EVALUATING THE DATA QUALITY OF EYE TRACKING SIGNALS FROM A VIRTUAL REALITY SYSTEM: CASE STUDY USING SMI'S EYE-TRACKING HTC VIVE

A PREPRINT

Dillon J. Lohr
Department of Computer Science
Texas State University
San Marcos, TX 78666, USA
djl70@txstate.edu

Lee Friedman
Department of Computer Science
Texas State University
San Marcos, TX 78666, USA
l_f96@txstate.edu

Oleg V. Komogortsev
Department of Computer Science
Texas State University
San Marcos, TX 78666, USA
ok11@txstate.edu

December 4, 2019

ABSTRACT

We evaluated the data quality of SMI's tethered eye-tracking head-mounted display based on the HTC Vive (ET-HMD) during a random saccade task. We measured spatial accuracy, spatial precision, temporal precision, linearity, and crosstalk. We proposed the use of a non-parametric spatial precision measure based on the median absolute deviation (MAD). Our linearity analysis considered both the slope and adjusted R-squared of a best-fitting line. We were the first to test for a quadratic component to crosstalk. We prepended a calibration task to the random saccade task and evaluated 2 methods to employ this user-supplied calibration. For this, we used a unique binning approach to choose samples to be included in the recalibration analyses. We compared our quality measures between the ET-HMD and our EyeLink 1000 (SR-Research, Ottawa, Ontario, CA). We found that the ET-HMD had significantly better spatial accuracy and linearity fit than our EyeLink, but both devices had similar spatial precision and linearity slope. We also found that, while the EyeLink had no significant crosstalk, the ET-HMD generally exhibited quadratic crosstalk. Fourier analysis revealed that the binocular signal was a low-pass filtered version of the monocular signal. Such filtering resulted in the binocular signal being useless for the study of high-frequency components such as saccade dynamics.

Keywords Eye tracking · Data quality · HTC Vive · Head-mounted display · HMD · Virtual reality · VR · Spatial accuracy · Spatial precision · Temporal precision · Linearity · Crosstalk

1 Introduction

We recently acquired an eye-tracking virtual reality (VR) head-mounted display (HMD): SMI's tethered system based on the HTC Vive (ET-HMD). Our overall goal of this work was to fully and carefully characterize the data quality produced by this device. Many similar devices are already, or will soon be, on the market (e.g., FOVE, VIVE Pro Eye, Varjo VR-1). Our methods and results may be of interest to other researchers working with such devices.

One important reason for a detailed analysis of the data quality of eye trackers is that the manufacturers' published specifications are often not achievable in practice [7, 21, 39]. Studies of eye movements in reading need to know, as accurately as possible, which exact word is being read and even which letter is being fixated [42, 43]. The use of eye position as a selection device for options on a web page also seems like an example where very accurate tracking would be required. In the context of oculomotor biometrics (the effort to identify individual humans based on their eye movements) [13, 19], improved performance of the eye-tracking device is likely to lead to an improved estimate of inter-human variation. The need for increased data quality may be the motivation behind the recent increase in the number of publications including the words "eye," "tracker," and "quality" (Fig. 1).

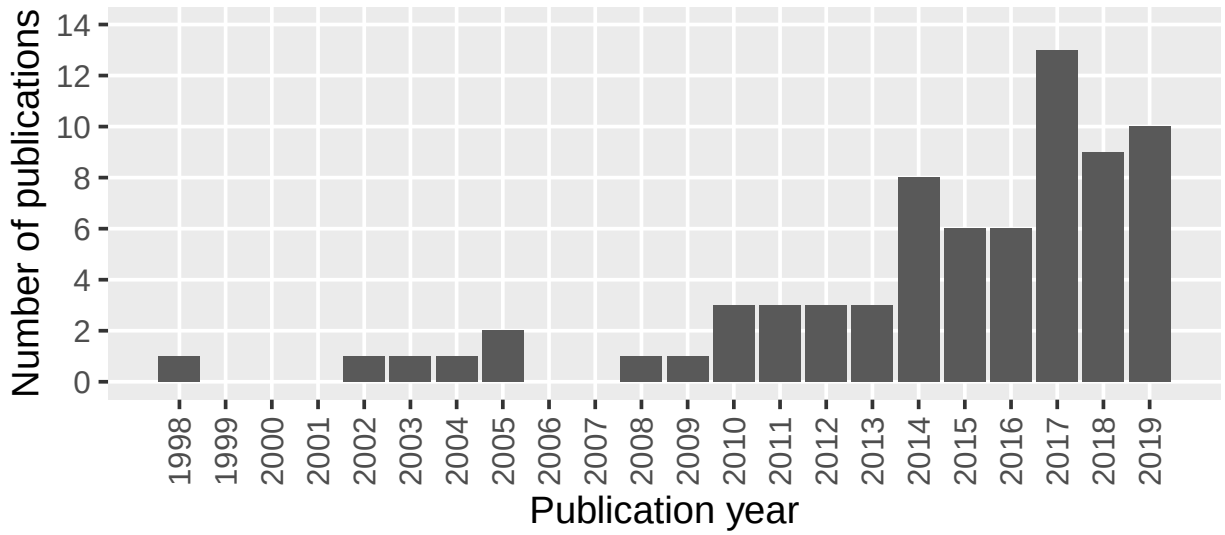

Figure 1: The number of articles in the PubMed database containing the words "eye," "tracker," and "quality" as of November 7, 2019. Papers are grouped by the year they were added to PubMed. Partial data for 2019 (approximately 10 months).

