vertically and -18.2 to 18.2° horizontally. Each target appeared once per task repetition, and participants pressed the space bar after saccading to and fixating on each target. The center point served as both the start and end points. We used a constrained randomization procedure for the large grid to ensure uniform saccade amplitude and angle distributions, maximizing the entropy of the saccade amplitude and angle histograms. The small grid task was similar but involved only a subset of 13 target points that were also used in the calibration procedure. The stimulus sequence was naively randomized within each block for each participant.

Task 2: Smooth pursuits

Bullseye targets were used for the smooth pursuit task. We used Ehinger et al. (2019)'s adaptation of the step-ramp smooth pursuit paradigm from Liston and Stone (2014) to investigate smooth pursuits. Participants fixated on a central bullseye target and pressed the space bar to start a trial, with the probe starting after a random delay sampled from an exponential function (mean 500 ms). The stimuli moved along linear trajectories at one of five speeds (16, 18, 20, 22, 24°/s) and trials ended when the target was 10° from the center. We used 24 different orientations spanning 360°, starting each stimulus such that it took 0.2 seconds to reach the center, minimizing catch-up saccades. Each smooth pursuit task consisted of 20 trials, with a total of 120 trials per experiment. Each participant encountered all possible combinations of speed and angle once, randomized throughout the experiment. Participants were instructed to follow the target with their eyes as long as possible.

Task 3: Free viewing

For the free viewing task, participants were presented with a total of 18 different natural images, primarily patterns from Backhaus (2016). Each of the six blocks comprised three randomly chosen images. The image order was randomized across the experiment, and each image was shown once only. In the beginning of each trial, a fixation cross target was presented at the screen center for an average of 0.9 seconds with a random jitter of 0.2 seconds. Afterwards, an image (900 × 720 pixels) was displayed for 6 seconds. Participants were instructed to explore the images freely.

Task 4: Microsaccades

To elicit microsaccades, a central bullseye fixation target was displayed for 20 seconds, with participants instructed to maintain fixation until the target disappeared.

Task 5: Blinks

For the blink task, a fixation cross target was used. Participants fixated on a central target and were instructed to blink each time they heard a 300 Hz sound for 100 ms. In each block the sound chimed seven times with 1.5-second pauses between sounds. Each sound onset was jittered by ±0.2 seconds to reduce predictability.

Task 6: Pupil Dilation

For the pupil dilation task, bullseye targets were used. To stimulate pupil size changes, we varied the monitor's light intensity while participants fixated on a bullseye target presented in the screen center. Each block consisted of four trials with a different luminance level (25%, 50%, 75%, and 100%). The order of the bright stimuli was randomized within each block. At the beginning of each trial, a black screen was displayed for 7 seconds (jittered by ±0.25 seconds) to allow the pupil to reach its largest size. Afterwards, one of the four target luminances was displayed for 3 seconds (jittered by

±0.25 seconds).

Task 8/9: Head Movements

For the head movement tasks, fixation cross targets were used. For the roll movement task participants tilted their heads to align their eyes with a rotated line displayed at seven different angles (-15°, -10°, -5°, 0° (horizontal), 5°, 10°, or 15° of visual angle). They pressed the space bar once their eyes were in line with the target to proceed to the next orientation.

For the yaw movement task, participants completed 15 head rotations to fixate on targets positioned horizontally at five locations (-17.6°, -8.8°, 0°, 8.8°, or 17.6° of eccentricity). They rotated their heads to align their noses with the target, fixated on it, and pressed the space bar to confirm. The target positions were randomized within each block.

Data analysis

Our data analysis follows the modular pipeline outlined by Ehinger et al. (2019), from which the following subsections are adapted. Data analysis was performed using Python 3 (Van Rossum & Drake, 2009), pyEDFread (Wilming et al., 2024), NumPy (Harris et al., 2020), pandas (McKinney, 2010), and SciPy (Virtanen et al., 2020). Visualization was done using plotnine (plotnine development team, 2024) and Matplotlib (Hunter, 2007). Experimental code, data and data analysis code are available under *[tbd]*. Citations, Data Transparency, Analytic Methods (Code), Research Materials, Design and Analysis adhere to the Transparency and Openness Promotion (TOP) Guidelines (Nosek et al., 2015) endorsed by the American Psychological Association. The present study did not test specific hypotheses; rather, we focussed on an exploratory data analysis approach to compare various gaze parameters between both eye-tracking devices. Data analysis for the respective gaze parameters are described in detail below.

