Dear Editor Rima-Maria Rahal,

We appreciate the reviewer’s positive and constructive reviews on our manuscript #796.

We carefully revised the manuscript, taking into account the reviewer’s suggestions and concerns.

The most important revisions concern:

- Increased the planned sample size to a minimum of 25 participants to account for potential exclusions and ensured that the study duration would be adjusted based on piloting.
- Detailed the models and software used for data analysis, including specific methods for comparing various eye-tracking parameters. Clarified the use of robust linear mixed-effects models for accuracy decay and ensured all software packages and their versions were specified.
- Incorporated updated analysis pipelines suggested by Reviewer 2 and ensured the use of appropriate algorithms for eye movement classification.
- Added a commitment to share data and analysis scripts on a public repository and adhere to Transparency and Openness Promotion (TOP) Guidelines.
- Provided explanations in the study design template on how results will be interpreted given different outcomes.
- Clarified the methods for analyzing accuracy and precision, including the handling of head movement tasks and comparison of performance between central and peripheral targets.

We have addressed all the comments of the two Reviewers point-by-point as outlined below. Revised text in the manuscript is highlighted in blue.

Sincerely,

Valentin Foucher, Alina Krug, and Marian Sauter

Editor:

Dear Dr. Foucher,

thank you for your submission "Independent Comparative Evaluation of the Pupil Neon - A New Mobile Eye-tracker" to PCI RR, for which I have now received two independent reviews by experts in the field. Based on these reviews and my own reading of your manuscript, I would like to invite you to revise the proposal. There is much to like about the manuscript, but I will highlight the most salient opportunities for further improvement below:

- Clarify the analysis pipeline (review criterion 1C)
- Consider increasing the sample size / time planned for the study (review criterion 1C)
- Clarify the calibration / validation procedure for Pupil Labs Neon (review criterion 1C)

Response: Thank you for your positive feedback. We have revised the manuscript according to the reviewers comments, as pointed out in detail below.

Reviewer 1: Lisa Spitzer

I want to congratulate the authors for this interesting Stage 1 RR, which I enjoyed reading and reviewing very much.

Summary: The authors aim to provide detailed benchmark information for the new mobile eye-tracker Pupil Neon, using the EyeLink 1000 Plus as a reference. For this reason, they plan to utilize the extensive test battery provided by Ehinger et al. (2019), taking into account not only accuracy and precision, but a broad range of different eye-tracking parameters. Participants will absolve multiple blocks of this test battery, while their eye movements will be measured simultaneously with both eye-trackers.

Response: Thank you for your positive feedback.

Major points

The authors are planning to use only a very small sample size. Given that sample sizes are typically smaller in eye-tracking studies and no specific hypotheses are tested, this might be enough, however, since data quality can vary greatly between participants, I recommend targeting a slightly larger sample / increasing the time for data collection.

Response: Thank you for this comment. Since our study is a within-subject comparison (both eye-trackers at the same time), between-subject data quality hopefully should not be that big of a concern. However, to be on the safe side regarding potential exclusions of participants for technical reasons, we increased the planned sample size to a minimum of 25 participants. We changed this accordingly in our study design template:

“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).”

I strongly encourage the authors to be more precise in describing the models used for their data analyses. For example, they should report all (random/fixed) effects used for the rLMM for task 1/7/10; which methods will be used to compare the number of fixations, fixation durations, and saccadic amplitudes between eye-trackers for task 3, how exactly the number of blinks and blink durations will be evaluated for task 5, how the normalized pupil areas will be compared for task 6, how the accuracy will be compared in the movement tasks etc. In addition, the used analysis software, including packages and version numbers, should be reported.

Response: Thanks for this feedback. We believe to have extensively improved the analysis description by detailing the tests and comparisons between eye-trackers in the “Data Analysis” subsections. References to the used analysis software and packages were moved from the “Experimental setup” section to the “Data Analysis” section to improve clarity. We hope the reviewers will be satisfied with these modifications.

“Data analysis was performed using Python 3 (Van Rossum 2009), pyEDFread (Wilming 2024), NumPy (Harris 2020), pandas (McKinney 2010), and SciPy (Virtanen 2020). Visualization was done using plotnine (plotnine development team, 2024) and Matplotlib (Hunter 2007).”