When we refer to data quality, we are specifically referring to measures of spatial accuracy, spatial precision, temporal precision, linearity, and crosstalk, each defined in Table 1 below. Hooge et al. [23] argued in favor of using the term

Table 1: Definitions of Data Quality Terms Used in This Paper

Term	Definition
Spatial accuracy	The ability of the eye tracker to correctly measure the gaze position
Spatial precision	The ability of the eye tracker to reliably reproduce a gaze position measurement
Temporal precision	The ability of the eye tracker to produce new gaze position measurements at a consistent rate
Linearity	The consistency of spatial accuracy across the spatial range of the recording device
Crosstalk	The extent to which movement in one direction (horizontal or vertical) artifactually influences movement in the orthogonal direction

"systematic error" instead of "spatial accuracy" and the term "variable error" instead of "spatial precision." Their view was that it is counterintuitive for a *high* accuracy to be characterized by a *small* systematic error, and likewise for a *low* precision to be characterized by a *large* variable error. Indeed, it may be confusing, especially for new researchers, that a *low* spatial accuracy actually means that the calculated value is *large*. Although we agree with the semantic argument by Hooge et al. [23], we will use the terms "spatial accuracy" (or just "accuracy") and "spatial precision" (or just "precision") to be consistent with the majority of the literature.

Regarding spatial precision, Hornof and Halverson [24] suggested that:

> "Variable error can be seen in the spread (or *dispersion*) of the recorded gazepoints around the actual fixation maintained by the human. Systematic error is the disparity (or *drift*) between the average gazepoint location and the actual fixation maintained by the human. Variable error indicates a lack of *precision*. Systematic error indicates a lack of *accuracy*" (p. 592).

We disagree with a formulation of precision as a measure of spread and accuracy as a measure of distance. It defines accuracy and precision as two fundamentally different types of things. We prefer to think of both accuracy and precision as average-distance measures with different reference points. Accuracy should be defined as the mean (or median) of the distances between each sample and the target position. Precision should be defined as the mean (or median) of the distances between each sample and the central tendency of all samples. That is, the reference point for accuracy is the target position, while the reference point for precision is the central tendency of the samples.

1.1 Historical efforts to assess eye-tracker quality

Assessment of the accuracy of eye-movement measuring equipment is probably as old as the measurement of eye movements.

In the first paper to employ a corneal reflection and the first paper to make reasonably accurate estimates of the average velocity of saccades, Dodge and Cline [12] employed a photographic method to measure eye movements. The method involved recording eye movements on a photographic plate that was falling at a rate which was constant within each recording but varied from recording to recording. Using a pendulum as a time source, Dodge and Cline were able to determine that the average error in temporal precision was 0.5 ms. In the paper introducing the scleral search coil method, Robinson [48] evaluated his new system and found that the device had an accuracy and linearity of about 2 percent of full scale, a spatial resolution of 15 seconds of arc, and a bandwidth of 1000 Hz. Merchant [36], in one of the earliest reports to use the video-oculography (VOG) method to estimate eye position using pupil position and corneal reflection position, found his device to have a sample-to-sample root mean square error (RMS) noise level of 0.3° and channel crosstalk of 20% maximum. However, Merchant's system was substantially nonlinear with respect to eye position. In the paper introducing the dual-Purkinje-image (DPI) tracking method, Cornsweet and Crane [10] reported that the device had a spatial resolution and spatial accuracy of about 1 minute of arc. Collewijn et al. [9] developed a more comfortable scleral search coil method and reported that the noise level of the system at maximum sensitivity was equivalent to about 0.5 minutes of arc. In an early general discussion of the topic of recording quality, McConkie [35] encouraged standardization in the reporting of data quality for eye movement research. He also described a method of measuring drift in a recording device with an artificial eye. Reulen et al. [46] developed an advanced infrared limbus tracker and provided an early example of a reasonably comprehensive performance evaluation, including measures of spatial resolution, spatial precision, linearity, and crosstalk. Drift was assessed qualitatively.

1.2 Current state of the field

1.2.1 Spatial accuracy

There seems to be little debate regarding the definition of spatial accuracy (see Table 1), although there is some variability in the exact term (e.g, "systematic error" or "offset error" [23, 24, 54]) used to label this quality.

The first step in the measurement of accuracy is the determination of which data points to include in the calculation, and there are various approaches to this. For example, Blignaut and Wium [6] ignored the first 1000 ms of data after presenting a target, then used the next 500 ms. Akkil et al. [2] had two parts to their study: a system-controlled routine and a participant-controlled routine. For the system-controlled routine, Akkil et al. ignored the first 500 ms of data after presenting a target and used the next 500 ms of data. For the participant-controlled routine, they identified the moment when the mouse cursor made the final approach to the target and collected roughly 500 ms of data. Blignaut and Beelders [5] used data in the period immediately before a mouse-up event during a participant-controlled routine, with the duration of the period typically between 150-300 ms. Kasprowski et al. [27] ignored the first 700 ms of data after presenting a target, then used the next 1100 ms. Blignaut et al. [7] used 250 ms of data preceding a mouse click. Nyström et al. [39] ignored the first 400 ms of data after presenting a target, then used multiple criteria, including temporal and spatial criteria, for determining which samples to include in their calculation.