Preprocessing

Data Export and Transformation: The raw EyeLink 1000 gaze data were exported using the EyeLink Data Viewer software and transformed into dataframes, which include calibrated gaze data mapped to the monitor coordinates. The Pupil Neon eye-tracking data were automatically sent to the Pupil Labs cloud after each recording session. Notably, there is no explicit calibration procedure for the Pupil Neon. In the cloud, each recording is associated with the video from the scene camera and saved in a workspace. After attributing the recording to a project, the data can be normalized from head coordinates to world coordinates using the "Marker Mapper Enrichment" (see Coordinate System Conversion section). The gaze data in normalized coordinates associated with the recording time range of interest was then exported from the cloud for local eye movement analysis.

Coordinate System Conversion: Since the Pupil Neon is a mobile eye-tracker (head coordinate frame) and the EyeLink 1000 is a desktop eye-tracker (world coordinate frame), the initial step involved converting both datasets to the same coordinate system. Four QR markers were placed at the corners of the monitor to detect the display. These markers were detected by the Pupil Neon scene camera and used to create a new world coordinate frame. This conversion was performed directly from the Pupil Labs Cloud using the "Marker Mapper Enrichment" feature. The gaze data in the world coordinate frame were then exported as a dataframe, with the bottom left corner of the screen as the frame origin.

Gaze data synchronization: Trigger messages were sent during the experiment to mark task events. To ensure synchronized gaze information from both eye-trackers, a trigger with the computer's timestamp was sent at the beginning of the recording phase before the first calibration to both devices. Gaze data from both eye-trackers were synchronized by matching the recording start timestamps. If time drifts were detected between recordings, synchronization was adjusted by estimating the slope difference for each event trigger.

Data Cleaning: Samples marked as corrupted or where no pupil was detected were excluded from further analysis, as the ones where the gaze point was outside the monitor area since the experiment was performed on the screen. During this data cleaning phase, *[tbd]* % of the data was removed for Eyelink 1000 (*[tbd]* samples), and *[tbd]* % for the Pupil Neon (*[tbd]* samples).

Eye Movement Classification

Eye movements were defined and classified across both datasets using an updated version of Ehinger et al. (2019) algorithmic pipeline, which applies identical algorithms to both eye-trackers wherever possible. This approach ensured consistency in the comparison of devices. Finally, the gaze position, eye movements, and pupil diameter associated with each task were compared concurrently between the two eye-trackers to evaluate their performance and consistency.

Blink Classification: Blink classification differed between the two eye-trackers. The EyeLink 1000 reports blinks when the pupil is missing for several samples. The thresholds for minimum blink duration classification can be accessed and modified. In our study, binks were defined by missing data for at least 100ms. In contrast, the Pupil Neon uses ML signal reconstruction for classification, meaning there are no missing samples (Pupil Labs blink detector, algorithm description 31.10.23). For this reason, a similar blink classification pipeline was not possible, leading us to use the proprietary algorithms. In the Pupil Neon, a machine-learning model is trained on the eye-camera video to classify eyelid opening, eyelid closing, or neither eyelid opening nor closing (see their algorithm description for more details on the parameters). After each frame is labelled, a post-processing procedure defines the eye blinks using the temporal sequence of the eyelid events. Especially, each blink is defined by onset and offset and a minimum blink duration of 100 ms. The samples associated with the 100 ms before and after a blink event were also marked as blink samples (Costela et al., 2014) and were not considered for subsequent analysis from sample data, such as saccade classification.

