

A study of artificial eyes for the measurement of precision in eye-trackers

Dong Wang¹ · Fiona B. Mulvey² · Jeff B. Pelz¹ · Kenneth Holmqvist²

Published online: 6 July 2016
© Psychonomic Society, Inc. 2016

Abstract The precision of an eye-tracker is critical to the correct identification of eye movements and their properties. To measure a system's precision, artificial eyes (AEs) are often used, to exclude eye movements influencing the measurements. A possible issue, however, is that it is virtually impossible to construct AEs with sufficient complexity to fully represent the human eye. To examine the consequences of this limitation, we tested currently used AEs from three manufacturers of eye-trackers and compared them to a more complex model, using 12 commercial eye-trackers. Because precision can be measured in various ways, we compared different metrics in the spatial domain and analyzed the power-spectral densities in the frequency domain. To assess how precision measurements compare in artificial and human eyes, we also measured precision using human recordings on the same eye-trackers. Our results show that the modified eye model presented can cope with all eye-trackers tested and acts as a promising candidate for further development of a set of AEs with varying pupil size and pupil-iris contrast. The spectral analysis of both the AE and human data revealed that human eye data have different frequencies that likely reflect the physiological characteristics of human eye movements. We also report the effects of sample selection methods for precision calculations. This study is part of the EMRA/COGAIN Eye Data Quality Standardization Project.

Keywords Eye movements · Artificial eye · Precision · Data quality · Eye-tracker noise · Power-spectral density

High-quality eye movement data are a prerequisite for the valid measurement of fixation durations, saccade amplitudes and velocities, and many other behavioral measures in eye movement research. Spatial accuracy and precision are two of the most important aspects of eye data quality. *Accuracy* is defined as the difference between the tracker-estimated gaze position and the actual gaze position, whereas *precision* is defined as the ability to reliably reproduce a measurement, given a fixating eye (ideally, a *stable* eye)—see Fig. 1. Accuracy and precision are two independent measurements of eye-tracking data quality—that is, they can be both good or poor, or one good and the other poor. The most commonly used measures of precision in eye-trackers are the sample-to-sample root mean square angular displacement [RMS(S2S)] and the standard deviation (*SD*) of samples in a given time window. These values will change not only dependent on the actual precision level of the tracker but also on the calculation used, and the samples or time period included in that calculation. Longer periods increase the probability of fixational eye movements (tremor, microsaccades, and drift), which in turn will increase imprecision. Holmqvist, Nyström, and Mulvey (2012) showed that artificially increasing imprecision via the addition of Gaussian noise from 0.03° to 0.30° results in an increase of up to 200 ms in calculated fixation durations. If this range of noise, as measured by RMS(S2S) precision, is representative for current eye-trackers, it means that identical eye movements recorded on different eye-trackers, or with different levels of precision in recordings from the same eye-tracker, can lead to different research results. This has a profound effect on the replicability and valid comparison of research results.

✉ Dong Wang
dxw1481@rit.edu

¹ Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, New York, USA

² Humanities Laboratory, Lund University, Lund, Sweden

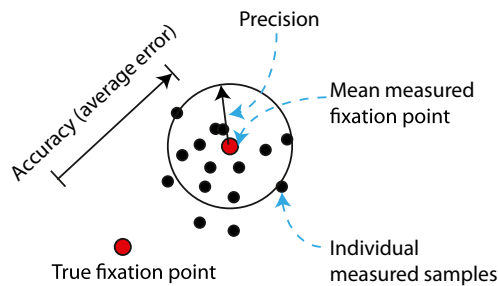


Fig. 1 Definitions of accuracy and precision, from eye-tracking data reproduced from Cleveland, Mulvey, and Pelz (2015). For a fixation, accuracy is calculated as the difference between the mean measured gaze position and the true fixation position. Note that the true fixation position can only be ascertained by feedback from the participant. Precision is a metric either of the sample-to-sample movement or of how spread-out the raw gaze samples in the fixation are

The precision of an eye-tracker is known to vary with the features of individual participants' eyes, the recording environment, the ability of the participant recorded to fixate, and system specifics. For instance, Nyström, Andersson, Holmqvist, and van de Weijer (2013) showed that when recorded using a SensoMotoric Instruments (SMI) HiSpeed 1250 tower eye-tracker, blue eyes result in data with a poorer RMS(S2S) precision than brown eyes, and that data recorded from participants with glasses are more imprecise than data from uncorrected eyes in this particular eye-tracker. Increasing the pixel resolution at the eye (e.g., by using a higher-resolution camera or placing it closer to the eye) makes precision better [lower RMS(S2S) or *SD*]. Since all measures of precision involve selecting samples or segments of time within which to calculate, sample selection methods are also a source of unknown variance in reported precision values. Selecting samples within a time window, but removing spatial outliers, will always result in a better precision value, as can extensive loss during the sample period. Although this might seem obvious, it is nonetheless common practice to remove outliers before calculating precision, both in research on eye-tracking methodology and in manufacturer-supplied values. This study draws on methods developed for the purposes of standardizing measures of eye-tracker data quality in examining the effect of sample selection, window size, artificial eye properties, and noise components across 12 commercially available eye-trackers.

When calculating precision for an individual on a particular eye-tracker, the samples recorded when the eye is fixating a stationary target are usually used. The intention is that the samples' spatial variation originating from eye movement is minimized as much as possible. Generally, a reasonably short period during fixation is chosen, but there is no consensus on how long the period should be, or indeed whether the exclusion of samples on the basis of outlier spatial extent is valid. The only way to completely exclude human eye movements from the calculation is to use an artificial eye (AE) positioned

in front of an eye-tracker (Abramov & Harris, 1984; Holmqvist et al., 2011; Reingold, 2014). Most AEs used by eye-tracker manufacturers are similar to doll's eyes, with an artificial pupil and a reflective surface that generates a first-surface reflection (corresponding to the corneal reflection). Some AEs are even simpler, and have only a dark circle to represent the pupil and a ball bearing to create the glint. AEs are designed according to how a human eye appears to the particular eye-tracker to be tested, which may be more or less complex, but are greatly simplified relative to a real eye. Video-based eye-trackers generally rely on the recognition of the pupil and corneal reflection(s) to track eye position, sometimes relying on separate face detection to help locate the eye in space. With these requirements, for some eye-trackers, manufacturers use a printed image of a face with two AEs stuck in place to test the quality of the data their systems produce. It is a challenge for eye-trackers that require a bright pupil reflection (elicited by on-axis illumination) to find any AE model that produces the kind of image their internal eye model recognizes. Bright pupil tracking is used by several major eye-tracker manufacturers, including Tobii, Applied Science Laboratory (ASL), SmartEye, and LC Technologies; therefore, there is a need to develop AEs that work with these systems, too. As a unique case, the dual-Purkinje imaging (DPI) eye-tracker (Crane & Steele, 1985) does not use any tracking of the pupil, but instead relies on the fourth Purkinje reflection at the back of the crystalline lens, along with the corneal reflection, to determine eye position and orientation.