There is also variation in the quantity and position of targets used. The most commonly used target arrangement is a square grid containing 9 to 25 points [27], but even these somewhat standard arrangements vary in their visual angle span. On one end of the spectrum, Akkil et al. [2] used a 4-point arrangement with each point located 20% of the screen dimension away from the corners. On the other end, Blignaut et al. [7] employed a 40-point arrangement (8 columns and 5 rows) spanning 38.4° horizontally and 24° vertically.

1.2.2 Spatial precision

Aside from the obvious benefit of precise measurement, the spatial precision of a device can influence experimental results from human studies such as estimated fixation durations [22, 45], event detection [39], or investigating imperfections in the oculomotor system [39]. Spatial precision can be measured with an artificial eye [33, 44, 52, 56] or from a human recording [6, 33, 52, 56], but Holmqvist et al. [22] claim it is better to use both artificial eyes and human subjects when assessing spatial precision. Using an artificial eye provides a spatial precision measurement without the complication of human oculomotor noise [22], including drift, tremor, and microsaccades [32, 34]. However, as noted by Holmqvist et al. [22], artificial eyes do not have the same iris, pupil, and corneal reflection features as human eyes. Also, human eyes can vary greatly (e.g., eye color, pupil size, corneal size, degree of corneal bulge). Therefore, testing with real eyes can provide precision data for a realistic population of potential test subjects.

Spatial precision is frequently measured either as RMS or as an estimate of dispersion among a set of samples [22, 52]. Holmqvist et al. [22] show how RMS measures are more resistant to vibrations in the environment than dispersion as assessed with a standard deviation (SD). However, Blignaut and Beelders [4] state that a disadvantage of using RMS is that it varies with sampling rate. Given the time-series nature of the data stream, there is an implied temporal dependency in the calculation of RMS. That is, RMS is a function of the ordering of the samples. If one randomly shuffled the sample positions, one would in all likelihood get a different RMS.

In a more thorough overview of spatial precision measures, Holmqvist et al. [21] provided eight additional measures of spatial precision. Blignaut and Beelders [4] briefly critiqued these eight measures, rejected four of them, considered three to be overly complex and difficult to interpret, and provided a detailed evaluation of RMS, SD, and bivariate contour ellipse area (BCEA). BCEA defines the area of an ellipse that encompasses some proportion of a set of points [21]. Blignaut and Beelders [4] reduced BCEA to a one-dimensional value they called r(BCEA) by approximating the ellipse with a circle. They proposed that the use of r(BCEA) is intuitive, independent of sampling rate (unlike RMS), and independent of the arrangement of samples within a fixation.

1.2.3 Temporal precision

For most eye-movement recording devices, the nominal interval between samples is measured in milliseconds, and frequently only the nominal sampling rate is available to the user. This is certainly the case with our EyeLink 1000 (SR-Research, Ottawa, Ontario, CA), which has a nominal sampling rate of 1000 Hz and, according to staff at SR Research, produces samples precisely 1 ms apart. The ET-HMD has a nominal sampling rate of 250 Hz, but it provides timestamps for samples with nanosecond precision. (We are unaware of how common such precise timestamps are.) When we use the term "temporal precision," we mean the variation of the nominal sampling rate estimated by sub-millisecond timestamps. This definition of temporal precision was used by Abdulin et al. [1] (using timestamps precise to 10^{-7} seconds) in the evaluation of their custom eye-tracking device. As we will discuss in Sect. 2.2.3, others use this term in a different context, befitting gaze-contingent research.

1.2.4 Linearity

Several studies have noticed that spatial accuracy can be a function of target position [5, 7, 24, 39, 46]. Since at least 1975 [57], this has often been referred to as "linearity." Although this term for a dependence of accuracy on target position has not always been referred to as linearity, the relationship has been studied repeatedly in the literature.

Reulen et al. [46] used the term linearity and specifically measured it by fitting a line to a scatter plot relating measured eye position to target position (both horizontal and vertical). Residuals from these lines were expressed as a percentage of the position range. Linearity was expressed as the maximum residual. For their device, linearity was 3% for the horizontal direction and 2% for the vertical direction. In our view, it seems that such a measurement would be highly sensitive to outliers. Blignaut and Beelders [5] primarily used a graphical analysis to illustrate the linearity of their device, though they did not specifically use the term linearity. We think this is a good approach, but linearity could be further characterized statistically.

1.2.5 Crosstalk

In some of the earlier references to crosstalk (also known as "cross-coupling"), little detail is provided on the precise calculation. For example, the first formal measure of crosstalk in eye movements that we are aware of was by Merchant [36] who reported a crosstalk of 20% maximum but provided no further details. Similarly, Young and Sheena [57] discussed crosstalk as a problem with electrooculography (EOG) without indicating the measurement method.