“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’s (2019) recommendations, the model was defined by LMMaccuracy ~ 1 + et session (1 + et session | subject \ block) and evaluated with the robustlmm R package (Koller, 2016).”

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

“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.”

“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).“

“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 (2006) algorithm specifically for each block to assess the variance of reported microsaccades.”

“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.” [...] “The mean difference in blink number and durations between the two eye-trackers was assessed using the 95% bootstrap confidence interval (95% CI).”

“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).”

“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).”

Minor points

L84: an eye-tracker’s (spelling)

Response: Thank you, we corrected this.

L170: Will participants with (hard/soft) contact lenses be excluded?

Response: Thank you for pointing this out. We will indeed exclude participants wearing hard contact lenses due to the difficulties they can pose for pupil center estimations and alterations to the position of the corneal reflection. Participants wearing soft contact lenses however will not be excluded since soft contact lenses are less problematic with regard to gaze data accuracy (Klein & Ettinger, 2019). Participants with soft contact lenses can abort the experiment at any time in case they experience discomfort or a drying effect caused by the illuminators (Eyelink 1000 User Manual, 2010). In this case, a new participant will be recruited. We specified this in the description of inclusion criteria:

"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."

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

Klein, C., & Ettinger, U. (Eds.). (2019). Eye movement research: An introduction to its scientific foundations and applications. Springer Nature.

L172ff: Please provide an explanation why the study was deemed exempt from ethical approval.

Response: In the opinion of the Commission, the project is not subject to consultation within the meaning of Art. 23 of the Declaration of Helsinki, i.e. no vote is required. Typically, all similar behavioral studies are declared exempt by our ethics commission but there is no specific reason indicated. The letter reads (we know the reviewer speaks German): "*[....] teilen wir mit, dass dieses Projekt nach Einschätzung der Kommission nicht beratungspflichtig im Sinne des Artikel 23 der Deklaration von Helsinki ist. Das heißt hierfür somit kein Votum erforderlich ist. Eine inhaltliche Stellungnahme oder Bewertung aus berufsrechtlicher bzw. ethischer Sicht erfolgt unsererseits daher nicht. Ein Aktenzeichen wird nicht vergeben. Dieses Schreiben können Sie als Nachweis bei eventuellen Nachfragen nach einem Ethikvotum bei Institutionen der Wissenschaftsförderung und wissenschaftlichen Fachzeitschriften vorlegen*."

L176: How can participants be excluded based on calibration accuracy for the Pupil Labs Neon glasses, if there is no calibration for this model?

Response: Thank you for pointing this out. We deleted the respective part in the manuscript.

L193-194: Why is the participant-monitor distance not yet determined?

Response: Thank you for this comment. The reason for which we didn't precise the participant-monitor distance in the manuscript was because the experiment is not set up in our lab yet. However, we can already inform that the participants will be seated at 60cm from the monitor. We added this information to the manuscript.

“The participants were seated at a distance of 60 cm from the screen.”

L198: Why use monocular recording instead of binocular for the EyeLink?

Response: We will use monocular tracking instead of binocular tracking for the Eyelink 1000 in order to use the same gaze parameters calculation methods as Ehinger et al. (2019) to ensure maximum comparability between the eye-tracker evaluation studies. We added the reference to Ehinger’s study for clarity.

“The desktop-mounted EyeLink 1000 Plus (SR Research Ltd.) recorded monocular movements of the dominant eye at 1000 Hz in head-free mode (cf. Ehinger et al., 2019).”

In L193, the authors describe that a head-chinrest is used, but it is not described whether/how this is removed for the head movement tasks.

Response: Thank you for pointing this typo out. There is no head-chinrest in our experiment since we use theEyelink 1000 in free-view mode to match the Pupil Neon usage. We deleted it from the manuscript.

L253: its (spelling)

Response: Thank you, we corrected this.

L324/327: Please add unit (degree).

Response: Thanks for helping us improve the clarity of our manuscript, we precised the units and updated the locations according to the monitor size.