In principle, we may assume that noise in data from AEs and human fixating eyes is different. Theoretically, the total noise from human eye data, even during careful fixation, would include both biological components and the system-generated noise. Although this may suggest that precision estimates are constant across samples for AEs, the exact implementation of image recognition techniques in the various eye trackers may result in trial-by-trial variations in the precision estimates, due to small vibrations, minor illumination changes, or camera noise. A range of AEs have been developed by both manufacturers and researchers, but little is known about how these models compare. We therefore set out to compare four sets of AEs on a range of commercially available eye-trackers, including tower-mounted and remote video-based eye-trackers, and the DPI (12 trackers in total). Three sets of AEs were provided by their manufacturers (Tobii, SMI, and SR Research), whereas the fourth AE was a modified version of a commercially available eye model originally designed for ophthalmic imaging training (Ocular Instruments, 2013). On the basis of what is known about how human variations affect noise characteristics, we would expect different precision measurements across the different AEs, due to variations in iris color, pupil size, and reflectivity. To establish how the precision measured in AEs compares to the

precision from human eyes, we measured the precision of the eye-trackers in a set of 20 human participants. We also examined the influence of the number of samples included in the calculation to study how precision is affected by sample selection.

Precision is influenced by noise, but it is unclear how different aspects of the noise contribute to precision. We therefore carried out an analysis of the spectral qualities of the noise in the various eye-tracking systems. Previous work by Findlay (1971), using eye movement signals recorded on magnetic tape, suggests that although the methods of recording he used are very different from current eye-trackers, the nature of the system noise component described is the same random spatial displacement reported from current trackers (Mulvey et al., 2015). According to Findlay, the statistical characteristics of the saccadic component allow its mean power to be estimated, even below the system noise level. Biological tremor is described as having similar characteristics to system noise, so their power spectra add. A corrected tremor component is estimated by subtracting the system-noise power from the observed random-component power. Coey, Wallot, Richardson, and Van Orden (2012) measured the noise of an ASL eye-tracker, and through fractal analysis showed that noise from humans had a fractal structure (i.e., pink or $1/f$ noise), whereas noise from an AE was white—that is, it has the same energy at all frequencies. The fractal structure is characterized by a scaling relation between the power of changes in the measured variable and the frequency with which changes of that size occur. In more recent research from Wallot, Coey, and Richardson (2015), eye movement spectral characteristics were generalized via power-law scaling, and the results indicated that this characteristic reflects a relative demand for voluntary control during visual tasks. To validate that these spectral characteristics are a pure effect of human eye movement, we need to verify that the eye-trackers used in such studies do not yield power-law scaling, which we can measure with AEs. In our study, we analyzed the noise characteristics from both human and AE data using power-spectral density analyses to test whether the statements above hold for a variety of trackers and AEs.

Procedure

AEs and eye-trackers

Four sets of AEs were used. Three of them were provided by the manufacturers: SR Research, SMI, and Tobii, shown in Fig. 2A, C, and D. They are regularly used by the above manufacturers to test the precision of their eye-trackers. We also used an ocular-imaging eye model modified from Ocular Instruments (Model OEMI-7), shown in Fig. 2B. This AE has not previously been used with eye-trackers. We tested and

modified this AE for our purposes, since its structure more closely resembles the human eye—with an anterior chamber, crystalline lens, fluid filled globe, reflective retina, and fundus. However, since the original material of this eye model has different refractive indices than the human eye, the reflectance of the back surface of the lens (the fourth Purkinje image) and the bright pupil reflection is much brighter than ideal. We modified this AE as we describe in the Appendix to bring the values closer to those in the human eye.

We tested 12 commercial eye-trackers in this study, including three tower-mounted eye-trackers, eight remote eye-trackers, and the DPI. Their sample rates and manufacturer-provided precision values are shown in Table 1. Among the video-oculography (VOG) systems, the EyeFollower is the only bright-pupil system tested, whereas the others are dark-pupil systems. Bright-pupil imaging is achieved with a different illumination geometry, in which the eye camera illuminator is coaxial with the eye camera, causing a bright reflection from the retina in the eye image. Dark-pupil systems place the illuminator off-axis, so the pupil appears dark. The eye-trackers tested were either part of the equipment available locally, or were on loan from manufacturers in support of the standardization effort.

Calibration and data recording

To calibrate the eye-trackers before data recording, different procedures were carried out on different types of systems. For tower systems, we performed both (a) human calibration and (b) AE calibration. For the first step, we calibrated the system with a human participant, and then we replaced the human with a single AE in the recommended geometry with the tracker (using a tripod for the SMI tower systems, and mounting it directly on the chinrest on the EyeLink). Next, the pupil and CR thresholds were adjusted to work with the AE, and recording began. When calibrating with AEs, we mounted the eye model on a gimbal mount with a laser diode attached to the top of the eye model, as is shown in Fig. 3. The brass housing of the laser diode was hot-glued to the outside of the gimbal's inner ring so that the relative position between the laser diode and the eye model was fixed. Note that we did not do an optical alignment, so there might be a small angle between the optical axis of the AE and the laser diode, which is consistent like angle kappa (kappa is the angle between the optical axis and the visual axis; Artal, Benito, & Tabernero, 2006). The distance from the exit pupil of the laser diode to the center of the AE was 17.5 mm. We used a calibration target sheet attached in front of the screen with a projection of each calibration point for the laser to aim at. After the calibration, the AE was rotated to a “gaze” point at roughly the center of the screen, and the recording began.

For remote systems, we calibrated using human observers only and recorded data using the AEs. Since all of the remote

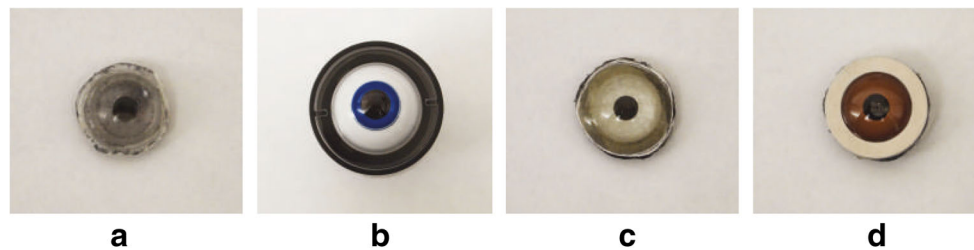


Fig. 2 Artificial eyes tested. (A) Eye model from SR Research. (B) OEMI-7 from Ocular Instruments. (C) Eye model from SMI. (D) Eye model from Tobii. The eyes in panels A, C, and D have a pupil size of

about 5 mm, and the one in panel B has a pupil size of 7 mm and is capable of bright pupil reflection and 4th-Purkinje reflection

systems tested were binocular systems, we put two AEs of each type on a printed face with an interpupillary distance of 65 mm and mounted the “face” on a tripod. For the AEs from SMI, SR, and Tobii, the calibration distance from the human eyes to the stimulus monitor was 60 cm, and the tripod was positioned so that the distance between the AEs and the monitor screen was also 60 cm, to comply with the manufacturer-recommended setup. For the OEMI-7, we recorded at 70 cm, because tracking at 60 cm was poor, which might have been due to the large pupil size of the OEMI-7. All systems were powered up for at least 10 min (the DPI was powered up for 30 min) before the experiment, as some manufacturers recommended.

The DPI data were recorded and calibrated offline after recording. The Generation V DPI tracker is linear within the

central 5°, so calibration consisting of gain/offset was sufficient (Crane & Steele, 1985).

Two recordings with independent calibrations were made on each of the other systems using each AE. Each recording contains 60 s of data. The workflow of the experiment was as follows: The tracker’s native calibration routine was initiated at the beginning of the experiment. Following calibration, the AE was manually aimed at the center of the stimulus screen, and 60 s of static data were recorded.