To describe the approach employed by Reulen et al. [46] for the measurement of crosstalk, consider horizontal crosstalk as an example. First, they plotted both the horizontal and vertical eye-movement signals against the vertical target signal. Next, they performed a linear regression on both eye-movement signals. This produced two slopes: one for horizontal eye position versus vertical target position, and one for vertical eye position versus vertical target position. Horizontal crosstalk was then expressed as the ratio of the horizontal-vertical slope to the vertical-vertical slope. Analogous measurements were made for vertical crosstalk. They found that this ratio was about 1/10 for their device, meaning an eye rotation of 10° in one direction caused 1° crosstalk in the orthogonal direction. This approach assumed crosstalk was linear.

Working with a simulated model, Rigas et al. [47] measured crosstalk as the absolute ratio (expressed as a percent) between the observed movement in one direction (horizontal or vertical) and the ground truth movements in the orthogonal direction. Their approach provided no information on the shape of the crosstalk effect (e.g., linear, quadratic, or cubic).

1.2.6 Recalibration

Generally, researchers use a manufacturer-supplied calibration routine (MSC) at the start of each recording (though, there are efforts to make calibration-free eye-tracking devices [28, 38]). Many researchers have started providing their own user-supplied calibration routines (USCs) after the MSC and before task data is collected. Typically, during a USC, between 4 and 40 calibration targets are presented, distributed over some horizontal and vertical range. However, some methods do not require calibration targets at all [51]. USCs that do include targets determine when the subject is fixated on a target using one of three approaches [14]: algorithm-controlled [2], operator-controlled [39], or participant-controlled [7, 23, 39].

After selecting samples during stable fixation, the actual recalibration of eye position takes place, and there are several methods for this. One approach is to use a simple linear mapping, but this does not account for interactions between the two dimensions [3]. More complex polynomial mappings can also be done [3, 27], with the caveat that more calibration targets (and, thus, more time) are required to fit the polynomials. Others create a 3D eye model to fit the measured eye movements [16, 37]. Another approach is to perform Procrustes analysis [49], where the eye-position samples are translated, scaled, and rotated to fit the target positions.

Another issue is the time between the USC and task-related data collection. There is reason to believe that calibrations may deteriorate over time [24], and therefore the shorter this interval, the better. In the present study, our USC was part of our experimental task, and so there was no delay between the collection of calibration data and the collection of the experimental data.

1.2.7 Filtering

There is a general awareness of the potential influence of manufacturer-supplied filters of gaze position signals. For example, in a study of spatial accuracy ("systematic error") and spatial precision ("variable error"), Hooge et al. [23] stated:

> "The values for the decrease in the systematic error, and the increase in the variable error here presented, may be specific to the eye tracker, the experimental conditions and the populations we employed. Values may be different for different eye trackers (for example because some manufacturers apply filtering to the raw data to decrease the variable error), ..."

Also, there is general knowledge about the potential negative effects of filtering on eye-movement signals. For example, Reingold [44] made the following points:

> "The emphasis on the tracking of "stationary" biological or artificial eyes as a primary method for eye tracker data quality evaluation has the unfortunate consequence of creating an incentive for manufacturers to produce systems that use heavy filtering (i.e., denoising algorithms) that, while making the eye look stable during fixations, severely distort the eye movement signal in terms of the velocity profile of the motion. Although appearing to improve static accuracy, it is often not appreciated that such filtering destroys important aspects of data quality including the temporal accuracy of identifying the beginning and end of fixations, the number of fixations detected (see Holmqvist et al., 2012), the kinematics of saccadic eye movements, the ability to detect small saccades, and eye movements produced while looking at dynamic stimuli (e.g., smooth pursuit)."

For other comments on the negative effects of filtering, see Kolarik et al. [29] and Nyström et al. [39]. Therefore, we thought that it might be important to check for, and assess, any potential pre-filtering of our signals. To this end, we performed a Fourier analysis of our various signals. We propose that this sort of analysis should be considered as a part of any comprehensive assessment of eye-tracker quality.

1.3 Present experimental plan

In the present report, we applied new and potentially useful analyses to the characterization of the ET-HMD. Of course, we assessed spatial accuracy, spatial precision, temporal precision, linearity, and crosstalk. All of our assessments were based on the analysis of a random saccade task. We introduced a method to remove the saccade latency for analysis of fixations during calibration, providing more data for analysis. For spatial precision, we employed the non-parametric median absolute deviation (MAD) rather than the SD so that our measure was robust to various underlying distributions. In addition to a nominal sampling rate of 250 Hz, the ET-HMD provides actual timestamps with nanosecond precision. We characterized temporal precision in the present study as the SD of intersample intervals (ISIs). Our user-supplied calibration was prepended to our task to minimize the time between recalibration and the recording of

experimental data. We presented a novel and sound "binning" method for the selection of samples within fixations to include in recalibration. For linearity, we fit a line to the relationship between target position and eye position and described the degree of linearity with an adjusted R-squared (R^2_{adj}) from a linear regression analysis. Since an ideal system would have a slope of 1.0 for this fit, we compared the measured slope and its confidence limits to this ideal. For our crosstalk analysis, we fit 4 multiple regression models (linear only, quadratic only, both linear and quadratic, and intercept only) and chose the best model using the Akaike information criterion (AIC). The ET-HMD provides 3 signals (a left eye signal, a right eye signal, and a binocular signal) and all of these were compared on all quality metrics. We also provided a Fourier analysis of the binocular and monocular signals and noted that the binocular signal is a substantially low-pass filtered version of the monocular signal.