Saccade Classification: Saccades were classified using the velocity profile of eye movements to extract saccades, following the methods of Engbert and Kliegl (2003) and Engbert and Mergenthaler (2006). The algorithm was derived from Ehinger et al. (2019) pipeline with the hyperparameter lambda adjusted to a value of 5 for saccades classification. Unlike Ehinger et al. (2019) method, we did not interpolate the samples since the EyeLink 1000 and the Pupil Neon had constant sampling rates of 1000 Hz and 200 Hz, respectively. This classifier was applied to the sample data.

Note: After personal communication with B. Ehinger, the saccade classification pipeline will be updated from Engbert-Mergenthaler to REMoDNaV algorithm, which still uses the velocity profile of eye movements to extract saccade. The filter settings will be optimized by systematically adjusting parameters to minimize false positives in saccade detection and improve the accuracy of fixation identification, starting from REMoDNaV default values.

Fixation Classification: Samples not classified as blinks or saccades were labelled as fixations. Fixations shorter than 50 ms were removed from the dataset. Since an evaluation of the fixations classification is beyond the scope of the present study, we decided to focus on the performance comparison between devices while acknowledging that the eye movement classification described in this study is not optimized for mobile eye-trackers due to head movements. However, we should make it clear that the tasks do not include any moving objects, and the participants' heads were generally still despite no headrest restrictions. Following Ehinger et al. (2019) analysis pipeline, the analysis of the gaze data during the Head Movement task was only performed after the movement but not during.

Smooth Pursuits classification: An exception to this eye movements classification was the Smooth Pursuit task, in which smooth pursuits were defined by gaze movements with similar direction and velocity as the moving target. Please see "Task 2: Smooth pursuits" for further details.

Pupil Size: The Eyelink 1000 computes the pupil size by counting the number of pixels that are detected inside the pupil ellipse boundaries. Thus, the pupil size is given in area. The Pupil Neon uses a deep learning algorithm (referred to as NeonNet) to compute from the eye videos, and for each eye separately, a 3D model of the eyeball from which the pupil sizes in diameter (mm) are extracted (Pfeffer & Dierkes, 2024). The accuracy of the pupil diameter measurements is also improved by specifying the user's inter-eye distance in the user's profile before the recording, which we did. The pupil diameter reported in the 3D eye-state measurements was converted into pupil area using $A = \frac{1}{4} \cdot \pi \cdot l_1 \cdot l_2$ where A denotes the ellipsis area, l_1 denotes the semi-major axis and l_2 denotes the semi-minor axis. The pupil size was then normalised to the median of a baseline period before the bright stimulus onset, accounting for fluctuations due to attention or alertness.

Note: The 3D eye-state measurements from Pupil Labs currently give us the pupil diameter D. However, we do not know yet if we can access the two ellipse axes l_1 and l_2 directly (ongoing communication with Pupil Labs). If we can access l_1 and l_2, then the pupil size will be converted into pupil area as described above, before being baseline-normalized. If we cannot access l_1 and l_2, then we will standardize the pupil size from both eye-trackers using the z-score, and then perform the baseline normalization.

Measures of Gaze Data Quality

Spatial Accuracy: Spatial accuracy refers to the distance between the measured gaze point and the target position (Holmqvist et al., 2012). It should be noted that the actual gaze point might differ from the target position (e.g. due to misalignment of the fovea despite the subjective direction of gaze towards the target), but we consider here the target position as a proxy for the actual gaze point. This distance is often expressed by an angular difference which can be computed by the cosine between the mean gaze point vector and the target location vector. The vectors were converted from the Spherical coordinate system to the Cartesian system to compute the cosine distance which results in an angular difference between 0 and 180 degrees. The accuracy was monitored by first calculating the 20% winsorized mean angular difference between the estimated gaze point and the target location for each participant over blocks, and then reporting the 20% winsorized mean and the interquartile range (IQR) over the already averaged values for both eye-trackers. Participants may make small eye movements during fixations or catch-up saccades for the ones with large amplitude, which can especially happen during the calibration or the grid tasks. In such cases, multiple candidates could be considered to attribute fixations' coordinates. Similarly to Ehinger et al. (2019) method, we decided to select the last ongoing fixation that happened just before participants pressed the space bar.