“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)”

“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).”

L361: For the Stage 2 RR, I would like to know how much data had to be excluded due to data loss or corrupt data. Therefore, the authors might add a gap text or similar in which these results are reported or keep in mind to describe this in the results part of the Stage 2 RR.

Response: Thank you for reminding us to look at the missing data. We added a gap text in the Data Cleaning paragraph.

“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).”

____

L372: “The EyeLink 1000 reports blinks when the pupil is missing for several samples” – please be more specific.

Response: Thank you for this comment. We added the information about the blink classification:

“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.”

____

L400ff: Please refer to the subsection “Task 2: Smooth pursuits” in L449 for details (at first, I thought that the explanation provided in L400-402 was the only description for the smooth pursuit task)

Response: Thank you for improving the clarity of the manuscript. We added this precision:

“Please see "Task 2: Smooth pursuits" for further details.”

____

L422: I would make a distinction here between the “actual gaze point” and the fixation target - the actual gaze point may differ from the target position (e.g. due to misalignment of the fovea despite the subjective direction of gaze towards the target), but the target can be used (and is used here) as a proxy for the actual point-of-gaze

Response: Thank you for this valuable comment. Indeed we approximate here that the actual gaze point and the target location are similar since we can hardly control for potential misalignment or other variation. Then we compute the angular difference between the measured gaze point and the target location to monitor the spatial accuracy. We added this precision in the “Spatial accuracy” paragraph:

“Spatial accuracy refers to the distance between the measured gaze point and the target position. 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.”

____

(How) was performance aggregated across blocks/participants – also incorporating winsorized means?

Response: Thanks for this insightful question. Performance was monitored by first calculating the winsorized mean for each participant over blocks, and then reporting a second winsorized mean over the already averaged values. The interquartile range (IQR) of the averaged values is also reported. We added this information in the “Spatial accuracy” and “Spatial precision” paragraphs:

“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.“

“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.“

____

L442: When interpreting the results for task 10, please be clear that the accuracy will not only deteriorate due to time, but also because the movement tasks were performed before task 10, which likely changed the head positions, affecting the recording.

Response: Thanks for this insight. Indeed, while task 7 is monitoring accuracy decay over time, task 10 is additionally monitoring the influence of head movements on accuracy decay. This will be taken into account during interpretation. We also precised this in the description of Task 1 / Task 7 / Task 10:

“Task 10 is additionally monitoring the influence of head movements on accuracy decay.”

____

L442: It might be worth to also consider target distance from monitor center (as it was shown that performance might be worse in monitor corners, e.g., Spitzer & Mueller, 2022)

Response: Thank you for this advice. Indeed performance might be worse in monitor corners. We could look at the performance decay over two different groups of points - the center and the corner ones - in order to see if the eye-trackers performances are impacted differently by target distance from center. We added this precision in the “Task 1/7/10” subsection of the “Task-specific analyses” section.

“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.”

____

L476: Did you only analyze the fixation location itself, or also the accuracy? If accuracy was inspected, please report so.

Response: Thanks for this comment. The accuracy was also analyzed from the fixation location. We precised this analysis with more details in the “Task 8/9” subsection of the “Task-specific analyses” section.

“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, and then reporting the 20% winsorized mean over the already averaged values. 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, and then reporting the 20% winsorized mean over the already averaged values."

Of course, the comparison is limited due to this study incorporating other participants, a different lab etc., but I would still be interested in a comparison of the results found here with the results of the Pupil Labs model measured by Ehinger et al. (2019), which the authors could possibly include in the discussion section of the Stage 2 RR

Response: Thank you for the suggestion, we are indeed planning to discuss our results with regard to those obtained by Ehinger et al. (2019).

From personal experience (we performed a similar study, also using an adapted version of the Ehinger test battery), the experiment might take longer than 60 minutes. The authors might consider doing a test run, and planning longer time slots per participant.

Response: Thank you for your insights, we agree and will perform a piloting session to determine the duration of the experimental procedure as soon as the implementation of the experiment is completed. We adapted the respective part of the manuscript:

"The experimental session lasted approximately [tbd] minutes."