2 Methods

2.1 Participants

Twelve participants (9 males, 3 females, median age: 20, range: 19-66) willingly took part in this study. Five participants normally wore glasses but removed them for this study, three participants wore contact lenses during the study, and four required no vision correction.

2.2 Definitions of data quality

2.2.1 Spatial accuracy

Following Holmqvist et al. [22], this is a measurement of the distance (in degrees of the visual angle) between the actual gaze position (where the subject is actually looking) and the measured gaze position (where the eye tracker reports the subject is looking) during stable fixation. For the present purposes, once a subject has fixated on a target, we assume the subject is actually looking at the target and that the target position can be treated as the actual eye position. A separate measurement may be taken from the horizontal and vertical distances, or a single, combined measurement for both directions may be obtained (e.g., by using the Euclidean distance). The dependent variable for statistical analyses of spatial accuracy is the mean accuracy per subject across fixations.

Consider a series of n gaze samples recorded during a stable fixation. Each sample includes a measured gaze position (x_i^g, y_i^g) and a target position (x_i^t, y_i^t), each in units of degrees of visual angle. Note that since the target position remains constant during stable fixation, (x_i^t, y_i^t) is the same for all i. The spatial accuracy of the series of gaze samples is calculated using one of the following equations:

$$\theta_H = \frac{1}{n}\sum_{i=1}^{n} |x_i^g - x_i^t| \tag{1a}$$

$$\theta_V = \frac{1}{n}\sum_{i=1}^{n} |y_i^g - y_i^t| \tag{1b}$$

$$\theta_C = \frac{1}{n}\sum_{i=1}^{n} \sqrt{(x_i^g - x_i^t)^2 + (y_i^g - y_i^t)^2} \tag{1c}$$

where θ_H is horizontal accuracy, θ_V is vertical accuracy, and θ_C is combined accuracy (taking into account both horizontal and vertical components).

2.2.2 Spatial precision

Although some define spatial precision as the variation in the measured gaze position signal [4], as noted above, we prefer to think of both accuracy and precision as average-distance measures with different reference points. For accuracy, the reference point is the target position. For precision, the reference point is the central tendency of a set of sample positions. As with spatial accuracy, spatial precision must be measured during stable fixation. A separate measurement may be taken across the horizontal and vertical position signals, or a single, combined measurement may be obtained (e.g., by using the Euclidean distance to the central value). The dependent variable for statistical analyses of spatial precision is the mean precision per subject across fixations.

Spatial precision is commonly measured in one of two ways, using either the SD of sample positions or the RMS (see Equations 2 and 3 in Holmqvist et al. [22]). Both of these approaches give a higher weight to larger deviations than to smaller ones due to the squaring of deviations, which is not necessarily desirable [15]. Therefore, we have chosen a

spatial precision measurement based on the MAD of sample positions. This measure gives equal weight to deviations of all sizes and is robust to various shapes of underlying distributions (e.g., skewed or non-Gaussian).

Consider a series of n gaze samples recorded during a stable fixation. Each sample includes a measured gaze position (x_i^g, y_i^g). We denote the geometric median of the sample positions as the point $(\tilde{x^g}, \tilde{y^g})$ (computed using R package `Gmedian`, Cardot [8]). The spatial precision of the series of gaze samples is calculated using one of the following equations, where $M(x)$ is short for *median*(x):

$$\text{MAD}_H = M(|x_i^g - M(x^g)|),\ i = 1 \ldots n \tag{2a}$$

$$\text{MAD}_V = M(|y_i^g - M(y^g)|),\ i = 1 \ldots n \tag{2b}$$

$$\text{MAD}_C = \sqrt{M(|x_i^g - \tilde{x^g}|)^2 + M(|y_i^g - \tilde{y^g}|)^2},\ i = 1 \ldots n \tag{2c}$$

where MAD_H is horizontal precision, MAD_V is vertical precision, and MAD_C is combined precision (taking into account both horizontal and vertical components).

2.2.3 Temporal precision

In one sense, temporal precision can be defined as the variability of ISIs (we call this sense the "ISI sense"). In another sense, as Holmqvist et al. [22] have suggested, temporal precision can be thought of as the variability in "system latency," or the difference between the time of the actual movement of the eye and the time reported by the eye tracker (we call this sense the "Holmqvist sense"). It is easy to see that temporal precision in the Holmqvist sense is nearly perfect with analog systems such as EOG, infrared limbus tracking, and scleral search coil. Although the ET-HMD does produce timestamps for every video frame with nanosecond precision, there is some variability in the intersample intervals. We have no ground-truth measurement of when eye movements are initiated, so we cannot measure temporal precision in the Holmqvist sense. We can, and do, measure temporal precision in the ISI sense.

The nominal sampling rate of the ET-HMD is 250 Hz, but the ET-HMD does not always achieve precise 4 ms intervals between timestamps.[1] The nanosecond timestamps were converted to milliseconds before calculating the ISIs.

Consider a series of n gaze samples. Each sample includes a timestamp t_i measured in milliseconds. An ISI is the difference between consecutive timestamps:

$$\Delta t_i = t_i - t_{i-1},\ i > 1 \tag{3}$$

We compute temporal precision as the sample SD of ISIs.