Spatial Precision: Spatial precision refers to the variability in gaze coordinate estimations, reflecting the noise in the data. The less dispersed the estimations are, the better the spatial precision. The measure of the spatial precision was assessed in two ways: by the root mean squared (RMS) of inter-sample distances and by the standard deviation (SD) of the sample locations, respectively monitoring the proximity of consecutive samples and the spatial spread (see Ehinger et al. (2019) for a more detailed description). The fixation spread was monitored by first calculating the 20% winsorized mean SD and RMS for each participant over blocks, and then reporting the 20% winsorized mean and the interquartile range (IQR) over the already averaged values for both eye-trackers.

Task-specific Analyses

Task 1/Task 7/Task 10: Accuracy on Large and Small Grids. Spatial accuracy was evaluated by computing winsorized means on the offset between the displayed target and the mean gaze position of the last fixation before the new target appeared, and spatial precision was assessed by computing winsorized means on RMS and SD measures (see Spatial Accuracy and Spatial Precision sections). The mean difference in accuracy between the two eye-trackers was assessed using the 95% bootstrap confidence interval (95% CI). Spatial accuracy was compared between two groups of points - the center ones and the edge ones - in order to evaluate the impact of target distance-from-center on eye-trackers performances. Spatial accuracy was also measured at multiple time points to evaluate accuracy decay: with no decay (directly after initial calibration), after some temporal drift (2/3 of the block elapsed), and after provoked head movements (yaw and roll task). The decay of accuracy over time was evaluated using a robust linear mixed effects model with conservative Wald's t-test p-value calculation to account for outliers. Following Ehinger et al. (2019) recommendations, the model was defined by LMMaccuracy $\sim 1 + \text{et session } (1 + \text{et session} \mid \text{subject} \setminus \text{block})$ and evaluated with the robustlmm R package (Koller, 2016).

Task 2: Smooth pursuits. To analyze smooth pursuit onsets and velocities, we generalized the Liston and Stone (2014) model to a Bayesian framework using STAN. The x-y gaze coordinates of each trial were rotated to align with the target direction, fitting data up to the first saccade exceeding 1° or up to 600 ms after trial onset. We used a restricted piece-wise linear regression with a logistic transfer function for the hinge, assuming normal noise. The analysis relied on classifying initial saccades accurately. Then the smooth pursuit detection was monitored by first calculating the mean posterior value of the hinge-point and velocity parameter for each trial, and then reporting the 20% winsorized mean and the interquartile range over blocks and subjects for both eye-trackers. The mean difference in smooth pursuit onsets and velocities between the two eye-trackers was assessed using the 95% bootstrap confidence interval (95% CI). Additionally, we recorded the number of saccades during target movement to control for sampling rate bias.

Task 3: Free viewing. The free-viewing task was analysed by first calculating the 20% winsorized mean fixation number, fixation durations, and saccadic amplitudes for each participant over blocks, and then reporting the 20% winsorized mean and the interquartile range over the already averaged values for both eye-trackers. The mean difference in fixation number, fixation durations, and saccadic amplitudes between the two eye-trackers was assessed using the 95% bootstrap confidence interval (95% CI). Additionally, we visually compared gaze trajectories to assess the spatial inaccuracies. The first fixation on the cross was excluded, and we smoothed a pixel-wise 2D histogram with a Gaussian kernel (SD = 3°) to analyze central fixation bias.

Task 4: Microsaccades. The microsaccades detection was monitored by first calculating the 20% winsorized mean microsaccades number and amplitudes for each participant over blocks, and then reporting the 20% winsorized mean and the interquartile range over the already averaged values for both eye-trackers. The mean difference in microsaccades number and amplitudes between the two eye-trackers was assessed using the 95% bootstrap confidence interval (95% CI). Additionally, we visually compared the main sequences using the Engbert and Mergenthaler (2006) algorithm specifically for each block to assess the variance of reported microsaccades.

Task 5: Blinks. The blink detection was monitored by first calculating the 20% winsorized mean blink number and durations for each participant over blocks, and then reporting the 20% winsorized mean and the interquartile range over the already averaged values for both eye-trackers, noting the use of different blink classification algorithms (see section "Eye Movement Definition and Classification"). The mean difference in blink number and durations between the two eye-trackers was assessed using the 95% bootstrap confidence interval (95% CI).