I have not found a description of how/where the data and analysis scripts will be shared. Please add.

Response: Data and scripts will be uploaded to a publicly accessible repository, such as OSF, figshare or github or similar. Within the PCI-RR process, we have already created an OSF repository. If the storage limits do not prohibit us from using it for the final manuscript, this would be a way of making it available. We will make the final "where" decision when we know the volume of the data. Aside from technical considerations, we will upload the experimental code, anonymized raw data and analysis code. We added a description under the section "Data analysis":

"Experimental code, raw data and data analysis scripts 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."

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Yarkoni, T. (2015). Promoting an open research culture. Science, 348(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

In addition to the study design template, please include a statement that you are not testing hypotheses in your manuscript text.

Response: Thank you, we added a short statement in the section “Data Analysis”:

“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.”

____

I understand that it makes sense to leave some of the study design template cells empty, but still I think that there is merit in providing some explanation in the column “interpretation given different outcomes” - e.g., what will be your conclusion given significant effects in the rLMM computed for task 1/7/10?

Response: Thanks for your comment. We provided some interpretation of the outcomes in the Study Design Table accordingly, while also updating the Analysis Plan column according to the other comments of this review.

____

____

Reviewer 2: Benedikt Ehinger

In this registered report, Foucher et al. will investigate the performance of the Pupil Neon eye tracking glasses against the current "gold standard", Eyelink 1000. For this, they closely follow our previously published EyeTracking benchmark.

The paper, and the choice of eye-tracker, is well motivated and a very valuable thing to investigate for the community. The paper further is well written and reasoned, and I have only some smaller comments.

I'm now very excited to see the outcome of this comparison.

Response: Thank you for your positive feedback.

____

Major points

We are currently re-using our benchmark and are in the analysis phase of comparing the ViewPixx Trackpix3 against an eyelink. Because of this, we updated & upgraded our analysis pipeline (we also made the stimulation code compatible with Octave, if this is interesting to the authors, they can contact me for the code, I dont think it is in the public repo) to new python version & packages. We identified one major breaking change (besides the typical renaming + adding documentation to the analysis functions):

The engbert mergenthaler implementation which we took from the Donner' Lab, has a bug in the code, slightly miscalculating the velocity threshold. Due to this reason, and some other more conceptual ones, in our new pipeline we switched to REMoDNaV, which is a successor in spirit of the Engbert-Mergenthaler algorithm. An argument could be made, that yet another

class of event-classifier should be used (Drew & Dierkes 2024), but I think it would be ill-placed given the lab & head-fixed setup.
The improved analysis code can be found here: https://github.com/behinger/etcomp/tree/etcomp2/ . Note that we are still analyzing the data, and not all tasks have been fully analysed & ported to new versions of python. There are some drawbacks: We removed quite a bit of pupil-labs specific code, this is due to some internal misscommunication with the leading-author of that study. The authors should contact us in case they want to switch to the new pipeline so we can provide requirements/yml files we are using - we are just not there yet to have a proper project :)

Response: Thanks a lot for this precision. We would be delighted to improve the study by using the updated version of your analysis pipeline. We contacted the reviewer directly to discuss the incorporation of this new pipeline into Python. We added these changes to the manuscript in the “Eye Movement Classification” and “Saccade Classification” paragraphs:

“Eye movements were defined and classified across both datasets using an updated version of Ehinger et al. (2019) algorithmic pipeline.”

“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.”

____

Further, we introduced a reading task, given a collaboration with psycholinguists on this project - you can decide whether this is relevant/interesting for your study or not. For us we included it mostly because it allows some ecological validity tests for reading studies - but if it tells you more than the large-grid task, I cannot say.

Response: Thanks for this insight. We believe that the outcomes of a performance evaluation on a reading task would be very task-specific (e.g., as a searching task) and less likely to be used in different contexts. The grid task will give a fair evaluation of the Pupil Neon accuracy, for which the outcomes could be generalized to a reading task. For this reason, we decided not to include this reading task in the current version of this study, and let future studies choose whether such an additional task-specific test is necessary.