2.2.4 Linearity

When we discuss the term linearity, we are referring to the relationship between measured gaze position and actual gaze position. There are at least two senses of the term linearity relevant here. In one sense (hereafter referred to as "linearity fit"), we want to know how well the data fit a line. For this, we use the R^2_{adj}. In another sense (hereafter referred to as "linearity slope"), we want to know how close the slope of the relationship is to an ideal slope of 1.0. For this, we look at the measured slope and its 95% confidence limits. The dependent variable for statistical analyses of linearity slope is the slope estimate per subject across fixations. The dependent variable for statistical analyses of linearity fit is the R^2_{adj} per subject across fixations.

In Figure 2, we present a linearity analysis for a single subject for illustrative purposes. To assess horizontal linearity, we plot the horizontal gaze position versus horizontal target position. Note that only one point is drawn per fixation target, because we perform our linearity regressions using the gaze centroid for each fixation instead of the individual gaze samples. Both the linearity slope and linearity fit are shown in annotations on the figure. In this case, the linearity slope is significantly different from the ideal slope of 1.0, because the 95% confidence intervals do not include 1.0.

2.2.5 Crosstalk

This is a measure of how much the rotation of the eye in one direction (horizontal or vertical) affects the measured gaze position in the orthogonal direction. We refer to horizontal crosstalk from vertical movements as "horizontal crosstalk," and to vertical crosstalk from horizontal movements as "vertical crosstalk."

[1] In e-mail correspondence with our former contacts at SMI, we asked, "What is the general stability of the sampling rate of the eye tracker? Are there ever any dropped frames?" One contact responded, "It's possible, yes. Unlike some of our other products, the ET-HMD has to use the computer for image processing and eye tracking, so a resource problem could lead to dropped frames. I wouldn't expect this to be common unless there is a problem."

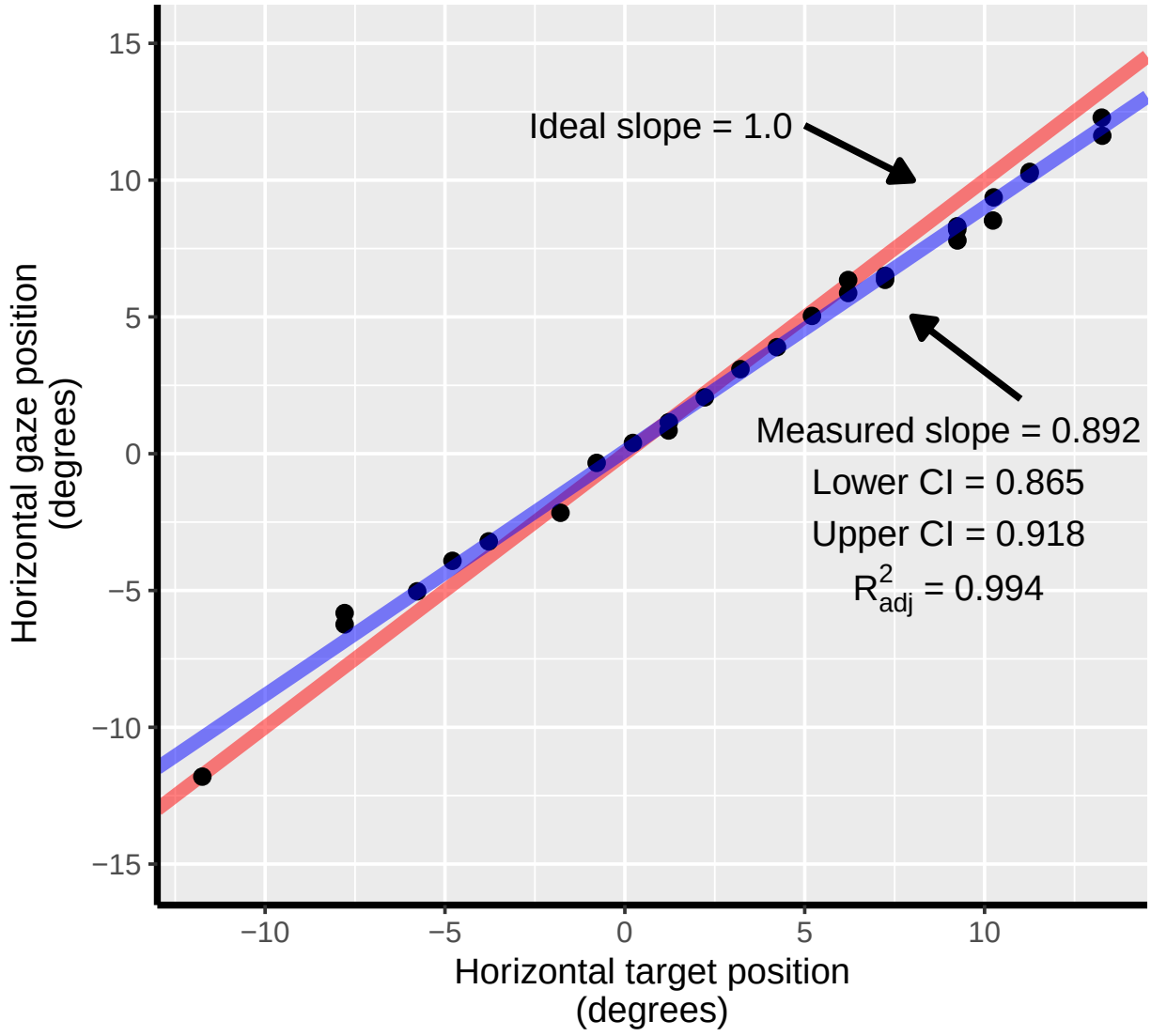

Figure 2: An example of linearity for illustrative purposes.