The ioHub (iSolver Software Solutions, 2013), a Python library in PsychoPy (Peirce, 2007), was used to program the experiment and record the data. We used ioHub instead of the experiment software provided by the manufacturers, because ioHub enables the user to run exactly the same experiment on all eye-trackers with careful temporal control and replication of stimulus properties, and to save the recordings into the same HDF5 format (HDFGroup, 2014; iSolver Software Solutions, 2013).

Table 1 Sample rates and manufacturer-reported precisions of the eye-trackers used in this study

Tracker	Sample Rate (Hz)	Manufacturer Reported Precision (Visual °)	Type
The EyeTribe	30	0.10	Binocular
Tobii X2-60	60	0.34	Binocular
Tobii T60XL	60	0.09	Binocular
EyeFollower	60	Not reported	Binocular
SMI REDm	120	0.10	Binocular*
SMI HiSpeed 240	240	0.01	Monocular
SMI RED250	250	0.03	Binocular*
Tobii TX300	300	0.14	Binocular*
SMI RED500	500	0.03	Binocular*
EyeLink 1000**	1,000	0.01	Monocular
SMI HishSpeed 1250	1,250	0.01	Monocular
DPI	1,000***	<0.02	Monocular

Some trackers can operate under more than one sample rate, but we list the sample rate we used. The reported precisions are from the manufacturer specification sheets and websites. The unit of precision is degrees of visual angle. * For binocular systems in which eye data are averaged across the eyes by default, this averaging was turned off in our study to get a true measure of recording quality. ** The EyeLink 1000 was set up in tower mode for this study. *** The DPI system is an analog system, and we sampled it at 1000 Hz.

Precision calculation

With the recorded data, precision was calculated as a function of the angular distances θ_i (in degrees of visual angle) between

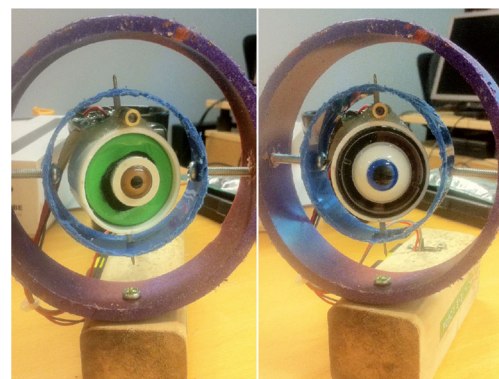


Fig. 3 Gimbal mount used to record AEs with tower setups—where the head would normally be in a fixed position via a chinrest or a bite bar. The gimbal eye can rotate horizontally and vertically. A laser diode is attached on top of the inner ring (17.5 mm away from the pupil center) to indicate gaze direction

successive data samples (x_i, y_i) to (x_{i+1}, y_{i+1}) (Holmqvist et al., 2012), of the form

$$\theta_{RMS(S2S)} = \sqrt{\frac{1}{n} \sum_{i=1}^n \theta_i^2}, \quad (1)$$

and the SD was calculated as

$$\theta_{SD} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_x)^2 + (y_i - \mu_y)^2}, \quad (2)$$

where μ_x and μ_y are the means of n sample locations. Both of these metrics are used in the eye-tracking field. RMS(S2S) tends to be sensitive to big sample-to-sample changes, but not to a gradual drift, and the SD represents the overall spread of the n data samples, so it is sensitive to both. Blignaut and Beelders (2012) also showed that RMS(S2S) is biased toward data from higher-sample-rate systems: With the same movement over time, high-sample-rate system would have better-precision data than low-sample-rate systems, as a result of the temporal proximity (and, given the nature of the behaviors recorded, therefore also spatial proximity) of successive samples.

For each tracker and each AE, 20 segments of 1,000 ms were randomly selected (under the constraint that the segments could not overlap temporally) from the data. Each time window was treated as a “fixation,” and the precision calculation was carried out within the time window. As the first step, we calculated the precision using different lengths of data samples (100, 175, 225, and 500 ms), which we will refer to as the *window size* in the rest of the article. Having several sample window sizes, as is illustrated in Fig. 4, allowed us to see the effects of sample rate and window size on RMS(S2S) and on SD . On the basis of our investigations, we chose a window size of 225 ms as the best compromise, providing a sufficient number of data points at lower sample rates while minimizing the probability of including fixational eye movements.

Comparison with human data

We also recorded human data with all listed trackers with 194 human participants in a separate study (Mulvey et al., 2015). To be comparable with the data from the AEs, which was directed toward the center of the screen, we only used the human gaze data related to a target at the center of the stimulus monitor from 20 participants. Since the participants of the larger study had a variety of pupil occlusion, eye makeup, and other quality-decreasing eye characteristics, we selected 20 from the original 194 participants for comparison with the AE measures. The selected participants wore no contact lenses or glasses, did not have laser surgery, did not take any medications or have any conditions known to affect fixational stability, did not have extensive loss during recording, wore no eye makeup, and had a fully visible pupil when they looked straight ahead. The minimum imprecision values of these 20 observers were treated as representative of good human precision. In the human eye data, the corresponding sliding windows (shown in Fig. 4) ran from 200 to 1,200 ms after target onset, to minimize the effect of saccade latency (Carpenter, 1988). Track loss of one or more sample in human data resulted in that window being excluded as a candidate for the precision calculation. The window with the lowest imprecision was selected for the calculation of minimum RMS(S2S) and SD across the 1,000-ms presentation time. We applied sliding windows of the same temporal extent as the AE recordings: 100, 175, 225, and 500 ms. For the human data, we expected window size and sample rate to have larger effects than would be seen with the AEs—compounded by an increasing probability of including fixational eye movements.

We did not average the left and right eyes on the remote systems, which is the default option for some eye-trackers. For many practical applications of eye-tracking, averaging makes very good sense (where an analysis is made with areas of interest [AOIs], for instance), whereas for other types of experiments, averaging invalidates certain measures (e.g.,

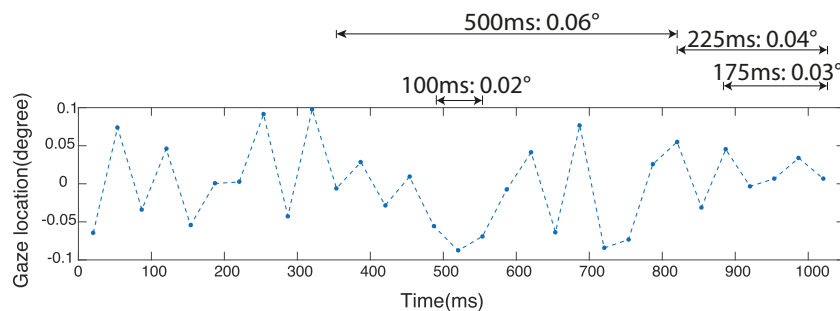


Fig. 4 Example of sample selection with a sliding window of varying duration for SD calculation. The vertical axis is the horizontal gaze position, and each point represents a gaze sample. The arrows at the top of the figure show the windows that resulted in the smallest SD s at each window size, with the SD value calculated in degrees of visual angle. Different window sizes result in the selection of different time periods

for the minimum SD , and these periods do not necessarily overlap with each other or with the periods that would be selected for the minimum RMS(S2S). Shorter sliding windows selected for minimum SD or RMS(S2S) tend to result in lower variation, due to the greater probability of selecting brief periods that randomly have low variation, due to the stochastic nature of the data

saccadic velocity profiles, saccade and fixation durations, or binocular coordination). We did not want the AEs to be averaged, because our results would then only be appropriate for a subset of studies that do not rely on the true dynamics of the eye, and hence we turned off the default averaging for those systems.