Task 6: Pupil Dilation. We analyzed the relative pupil areas for each luminance. The normalized pupil response was calculated by dividing the pupil signal by the median baseline pupil size before the bright stimulus onset. This adjustment was necessary due to variations in the baseline levels, indicating potential influences such as attentional processes or camera distance. The normalized pupil area is reported as a percent change from the median baseline. Then the measurement of the pupil size was monitored by first calculating the 20% winsorized mean normalized pupil area between 2s and 3s after luminance change for each participant over blocks and luminance levels, and then reporting the 20% winsorized mean and the interquartile range over the already averaged values for each luminance level for both eye-trackers. The mean difference in pupil areas between the two eye-trackers was assessed using the 95% bootstrap confidence interval (95% CI).

Task 8/9: Head Movements. For the roll movement task, the accuracy decay was monitored by first calculating the 20% winsorized mean gaze position 0.5 seconds before the button press for each participant over blocks, and then reporting the 20% winsorized mean and the interquartile range over the already averaged values for both eye-trackers. The gaze position was taken 0.5 seconds before the button press due to continuous fixation on the center of the line during the head movement which led to no new fixation detected.

For the yaw movement task, the accuracy decay was monitored by first calculating the 20% winsorized mean gaze position at the final fixation before the participants confirmed their yaw movement for each participant over blocks, and then reporting the 20% winsorized mean and the interquartile range over the already averaged values for both eye-trackers. For both roll and yaw tasks, the mean difference in accuracy between the two eye-trackers was assessed using the 95% bootstrap confidence interval (95% CI).

Acknowledgements

This project has received funding from the European Union's Horizon Europe research and innovation funding program under grant agreement No 101072410 and the Graduate and Professional Training Center Ulm's Early Career Incubator program.

Funded by
the European Union

References

Ahlström, C., Kircher, K., Nyström, M., & Wolfe, B. (2021). Eye tracking in driver attention research—how gaze data interpretations influence what we learn. *Frontiers in Neuroergonomics*, *2*. https://doi.org/10.3389/fnrgo.2021.778043

Backhaus, D. (2016, July). *Mobiles eye-tracking: Vergleichende evaluation einer mobilen eye-tracking brille* [Doctoral dissertation, Potsdam University]. https://doi.org/10.13140/RG.2.2.35488.60165

Baumann, C., & Dierkes, K. (2023). *Neon accuracy test report* (tech. rep.). Pupil Labs.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, *10*(4), 433–436.

Bulling, A., & Gellersen, H. (2010). Toward mobile eye-based human-computer interaction. *IEEE Pervasive Computing*, *9*(4), 8–12. https://doi.org/10.1109/MPRV.2010.86

Cornelissen, F., Peters, E., & Palmer, J. (2002). The eyelink toolbox: Eye tracking with matlab and the psychophysics toolbox. *Behavior research methods instruments & computers*, *34*(4), 613–617.

Costela, F. M., Otero-Millan, J., McCamy, M. B., Macknik, S. L., Troncoso, X. G., Jazi, A. N., Crook, S. M., & Martinez-Conde, S. (2014). Fixational Eye Movement Correction of Blink-Induced Gaze Position Errors [Publisher: Public Library of Science]. *PLOS ONE*, *9*(10), e110889. https://doi.org/10.1371/journal.pone.0110889

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management science*, *35*(8), 982–1003.

Drewes, J., Montagnini, A., & Masson, G. S. (2011). Effects of pupil size on recorded gaze position: A live comparison of two eyetracking systems. *Journal of Vision*, *11*(11), 494. https://doi.org/10.1167/11.11.494

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice*. Springer.

Duchowski, A. T. (2018). Gaze-based interaction: A 30 year retrospective. *Computers & Graphics*, *73*, 59–69. https://doi.org/https://doi.org/10.1016/j.cag.2018.04.002

Ehinger, B. V., Groß, K., Ibs, I., & König, P. (2019). A new comprehensive eye-tracking test battery concurrently evaluating the pupil labs glasses and the eyelink 1000. *PeerJ*, *7*, e7086.