The crosstalk measurements used in the past assume a linear relationship [46, 47], but we thought that it might be valuable to also consider the possibility of the crosstalk having a parabolic shape. For data to assess crosstalk, we used fixation positions from our random saccade task, which sampled more-or-less the target movement range. Our crosstalk assessments emerged from regression models, with the dependent measure as the offset of gaze position (in one direction, say horizontal) from target position (also horizontal in this case) and the independent variable was the target position in the orthogonal direction (vertical in this case). All models fit an intercept. Gaze offset is a similar concept to accuracy. However, accuracy is a mean of the absolute value of distances, whereas offset is a mean value of distances which can be positive or negative.

For our crosstalk assessments, we used a step-wise approach and chose the best fitting model from the following 4 models: (1) a linear only fit, (2) a quadratic only fit, (3) both a linear and quadratic fit, and (4) an intercept only fit (neither linear nor quadratic). When assessing the crosstalk for a single subject, the best model was decided by the `stepAIC` function of the R package `MASS` [55]. When assessing the crosstalk across subjects, the best model was decided by the `step` function of the R package `lmerTest` [31].

In Figure 3, we present a crosstalk analysis for a single subject for illustrative purposes. Note that only one point is drawn per fixation target, because we fit our crosstalk models using the gaze centroid for each fixation instead of the

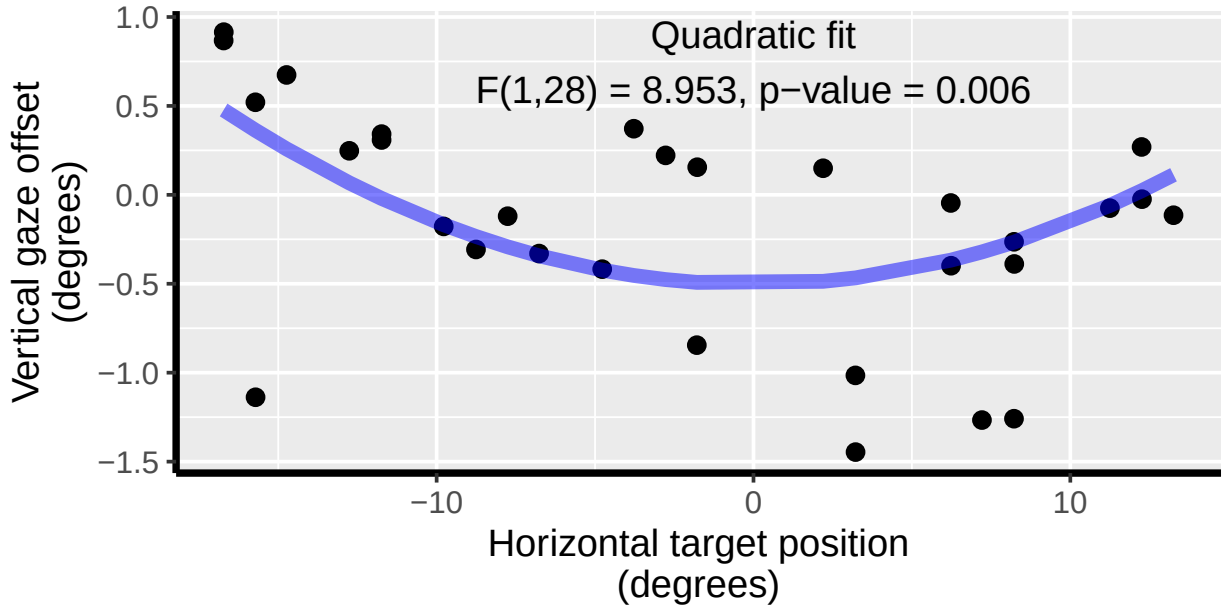

Figure 3: An example of crosstalk for illustrative purposes.

individual gaze samples. To assess vertical crosstalk, we plotted the offset of the gaze centroid from the vertical target position versus horizontal target position. In this case, the quadratic-only model was the best fit.

2.3 Eye-tracking hardware

2.3.1 ET-HMD

The ET-HMD is a modified HTC Vive with an embedded SMI eye-tracking device. The HTC Vive has dual Active-Matrix Organic Light-Emitting Diode (AMOLED) screens, each with a 3.6-inch diagonal and a resolution of 1080x1200 pixels (2160x1200 pixels combined). The embedded eye-tracking device by SMI tracks both eyes simultaneously with a sampling rate of 250 Hz and has a typical spatial accuracy of 0.2°. [2]

Each sample for each eye is associated with a 3-dimensional unit vector, $v = (v_x, v_y, v_z)$, that represents the direction the eye is looking. Using standard trigonometry, we converted these 3-dimensional unit vectors into degrees of visual angle, (x, y), with the help of MATLAB's `atan2d` function.

$$x = \texttt{atan2d}(v_x, v_z) \tag{4a}$$

$$y = \texttt{atan2d}(v_y, v_z) \tag{4b}$$

The ET-HMD provides three of these direction vectors: left eye direction, right eye direction, and binocular direction (called the "camera raycast"). We have shown below (see Sect. 3.5) that the binocular signal is a low-pass filtered version of the monocular signal. We adjusted the target position for monocular data following our observations shown in Figure 4, assuming an interpupillary distance of 62 mm. No such adjustment was necessary for the binocular data.