Power-spectral analysis

Coey et al. (2012) did a spectral analysis of eye data using an AE and human data on an ASL-D6 eye-tracker, and found that the AE data produced white noise, whereas the human data produced pink noise or $1/f$ noise, in which the power (P) is inversely related to the frequency (f). The power-spectral density of pink noise is in the following form:

$$S(f) \propto \frac{1}{f^\alpha}, \quad (3)$$

where f is the frequency and α is a scaling exponent. In a strict definition of pink noise, α should be equal to 1, which means that each frequency carries an equal amount of noise power. In behavior and the neural science literature, the term is sometimes used more loosely, where $0 < \alpha < 2$.

Because of track loss and temporal imprecision in eye-trackers, we could not assume that samples were regularly sampled in time—a requirement for traditional spectral analysis methods. The Lomb–Scargle periodogram (Lomb, 1976; Scargle, 1982) was developed specifically to allow spectral analysis with uneven sampling, so we adopted this method for the power-spectral density (PSD) analysis of the AE and human data. Prior to the PSD analysis, we removed data exceeding three SD s from the mean, as was suggested by Coey et al. (2012), but replacing missing samples with linear interpolation was unnecessary, because of the nature of the periodogram analysis.

Results

Precision of AEs and human data

We calculated the precision using both RMS(S2S) and SD , as is described in Eqs. 1 and 2, on 20 segments (1,000 ms per segment) of data from each AE, on each eye-tracker, with a sliding window across the entire segment, and report the smallest value. This was repeated with window sizes from 100 to 500 ms. Figure 5 illustrates how, even in the absence of any biological movement, the number of data points included in the RMS(S2S) calculations—that is, the combined effect of window size (number of samples being used in the calculation) and eye-tracker sample rate—had an effect on the calculated precision values. Calculating from larger window

sizes had a large effect on the reported RMS(S2S) values, especially in systems with low sample rates. SD values were more stable over numbers of data points than RMS(S2S) up to a window size of 500 ms, but the value was still affected by the number of data points submitted to the calculation. In panels a–f of Fig. 5, the Eyefollower and the DPI are not included because the SRR, SMI, and Tobii eyes were not recognized as eyes by these systems. In panels g and h, in which OEMI was used, all 12 systems were included using our modified OEMI, but both metrics increased when using this eye model for remote systems. For almost all eye-trackers and all window sizes up to 500 ms, the RMS(S2S) imprecision was significantly larger than the SD calculation. The difference was most apparent in eye-trackers with lower sample rates—for example, the Tobii X2 and the EyeTribe eye-tracker. For shorter window sizes, the RMS(S2S)-calculated imprecision was up to twice that calculated using SD for these systems. However, the rankings of eye-tracker imprecision were similar or identical across the two calculations.

We then compared precision under the two calibration methods with data from the three tower systems, to see whether calibrating with a human and subsequently substituting an AE for recording produced different results than calibrating and recording from AEs. These results provide an indication of the relative performances for human and AEs. For tower systems, imprecision recorded from an AE following calibration with a human eye gives a different result than when calibrating using AEs. On the basis of an independent-samples t test, when using the SMI eye model, human calibrated data has significantly higher RMS(S2S) than the AE calibrated data for all three systems, as we might expect [EyeLink: $t(59896) = 708.6$, $p = .02$, $d = 0.82$; HiSpeed 240: $t(14155) = 661.7$, $p = .01$, $d = 1.10$; HiSpeed 1250: $t(73747) = 208.1$, $p = .02$, $d = 0.73$]; whereas when using the Tobii eye model, only the EyeLink and HiSpeed 240 had a significantly higher human calibrated RMS(S2S) [EyeLink: $t(59893) = 186.47$, $p = .02$, $d = 0.64$; HiSpeed 240: $t(14156) = 89.28$, $p = .01$, $d = 0.73$]. When using the EyeLink eye model, the EyeLink system actually had higher AE calibrated RMS(S2S) than did the human eyes [$t(59892) = 222.57$, $p = .01$, $d = 0.54$]; For the OEMI-7 eye, no significant difference was found between the two calibration methods. Note that all precision calculations in this case were from AEs, and only the initial calibration methods (human vs. manually directed AE) differed.

We also calculated the RMS(S2S) and SD of the human data from recordings on the same trackers using the same sliding-window methodology as with AE data. The combined effects of sample rate and window size are illustrated in Fig. 6. With an increasing window size, the median as well as the variation in both measurements increased. As with AE data, the human data also showed an effect of sample rate on RMS(S2S) and SD . The three highest-sample-rate trackers (EyeLink, DPI, and SMI Hispeed1250) elicited lower