Einhäuser, W., Schumann, F., Bardins, S., Bartl, K., Böning, G., Schneider, E., & König, P. (2007). Human eye-head co-ordination in natural exploration. *Network: Computation in Neural Systems*, *18*(3), 267–297. https://doi.org/10.1080/09548980701671094

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

Engbert, R., & Mergenthaler, K. (2006). Microsaccades are triggered by low retinal image slip [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *103*(18), 7192–7197. https://doi.org/10.1073/pnas.0509557103

Funke, G., Greenlee, E., Carter, M., Dukes, A., Brown, R., & Menke, L. (2016). Which eye tracker is right for your research? performance evaluation of several cost variant eye trackers. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 1240–1244. https://doi.org/10.1177/1541931213601289

Gunawardena, N., Ginige, J. A., & Javadi, B. (2022). Eye-tracking technologies in mobile devices using edge computing: A systematic review. *ACM Computing Surveys*, *55*(8), 1–33.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in cognitive sciences*, *7*(11), 498–504.

Holmqvist, K. (2017). Common predictors of accuracy, precision and data loss in 12 eye-trackers. *The 7th Scandinavian Workshop on Eye Tracking*.

Holmqvist, K., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye tracker data quality: What it is and how to measure it. *Proceedings of the Symposium on Eye Tracking Research and Applications*, 45–52. https://doi.org/10.1145/2168556.2168563

Hooge, I. T. C., Niehorster, D. C., Hessels, R. S., et al. (2023). How robust are wearable eye trackers to slow and fast head and body movements? *Behavior Research Methods*, *55*, 4128–4142. https://doi.org/10.3758/s13428-022-02010-3

Huey, E. B. (1900). On the psychology and physiology of reading. i. *The American Journal of Psychology*, *11*(3), 283–302.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Kaduk, T., Goeke, C., Finger, H., et al. (2023). Webcam eye tracking close to laboratory standards: Comparing a new webcam-based system and the eyelink 1000. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02237-8

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in psychtoolbox-3? *Perception*.

Koller, M. (2016). Robustlmm: An r package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, *75*(6), 1–24. https://doi.org/10.18637/jss.v075.i06

Krauzlis, R. J. (2004). Recasting the smooth pursuit eye movement system. *Journal of neurophysiology*, *91*(2), 591–603.

Liston, D. B., & Stone, L. S. (2014). Oculometric assessment of dynamic visual processing. *Journal of Vision*, *14*(14), 12–12. https://doi.org/10.1167/14.14.12

MacInnes, J. J., Iqbal, S., Pearson, J., & Johnson, E. N. (2018). Wearable eye-tracking for research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices. *bioRxiv*, 299925.

Majaranta, P., & Bulling, A. (2014). Eye tracking and eye-based human–computer interaction. In S. Fairclough & K. Gilleade (Eds.), *Advances in physiological computing*. Springer. https://doi.org/10.1007/978-1-4471-6392-3_3

Mancini, M., Cherubino, P., Cartocci, G., Martinez, A., Di Flumeri, G., Petruzzellis, L., Cimini, M., Aricò, P., Trettel, A., & Babiloni, F. (2022). Esports and visual attention: Evaluating in-game advertising through eye-tracking during the game viewing experience. *Brain Sciences*, *12*(10). https://doi.org/10.3390/brainsci12101345

Martinez-Conde, S., Macknik, S., & Hubel, D. (2004). The role of fixational eye movements in visual perception. *Nature Reviews Neuroscience*, *5*, 229–240. https://doi.org/10.1038/nrn1348

Martinez-Conde, S., Otero-Millan, J., & Macknik, S. L. (2013). The impact of microsaccades on vision: Towards a unified theory of saccadic function. *Nature Reviews Neuroscience*, *14*(2), 83–96.

Mathôt, S., & Vilotijević, A. (2023). Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behavior Research Methods*, *55*(6), 3055–3077. https://doi.org/10.3758/s13428-022-01957-7

MATLAB. (2016). *R2016b*. The MathWorks Inc.