____

Minor points

L176: You write there are calibration accuracy limits for Pupil Labs Neon, but you nowhere describe any method to identify them (see question below)

Response: Thanks for making us aware of this ambiguity in the manuscript. This sentence was a mistake since there is no calibration for Pupil Neon. We removed this part in the new manuscript.

____

Figure 1: Nice improvements! I only found the large grid illustration confusing. Why does it not look the same as the small grids ones? It seems one point is dropping of the screen

Response: Thank you, for the large grid we used small dots as placeholders for all possible stimuli positions since displaying the actual stimuli material (fixation crosses) would have led to a crowded figure. The fixation cross below the screen figure is supposed to illustrate the actual stimulus material. We increased the size of the exemplary fixation cross and now refer to the large grid figure in the figure notes.

"This figure illustrates the task sequence within each experimental block. All possible stimuli positions are marked in gray, gray dotted arrows indicate stimulus movement. Gray markings were not shown throughout the trial. For the large grid task, fixation crosses served as stimulus material."

____

L267: Calibration/Validation of Pupil Labs Neon. Is there no settings whatsoever that you decide per subject that could influence accuracy? And, is there no recommended validation behavior?

Response: Thanks for your insightful comment. In their own study that tested the accuracy of the Neon, Pupil Labs refers to a "personal gaze offset correction" that can be done per subject and indeed seems to increase the accuracy. We were in personal communication with Pupil Labs to discuss this procedure. We decided to use it and added a sentence about it in the Eye-tracker calibration section.

"However, a personal gaze offset correction was performed for each participant to maximize Neon's accuracy. This offset correction was achieved directly on the companion device by fixating a single point at the center of the screen and applying the correction accordingly to the procedure described on Pupil Labs website."

https://docs.pupil-labs.com/neon/data-collection/offset-correction/

____

L366: As stated above, I would probably move to remodnav due to the bug in the mergenthaler algoritm (or fix the bug)

Response: As previously answered, we would be delighted to improve our manuscript by using an updated analysis pipeline. We contacted the reviewer directly to discuss the implementation of REMoVDNaV in the Python analysis script.

____

L414: You argue to convert PupilLabs Pupil measurement to area similar to eyelink. But I would argue, that the pupil-labs 'mm' output, is the actual more interesting and relevant output. So maybe calibrating the pupil for eyelink should be the goal, rather than "deconverting" the pupil-labs output back to ellipses / areas?

Response: Thanks for this feedback. We agree that using the "mm" output would give additional information on the actual measured size for each eye-tracker. The reasons that led us to choose the "area" unit were to keep the Eyelink 1000 as the reference eye-tracker by using its outcome units as default, and to follow the same analysis pipeline as Ehinger's (2019) analysis plan. Moreover, while converting Eyelink 1000 pupil size from "area" to "mm" is technically possible, it is not recommended to do it in head-free mode since the conversion requires the use of an artificial eye attached to the head support at a fixed distance from the monitor

(https://www.sr-research.com/support/thread-154.html?highlight=convert+pupil+size+to+mm). Finally, the baseline-corrected pupil size is used to evaluate the eye-trackers performances, which lowers the urge to use a specific unit. This will be something to discuss in the discussion part of the manuscript.

____

L420: There is a mistake in the formula in our paper (2*atan2 should be atan2) - we have a correction request pending since January for this. sorry for the inconvenience. The code is correct though.

Response: Thanks for notifying us of this mistake. We will take it into account during the conversion of the gaze points from screen coordinates to spherical angles.

____

Open question for analysis plan: Do you (winsorized) average the accuracy values per block, then take the winsorized mean over blocks, then the (bootstraped) winsorized mean over subjects? Afaik this is how we did it. You could also disregard blocks and immediately go for subjects. Maybe I missed it in the manuscript.

Response: Thanks for your question. For accuracy analysis, we plan to do the winsorized mean of gaze positions for each block, then take the winsorized mean over blocks for each participant, and then the winsorized mean over participants for both eye-trackers. Finally, the difference between the two eye-trackers is assessed using the 95% bootstrap confidence interval. Indeed we could disregard blocks and immediately go for subjects, but we believe having the “block step” would better capture the decay. We clarified our analysis plan in the new version of the manuscript.

____