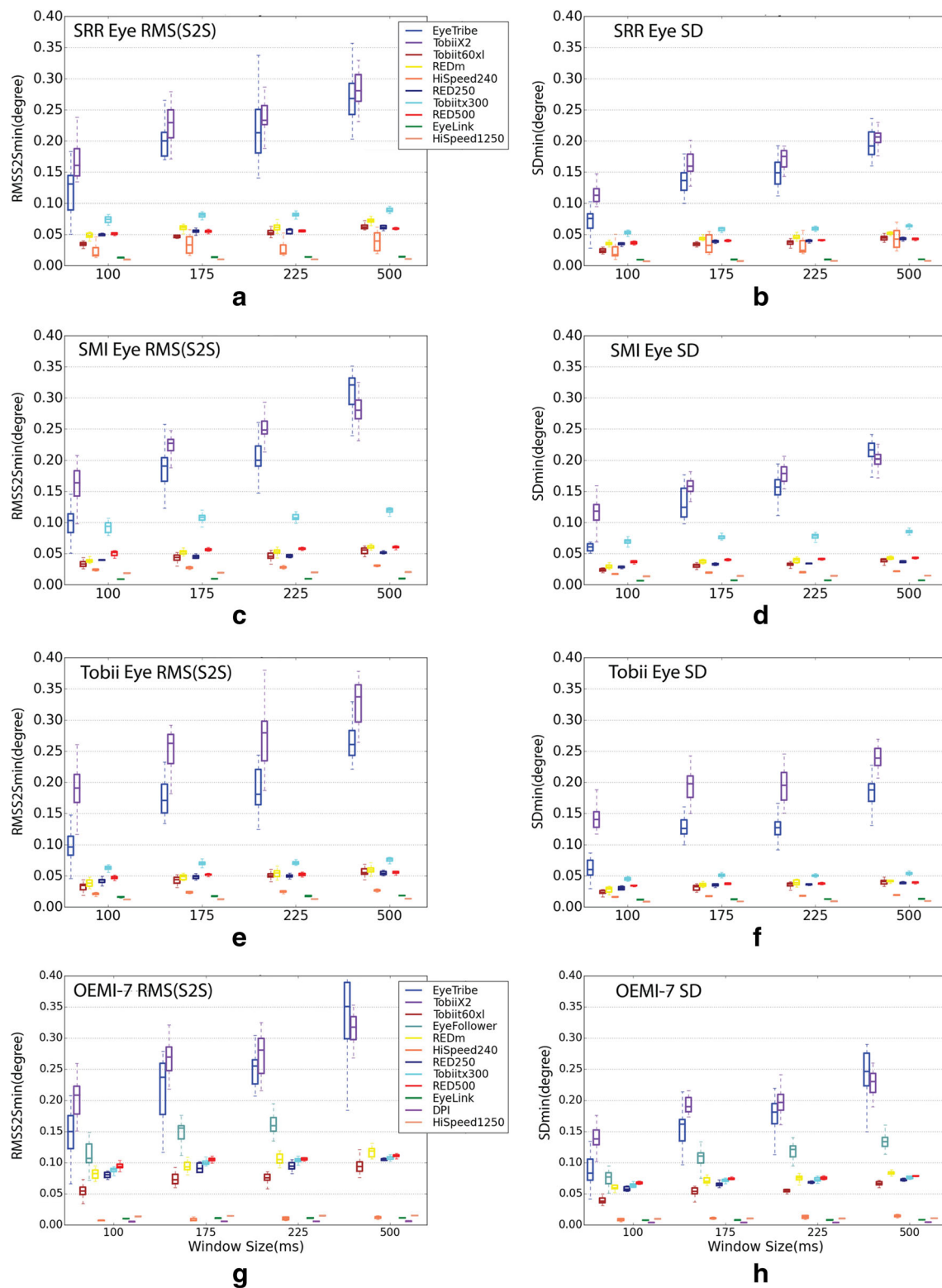


Fig. 5 Minimum RMS(S2S) (left column) and *SD* (right column) for sliding windows of various temporal extents on $20 \times 1,000$ -ms segments of AE recordings. The minimum from each sliding window is

reported; the horizontal axis is the window size, and the vertical axis is the precision metric. Data are shown for ten compatible trackers in the SRR, SMI, and Tobii panels, and the OEMI-7 results are for all 12 trackers

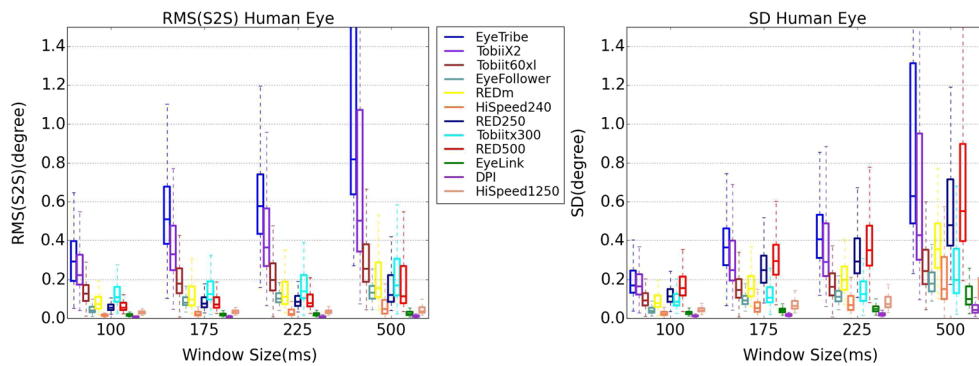


Fig. 6 Minimum RMS(S2S) (left) and minimum SD (right) values calculated from sliding windows of various temporal extent on 1,000-ms segments of human eye recordings. Participants fixated at the center

of the screen on 12 eye-trackers with varying sample rates. For each window size, from left to right, the trackers' sample rates increase

RMS(S2S) than SD across the different window sizes, whereas the lower-sample-rate trackers (EyeTribe, Tobii x2, Tobii T60XL and EyeFollower) had lower SD than RMS(S2S), which again demonstrated that RMS(S2S) is more biased toward higher-sample-rate systems than SD is.

We then explored how precision across eye-trackers compared between the AEs as a group and the group of human eyes. For comparability of the two data sets, we took five nonoverlapping temporal patches of 1,000 ms from each AE recording on each eye-tracker and compared these to 1,000-ms recordings of 20 humans fixating at the center of the screen on each of the eye-trackers tested, varying the sample window size and precision calculation as before. Comparing RMS(S2S) and SD from AEs and 20 human participants, with varying window sizes as in Fig. 7, we can see that as predicted, the difference between AEs and human eyes becomes larger as the window size increases. For lower-sample-rate systems—for example, the EyeTribe—both the RMS(S2S) and SD calculations showed rapidly decreasing precision with increasing window size. Higher-sample-rate systems generally had stable RMS(S2S) and SD for a range of window sizes. As the window size increased, SD produced a larger imprecision value than RMS(S2S) for all systems. As we would expect, the biological signal present from the small movements of human eyes, but absent from AEs, yielded a much higher imprecision value with increasingly large sampling periods in all systems. We can also see that the difference in precision between eye-trackers is much greater for real human recordings than for AEs. Part of the human “imprecision” (in this case, actually correct behavioral recordings rather than “noise,” per se) will be due to microsaccades and other fixational eye movements.

PSD analysis

To assess spectral noise, we conducted a PSD analysis on both the human and AE data for all the trackers. Figure 8

shows an example of the power spectrum from the same tracker with human and with AE data. The logarithm of the power and the logarithm of the frequency were plotted against each other. The scaling exponent α , as in Eq. 3, is defined as the negative of the slope of the regression line. In this figure, it is evident that the human data have a larger α value than the AEs, which aligns with the results from Coey et al. (2012) and Wallot et al. (2015). The same analysis was carried out on all AE and human data. From Fig. 8, it is obvious that human eyes elicit larger α for all the trackers. All of the AEs elicited α s closer to 0, which indicates white noise. Coey et al. (2012) found that human data are described by a scaling exponent in the range of 0.8 to 0.9, whereas our analysis produced scaling exponents mostly around 1.5. A larger α means more energy in lower frequencies, which would be expected from human data. When α equals 2, the noise becomes random-walk noise (Barnes & Allan, 1966). Figure 8 also illustrates that AEs contained some nonstochastic signal components, which could not be attributed to biological noise and were not simply a result of either the system or the AE model, but likely were a combination of both—that is, how well the AE matched the system's model of an eye. Here, we can see that some systems did better with human eyes than with AEs—most notably the EyeFollower on human versus the OEMI-7. Due to its reliance on bright pupil reflection, this system represents a type that could not be measured with the three existing simple AEs.

Effects on the ranking of eye-trackers

Table 2 shows the ranking of systems by measured precision. The three manufacturer eyes give roughly the same rankings, but with several exceptions, including that the SMI AE ranks the EyeLink 1000 as the most precise eye-tracker tested, whereas the SR Research eye puts the SMI HiSpeed 1250 at