MATLAB. (2021). *R2021a*. The MathWorks Inc.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Niehorster, D. C., Santini, T., Hessels, R. S., et al. (2020). The impact of slippage on the data quality of head-worn eye trackers. *Behavior Research Methods*, *52*, 1140–1160. https://doi.org/10.3758/s13428-019-01307-0

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, *10*(4), 437–442.

Pentus, K., Ploom, K., Mehine, T., Koiv, M., Tempel, A., & Kuusik, A. (2020). Mobile and stationary eye tracking comparison–package design and in-store results. *Journal of Consumer Marketing*, *37*(3), 259–269.

Pfeffer, T., & Dierkes, K. (2024). *Neon pupillometry test report* (tech. rep.). Pupil Labs. https://doi.org/10.5281/zenodo.10057185

plotnine development team, T. (2024). Plotnine: A grammar of graphics for python. https://doi.org/https://doi.org/10.5281/zenodo.1325308

Rayner, K. (2009). The 35th sir frederick bartlett lecture: Eye movements and attention in reading, scene perception, and visual search. *Quarterly Journal of Experimental Psychology*, *62*(8), 1457–1506.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, *124*(3), 372.

Rolfs, M. (2009). Microsaccades: Small steps on a long way. *Vision Research*, *49*(20), 2415–2441. https://doi.org/10.1016/j.visres.2009.08.010

Schleicher, R., Galley, N., Briest, S., & Galley, L. (2008). Blinks and saccades as indicators of fatigue in sleepiness warnings: Looking tired? *Ergonomics*, *51*(7), 982–1010. https://doi.org/10.1080/00140130701817062

SR Research Ltd. (2022). *Eyelink data viewer*.

SR Research Ltd., S. (2010). Eyelink 1000 user manual. https://www.sr-research.com/support-options/learning-resources/

Takahashi, R., Suzuki, H., Chew, J. Y., Ohtake, Y., Nagai, Y., & Ohtomi, K. (2018). A system for three-dimensional gaze fixation analysis using eye tracking glasses. *Journal of Computational Design and Engineering*, *5*(4), 449–457. https://doi.org/10.1016/j.jcde.2017.12.007

Thaler, L., Schütz, A., Goodale, M., & Gegenfurtner, K. (2013). What is the best fixation target? the effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42. https://doi.org/https://doi.org/10.1016/j.visres.2012.10.012

Titz, J., Scholz, A., & Sedlmeier, P. (2018). Comparing eye trackers by correlating their eye-metric data. *Behavior research methods*, *50*(5), 1853–1863. https://doi.org/10.3758/s13428-017-0954-y

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CreateSpace.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, D., Mulvey, F. B., Pelz, J. B., et al. (2017). A study of artificial eyes for the measurement of precision in eye-trackers. *Behavior Research Methods*, *49*, 947–959. https://doi.org/10.3758/s13428-016-0755-8

Wilming, N., Morton, N., & Ehinger, B. (2024). Github repository of pyedfread. https://github.com/s-ccs/pyedfread/tree/master

Zeromq. (2024).

Question	Hypothesis	Sampling plan	Analysis Plan	Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes
How accurate and precise are the Pupil Neon recordings when compared to the Eyelink 1000?	-	For logistical lab reasons, participants will be recruited in a time window of 2 weeks. We take however many we can get within that time with a minimum of 25 participants (cf. Ehinger, 2019).		-	-	Eyelink 1000 is the gold standard for eye-tracking measurements.
Accuracy tasks (large & small grid tasks)	-		Spatial accuracy is evaluated by computing the 20% winsorized mean (WS) offset between the displayed target and the mean gaze position of the last fixation before the new target appeared and its interquartile range (IQR), and spatial precision was assessed by computing WS on RMS of inter-sample distances and SD measures of sample locations and its IQR. The mean difference between the eye-trackers is assessed using the 95% bootstrap confidence interval (95% CI). The decay of accuracy over time was evaluated using a robust linear mixed effects model with conservative Wald's t-test p-value calculation to account for outliers.	-	The more significant is the LMM, the stronger is the decay of accuracy over time. Determine if Pupil Labs Neon and Eyelink 1000 accuracy and precision are different.	
Smooth pursuit task	-		First, calculate the mean posterior value of the hinge-point and velocity parameter for each trial, and then report the 20% WS and the IQR over blocks and subjects. The mean difference between the eye-trackers is assessed using the 95% CI. Number of saccades recorded during target movement to control for sampling rate bias.	-	Determine if Pupil Labs Neon and Eyelink 1000 smooth pursuit detection are different.	