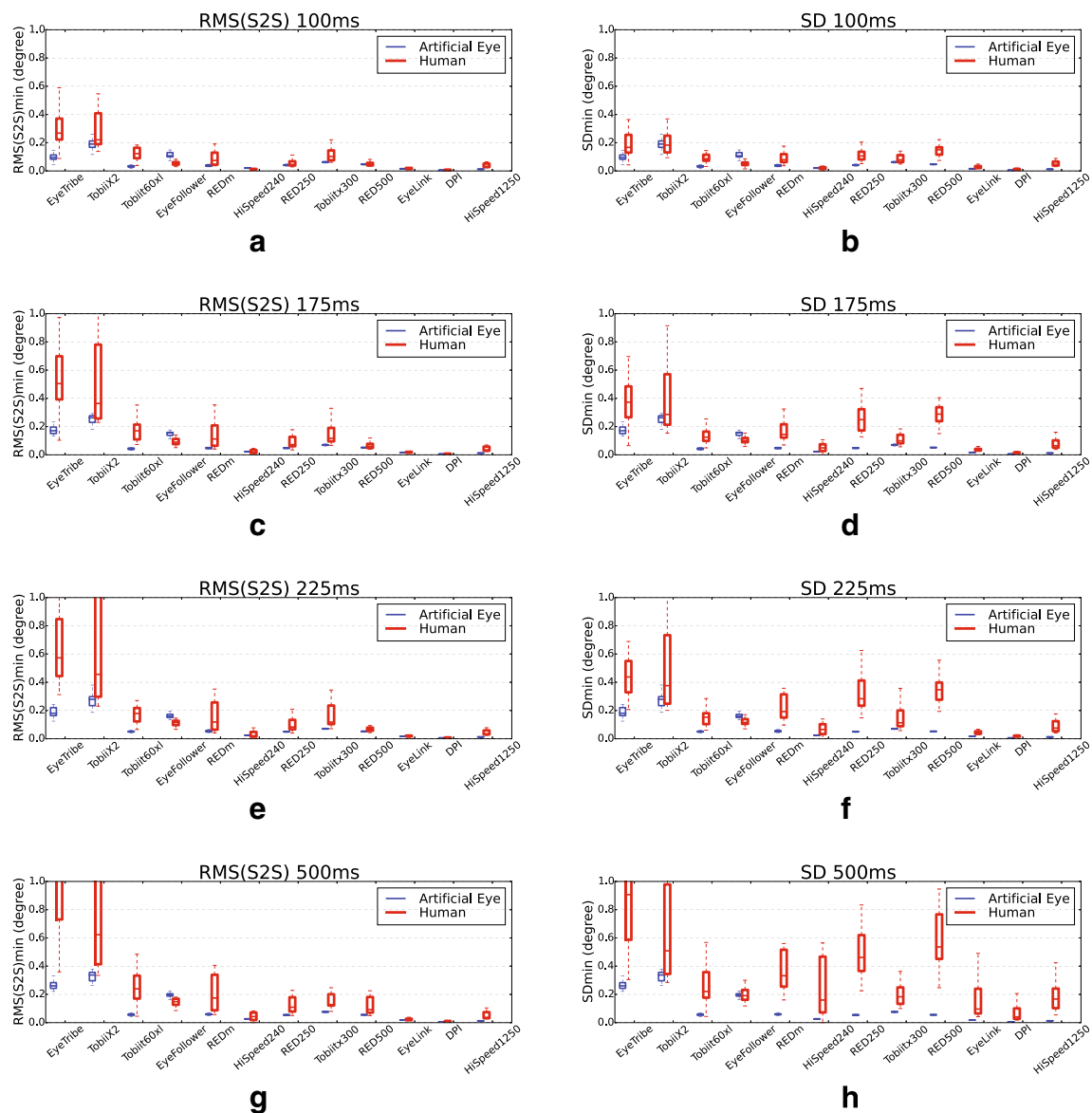


Fig. 7 Precision, calculated as RMS(S2S) (left) and as *SD* (right), from the group of 20 human participants and 20 comparable 1-s segments of AE recordings

the top. In contrast, the rankings from the human data differ much more.

Discussion

The main aim of the present study was to establish the reliability of different AEs in the assessment of the precision of different eye-trackers. Our results show that data from AEs have a lower (better) precision value, as both RMS(S2S) and *SD*, than human recordings for all of the eye-trackers tested except the EyeFollower. This difference reflects the fact that AE data do not include signal components of a biological origin. However, the four AEs that we tested reported variations in precision among themselves, which indicates that data

from them contain not only system noise, but also a component specific to each eye—most likely due to variation in iris brightness and pupil size. For a specific eye-tracker, we might take the smallest precision value from the four AEs as the system noise, or consider the AE that produced the smallest imprecision as the one that most closely modeled the internal eye model for this particular system, but the reality of recording varying human eyes or AEs from other manufacturers is that the internal eye model will not match any given human eye perfectly. When ranking the precision values from different eye-trackers, the AE and human recordings agreed that certain systems are better than others. They all ranked the SMI HiSpeed systems and the EyeLink in the first four places, and they all ranked the TobiiX2 and EyeTribe at the bottom, but the variations in Table 2 show that all eye-trackers are, to a

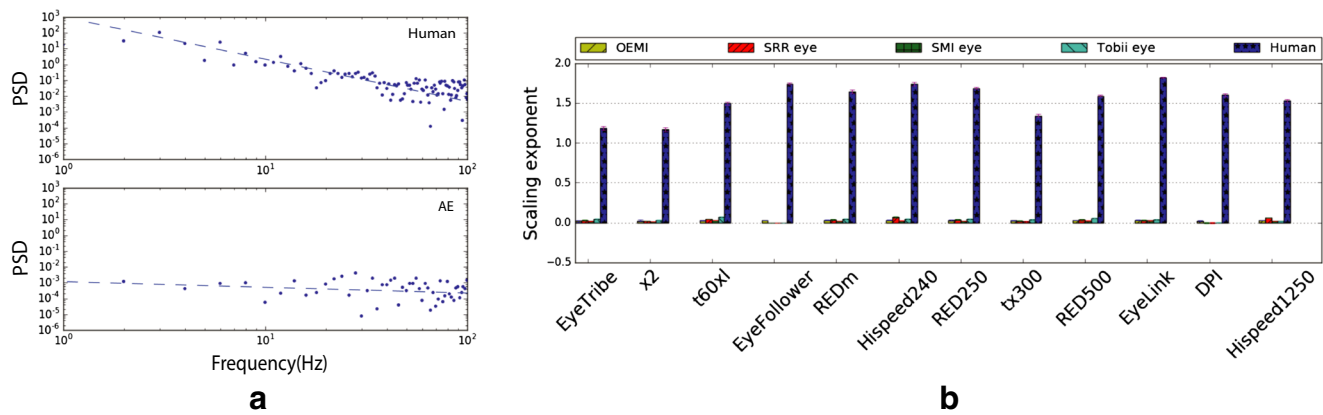


Fig. 8 (a) Example of power-spectral density (PSD) functions from humans and AEs. (b) Scaling exponents from AE and human data on all trackers. In panel a, the upper figure is human data recorded with SMI RED250, and the lower figure is the SMI AEs recorded with the same machine. The horizontal axis is frequency in log scale and the vertical axis is the PSD in log scale. The dashed lines represent the linear regression that fit the PSD in log scale. The negative of the slope of the regression line is the scaling exponent α . In panel b, the α level is an indication of

whether the signal shows stochastic noise (in which α would equals to 0), or pink/Brownian components with increasing value. Spatial-noise components differ with type of AE, as well as with the system. There is no repeating pattern in the AE values that can be attributed either to system or to AE model, but are likely some combination of both—or a result of system robustness to individual characteristics such as pupil size or pupil-iris contrast

greater or lesser extent, affected by properties of the (real or artificial) eye. For example, all four AEs rank Tobii T60XL in 4th or 5th place according to both precision metrics, whereas human data show a much lower ranking, at 10th with

RMS(S2S) and 7th with *SD*. Could this be a result of optimizing systems using artificial rather than human eyes? The AEs we tested provide an estimate of the relative precision of an eye-tracker, but the actual precision for a particular AE or eye

Table 2 System precision rankings with different AEs and human eyes, based on RMS(S2S) and *SD* and using a window size of 225 ms

(a) SRR AE, SMI AE and Tobii AE ranking of 10 eye-trackers

Ranking	SRR AE (n = 10)		SMI AE (n = 10)		Tobii AE (n=10)	
	RMS(S2S)	SD	RMS(S2S)	SD	RMS(S2S)	SD
1	SMI HS1250	SMI HS1250	EyeLink	EyeLink	SMI HS1250	SMI HS1250
2	EyeLink	EyeLink	SMI HS1250	SMI HS1250	EyeLink	EyeLink
3	SMI HS240	SMI HS240	SMI HS240	SMI HS240	SMI HS240	SMI HS240
4	TobiiT60XL	TobiiT60XL	SMI RED250	TobiiT60XL	TobiiT60XL	TobiiT60XL
5	SMI RED250	SMI RED250	TobiiT60XL	SMI RED250	SMI RED250	SMI RED250
6	SMI RED500	SMI RED500	SMI REDm	SMI REDm	SMI REDm	SMI RED500
7	SMI REDm	SMI REDm	SMI RED500	SMI RED500	SMI RED500	SMI REDm
8	TobiiTX300	TobiiTX300	TobiiTX300	TobiiTX300	TobiiTX300	TobiiTX300
9	EyeTribe	EyeTribe	EyeTribe	EyeTribe	EyeTribe	EyeTribe
10	TobiiX2	TobiiX2	TobiiX2	TobiiX2	TobiiX2	TobiiX2

(b) OEMI-7 and human eye ranking of 12 eye-trackers

Ranking	OEMI-7 (n = 12)		Human (n = 12)	
	RMS(S2S)	SD	RMS(S2S)	SD
1	DPI	DPI	DPI	DPI
2	SMI HS240	EyeLink	EyeLink	EyeLink
3	EyeLink	SMI HS1250	SMI HS240	SMI HS240
4	SMI HS1250	SMI HS240	SMI HS1250	SMI HS1250
5	TobiiT60XL	TobiiT60XL	SMI RED500	LCEyeFollower
6	SMI RED250	SMI RED250	SMI RED250	TobiiTX300
7	TobiiTX300	TobiiTX300	LCEyeFollower	TobiiT60XL
8	SMI REDm	SMI REDm	SMI REDm	SMI REDm
9	SMI RED500	SMI RED500	TobiiTX300	TobiiX2
10	LCEyeFollower	LCEyeFollower	TobiiT60XL	SMI RED250
11	EyeTribe	EyeTribe	TobiiX2	SMI RED500
12	TobiiX2	TobiiX2	EyeTribe	EyeTribe

Different trackers are printed in different colors to highlight variations

will depend further on system robustness to eye variance. We therefore conclude that the data recorded from currently existing AEs were, as expected, more precise, and can rank systems approximately, but not definitively, in terms of system noise.

The existing AEs from SMI, SR Research, and Tobii can be tracked by most video-based eye-trackers, but not by bright-pupil systems or the DPI. The modified OEMI-7 can be tracked on all of the eye-trackers we tested. If we were to propose a standardized set of AEs, the OEMI-7 has unique features that speak to its benefit. It provides a 1st and a 4th Purkinje reflection, as well as a retinal reflection. However, the material used to build the OEMI-7 (or any other AE currently in use) is fundamentally different from the human eye, resulting in several important differences in the features processed by eye-trackers. First, the brightness of the corneal reflection differs from those in humans. The brightness of the pupil under on-axis illumination also differs, which we could correct to some extent but not completely, and this might explain the poorer precision from the OEMI-7 on the EyeFollower than the average from human recordings. Also, the 4th Purkinje reflection's brightness was far higher than among humans, but our chosen solution for this is described in the [Appendix](#). Finally, the larger pupil size (7 mm) may also be a problem for some trackers, and although it was not modified here, it could be for a future set of AEs, with relevant characteristics systematically varied.

Despite the differences described, the OEMI-7 performed well when we compare the data from the tower-mounted eye-trackers: All the AEs resulted in similar precision values, close to the human precision values. For the noisier remotes, however, the AE data provided a broader distribution of precision values. The OEMI-7 eye, in particular, was an outlier for systems with a high RMS(S2S), which suggests that its construction is more problematic in terms of noise in the eye-tracker than the manufacturers' AEs. Furthermore, precision differed in the horizontal and vertical directions for some trackers when tracking with the OEMI-7, which may have been due to the large pupil size. The OEMI-7 has a 7-mm pupil, whereas human pupil size range from 2 to 8 mm with adaption, meaning that the OEMI-7 AE represents a model of a dark-adapted eye. The average human pupil size under room light viewing a monitor is about 5 mm, and perhaps the remote trackers cannot accommodate a 7-mm pupil. The OEMI-7 potentially allows for variation in pupil size and iris brightness, in that the iris is a single, detachable part of the model (though we did not change it in this study), and the liquid can be changed as a means of controlling the reflectivity of internal eye structures. A future customized version of the OEMI-7 with a different material and pupil size (or a 3-D-printed eye model) is worth investigation.

The finding of relatively white noise (with only slight variation in spectral qualities across systems) with the AEs versus

pink noise in the frequency domain in humans suggests a further role for AEs. One purpose of AEs for precision measurements might be to support investigations into the role of spectral noise. We found that although the precision values in RMS(S2S) and *SD* varied between eye-trackers, the results from the spectral analysis were quite consistent: All AEs elicit close to white noise on all eye-trackers. This result means that although different systems produce different noise levels [measured in terms of either RMS(S2S) or *SD*], they all produce mostly, but not exclusively, white noise, at least when the AE is directed to the center of the stimulus screen. Our results also clearly show that human data exhibit pink noise, as was previously reported for a single eye-tracker (Coey et al., 2012). We know that precision in terms of RMS(S2S) is of great importance to the detection of events such as fixations, saccades, and microsaccades, but the possible importance of PSD characteristics to the processing of eye movement data is less well-explored.

It is interesting, however, that the scaling exponent we found from human fixational data is much larger than in the study by Coey et al. (2012). A larger scaling exponent would indicate more energy in lower frequencies. One possible explanation is the presence of intrafixational eye movements. Ocular drift is smaller in amplitude than saccades and microsaccades, and much more frequent, and it provides a reference to support this claim. Another explanation might be that we had saccades of varying amplitudes to our central target in all human recordings, and the amplitude of the preceding saccade might have affected the subsequent fixational stability. In addition, the remote eye-trackers in our study are not usually used for recording microsaccades, because of their poorer precision and low sampling frequency. Another result is that the two eye-trackers with the poorest precision (the EyeTribe and the Tobii X2) also had the lowest scaling exponents—indicating the largest relative stochastic noise component. This suggests that in these trackers, the contribution of fixational eye movement to the spectral qualities was diminished, presumably due to relatively more system-based noise being present. Further studies will be needed to clarify the effects of spectral noise in event detection, and the relationship between the scaling component, eye properties, and precision.

Other than precision measurements, AEs are also the basis of robotic eye models that simulate human eye movement and test the spatial accuracy of eye-trackers recording the eye in motion, particularly in terms of faithful representation of saccade or smooth-pursuit dynamics. Measuring system latencies also requires AEs that can function with different systems and provide some kind of reproducible motion. Building AEs that vary in their eye shape, color, and pupil size will also be useful for testing the robustness of eye-trackers to varying eye characteristics.