Free viewing task	-		First calculate the 20% WS mean fixation number, fixation durations, and saccadic amplitudes for each participant, and then report the 20% WS and the IQR over the averaged values. The mean difference between the eye-trackers is assessed using the 95% CI. Visual comparison of gaze trajectories to assess the spatial inaccuracies.	-	Determine if Pupil Labs Neon and Eyelink 1000 fixation and saccade detection are different.	
Microsaccade task	-		First, calculate the 20% WS microsaccades number and amplitudes for each participant, and then report the 20% WS and the IQR over the averaged values. The mean difference between the eye-trackers is assessed using the 95% CI. Visual comparison of the main sequences using the Engbert (2006) algorithm to assess the variance of reported microsaccades.	-	Determine if Pupil Labs Neon and Eyelink 1000 microsaccade detection are different.	
Blinks task	-		First, calculate the 20% WS blink number and durations for each participant, and then report the 20% WS and the IQR over the averaged values. The mean difference between the eye-trackers is assessed using the 95% CI.	-	Determine if Pupil Labs Neon and Eyelink 1000 blink detection are different.	
Pupil dilation task	-		First, calculate the 20% WS pupil area between 2s and 3s after luminance change for each participant, and then report the 20% WS and the IQR over the averaged values for each luminance level. The mean difference between the eye-trackers is assessed using the 95% CI.	-	Determine if Pupil Labs Neon and Eyelink 1000 measurements of the pupil size are different.	
Head rolls task	-		First, calculate the 20% WS gaze position 0.5 seconds before the button press for each	-	Determine if Pupil Labs Neon	

			participant, and then report the 20% WS and the IQR over the averaged values. The mean difference between the eye-trackers is assessed using the 95% CI.		and Eyelink 1000 accuracy after roll movements are different.	
Head yaws task	-		First, calculate the 20% WS gaze position at the final fixation before the participants confirmed their yaw movement for each participant, and then report the 20% WS and the IQR over the averaged values. The mean difference between the eye-trackers is assessed using the 95% CI.	-	Determine if Pupil Labs Neon and Eyelink 1000 accuracy after yaw movements are different.	

Guidance Notes

- **Question**: articulate each research question being addressed in one sentence.
- **Hypothesis**: where applicable, a prediction arising from the research question, stated in terms of specific variables rather than concepts. Where the testability of one or more hypotheses depends on the verification of auxiliary assumptions (such as positive controls, tests of intervention fidelity, manipulation checks, or any other quality checks), any tests of such assumptions should be listed as hypotheses. Stage 1 proposals that do not seek to test hypotheses can ignore or delete this column.
- **Sampling plan**: For proposals using inferential statistics, the details of the statistical sampling plan for the specific hypothesis (e.g power analysis, Bayes Factor Design Analysis, ROPE etc). For proposals that do not use inferential statistics, include a description and justification of the sample size.
- **Analysis plan**: For hypothesis-driven studies, the specific test(s) that will confirm or disconfirm the hypothesis. For non-hypothesis-driven studies, the test(s) that will answer the research question.
- **Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis**: For hypothesis-driven studies that employ inferential statistics, an explanation of how the authors determined a relevant effect size for statistical power analysis, equivalence testing, Bayes factors, or other approach.
- **Interpretation given different outcomes**: A prospective interpretation of different potential outcomes, making clear which outcomes would confirm or disconfirm the hypothesis.
- **Theory that could be shown wrong by the outcomes**: Where the proposal is testing a theory, make clear what theory could be shown to be wrong, incomplete, or otherwise inadequate by the outcomes of the research.