As a general methodological remark, we note that the calculation of precision as RMS(S2S) values should be used with

caution, because they are biased toward higher sample rates, as was previously pointed out by Blignaut and Beelders (2012). *SD* measures of precision depend less on sample rate. Our results show that increasing window sizes lead to higher values of both RMS(S2S) and *SD*. This holds especially true for human data, for which both values increase up to five times when the window size increases from 100 to 500 ms. Therefore, it is important for reports of precision to describe the window size. Preferably, a standard window size would be defined for the comparison of systems.

Finally, our data also show how calibration with one eye model and recording with another can affect precision. In five out of 12 cases, tower-mounted trackers calibrated to a human eye before recording the AE had significantly higher (poorer) RMS(S2S) and *SD* values than recordings based on calibrating with the same AE, indicating that a change in eye image properties between calibration and recording affects precision in addition to accuracy. When we measured the precision of the EyeLink system with the AE supplied by SR Research, calibrating to the human eye elicited *lower* RMS(S2S) and *SD* values (better precision) than calibrating with the AE. In the remaining six cases, we found no significant difference in the precisions recorded from the AEs, whether calibration had been performed using it or a human eye. This variation may be due to the mapping functions, which map the pupil–corneal reflection vectors to the corresponding gaze locations. When recording using different eyes (whether artificial or real), the mapping function created for the eye calibrated is suboptimal for the eye used during subsequent recording, and small errors can be amplified, leading to decreased precision. If, however, the mapping functions are simply offset, there would be no significant effect on precision, only on accuracy.

We conclude that developing a standard set of complex AEs, in which pupil size and pupil–iris contrast would be systematically varied, is a promising approach toward the comparative measurement of system performance.

Author note The authors acknowledge the collaboration and input of members of the EMRA/COGAIN Eye Data Quality Standardization Committee at all stages of this research, www.eye-movements.org/eye_data_quality.

Appendix: Modification of OEMI-7

The OEMI-7 is a commercially available eye model that is designed for ophthalmic imaging training. It can be purchased online (Ocular Instruments, 2013) for approximately \$600. The cornea and lens of the OEMI-7 eye model is made of *para*-methoxy-*N*-methylamphetamine (PMMA). The refractive index of PMMA is 1.485 at 750 nm. The posterior and anterior chamber as shown in Fig. 9, are filled with distilled water. Given the refractive index of lens cortex and vitreous

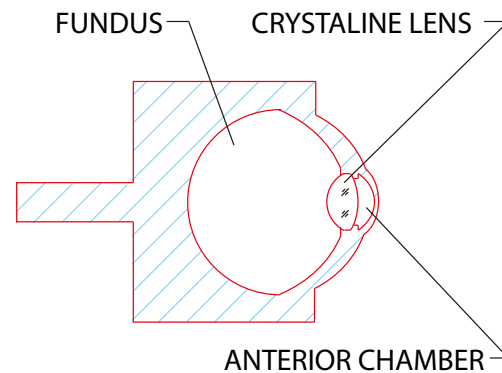


Fig. 9 Structure of OEMI7. (Picture from Ocular Instruments, 2013)

humor (Hecht, 1987), the reflectance at normal incidence at the back surface of the human lens can be calculated using the Fresnel Equation (Hecht, 1987, p. 100):

$$R = \left(\frac{n_2 - n_1}{n_2 + n_1} \right)^2 = \left(\frac{1.337 - 1.406}{1.337 + 1.406} \right)^2 = 0.000632,$$

where n_1 and n_2 are the refractive index of the two media, respectively. The reflectance at the back surface of the lens of the eye model equals to 0.00346 using the same formula, which means the eye model has a much brighter 4th Purkinje image than a real human eye, reflecting more than five times as much light. To render the eye model more similar to a human eye to the eye-trackers, we replaced the liquid with a solution with a refractive index of 1.412 filling the posterior and anterior chambers. A 60% glycerine–water solutions has a refractive index of 1.413 at 20°, so we carefully changed the liquid to a 60% glycerine–water solution.

References

- Abramov, I., & Harris, C. M. (1984). Artificial eye for assessing corneal-reflection eye trackers. *Behavior Research Methods, Instruments, & Computers*, 16, 437–438.
- Artal, P., Benito, A., & Tabernero, J. (2006). The human eye is an example of robust optical design. *Journal of Vision*, 6(1), 1–7. doi:10.1167/6.1.1
- Barnes, J., & Allan, D. (1966). A statistical model of flicker noise. *Proceedings of the IEEE*, 54, 176–178.
- Blignaut, P., & Beelders, T. (2012). The precision of eye-trackers: A case for a new measure. In *Proceedings of the symposium on eye tracking research and applications* (pp. 289–292).
- Carpenter, R. H. (1988). *Movements of the eyes* (2nd rev.). London, UK: Pion.
- Cleveland, D., Mulvey, F., & Pelz, J. B. (2015, August). *Revisiting eyetracker noise estimators*. Paper presented at the ECEM 2015 Conference, University of Vienna, Austria.
- Coey, C. A., Wallot, S., Richardson, M. J., & Van Orden, G. (2012). On the structure of measurement noise in eye-tracking. *Journal of Eye Movement Research*, 5(4), 1–10. doi:10.16910/jemr.5.4.5
- Crane, H. D., & Steele, C. M. (1985). Generation-V dual-Purkinje-image eyetracker. *Applied Optics*, 24, 527–537.

- Findlay, J. (1971). Frequency analysis of human involuntary eye movement. *Kybernetik*, 8, 207–214.
- HDFGroup (2014). Hdf5-1.8.14. Retrieved 2014-11-13, from www.hdfgroup.org/HDF5/
- Hecht, E. (1987). *Optics* (2nd ed.). New York, NY: Addison-Wesley.
- Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford, UK: Oxford University Press.
- Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye tracker data quality: What it is and how to measure it. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 45–52). New York, NY: ACM Press. doi:10.1145/2168556.2168563
- iSolver Software Solutions (2013). Io-hub 0.7. Retrieved 2014-10-10, from www.isolver-solutions.com/iohubdocs
- Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science*, 39, 447–462.
- Mulvey, F., Landwehr, N., Borah, J., Cleveland, D., Joos, M., Latorella, K., ... Wang, D. (2015). *How reliable is my eye-movement data? Results of system comparison and participant characteristics-based prediction of data quality from the EMRA/Cogain Eye Data Quality Standardisation Committee*. Paper presented at the ECEM 2015 Conference, University of Vienna, Austria.
- Nyström, M., Andersson, R., Holmqvist, K., & van de Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, 45, 272–288. doi:10.3758/s13428-012-0247-4
- Ocular Instruments (2013). Ocular imaging eye model and bracket. Retrieved 2014-10-10, from www.ocularinc.com/
- Peirce, J. W. (2007). Psychopy—Psychophysics software in Python. *Journal of Neuroscience Methods*, 162, 8–13. doi:10.1016/j.jneumeth.2006.11.017
- Reingold, E. M. (2014). Eye tracking research and technology: Towards objective measurement of data quality. *Visual Cognition*, 22, 635–652.
- Scargle, J. D. (1982). Studies in astronomical time series analysis: II. Statistical aspects of spectral analysis of unevenly spaced data. *Astrophysical Journal*, 263, 835–853.
- Wallot, S., Coey, C. A., & Richardson, M. J. (2015). Cue predictability changes scaling in eye-movement fluctuations. *Attention, Perception, & Psychophysics*, 77, 2169–2180. doi:10.3758/s13414-015-0983-5