2.3.2 EyeLink 1000

The EyeLink 1000 by SR Research is one of the leading eye-tracking devices currently on the market. It tracks one eye with a sampling rate of 1000 Hz and has a typical spatial accuracy between 0.25° and 0.50°. [3]

The EyeLink data we used in the present research was collected by monocularly tracking the left eye. We noticed the vertical data for our EyeLink was consistently 1.2° too low. We do not know the source of this. For the purposes of

[2] ET-HMD manufacturer-supplied specifications taken from `http://twittertechnews.com/wp-content/uploads/2016/09/smi_prod_eyetracking_hmd_HTC_Vive.pdf`.

[3] EyeLink 1000 manufacturer-supplied specifications taken from `http://sr-research.jp/support/EyeLink%201000%20User%20Manual%201.5.0.pdf`.

the comparisons performed in the present research, we simply added 1.2° in the vertical direction to every sample for all the EyeLink data. There was no such offset in the horizontal direction.

2.4 Description of the stimulus

We used Unity 2018.3.11f1 to create our stimulus for the ET-HMD. The stimulus was a small, solid-black sphere on a light gray background. It had a diameter of 0.5° and was positioned at an apparent depth of 1000 mm (the actual target diameter was 8.72 mm, which was 0.5° at the chosen distance).

The primary eye movement task was a random saccade task. Saccades could occur anywhere between 15° to the left and 15° to the right (−15° to +15° horizontally) and between 10° down and 10° up (−10° to +10° vertically), positioned relative to the nasal bridge, at a constant depth of one meter. The position of each saccade target was determined randomly. Thirty oblique saccades were performed, each with a minimum amplitude difference of 3° (radial) from the prior position. The duration of each fixation was chosen from a uniform random distribution (min = 1 s, max = 1.5 s).

A calibration task (described in Sect. 2.6) was prepended to the random saccade task to study the benefits of recalibration.

Headset position tracking was disabled so that, regardless of head orientation, the world center was always centered in the view. As a result, the stimulus was always positioned correctly relative to the nasal bridge.

For the EyeLink, the stimulus was a white ring on a dark background. The ring had an inner diameter of about 0.5° and an outer diameter of about 1°. The recordings we used from the EyeLink were collected during a previous study using a different set of participants, but the eye movement task was also a random saccade task that was very similar to the one used for the ET-HMD in this report. The main differences are that 100 saccades were performed instead of just 30, and no user-supplied calibration task was prepended to the random saccade task. Since the ET-HMD task contained only 30 random saccades, we used only the first 30 saccades from the EyeLink task.

2.5 Processing the gaze position signal

All targets were displayed relative to the nasal bridge, which is defined as the midpoint between the two eyes ("cyclopean eye"). Thus, assuming an interpupillary distance of 62 mm, for a target positioned −15° (horizontal), the left eye was actually viewing at −13.3°; and when the left eye was fixated on a target positioned +15° (horizontal), the left eye was actually viewing at +16.6°. We corrected for this disparity in the target position when evaluating each eye separately, but no correction was necessary when evaluating the binocular signal. See Figure 4 for an illustration of this phenomenon.

The full gaze position signal contained various eye movements, including fixations, saccades, and blinks. We wanted to measure data quality only when subjects were fixating. Typically, the human reaction time to the saccadic movement of a target (saccade latency) is around 200 ms [32, p. 113]. We first found the optimal temporal shift of the eye signal for each recording to align the eye and target movements as much as possible. To obtain the best overall estimate of saccade latency, we calculated the mean Euclidean distance between the measured gaze position and the target position at shifts of 1 sample (4 ms, given the sampling rate of 250 Hz), from 1 sample to 200 samples (4 ms to 800 ms). The shift resulting in the lowest mean Euclidean distance was chosen. We illustrated this process in Figure 5.

We neither manually nor algorithmically classified fixations and saccades. Instead, we modeled the start of each fixation as the end of each target step, and the end of each fixation as the beginning of the next target step. This was only reasonable because we had minimized the delay between the eye position data and the target position data, and also because all of the eye-tracking subjects were normal, healthy adults who followed the target very closely.

Each fixation duration was between 1000 ms and 1500 ms. We discarded the first 400 ms of each fixation, since this was a time of some instability as the saccade, small corrective saccades, and post-saccadic oscillations transitioned into pure fixation. The next 500 ms were employed for our data quality calculations.

2.5.1 Removing outliers

For these 500 ms, it was important to find and remove outliers prior to data quality calculation. We used two screening steps for outlier removal (see Table 2 for outlier statistics):

1. We computed the first and third quartiles and the interquartile range (IQR) of the Euclidean distances between each measured gaze position and the centroid of the measured gaze. Any sample 1.5×IQR below the first quartile or 1.5×IQR above the third quartile was discarded (Tukey's fences [53]).

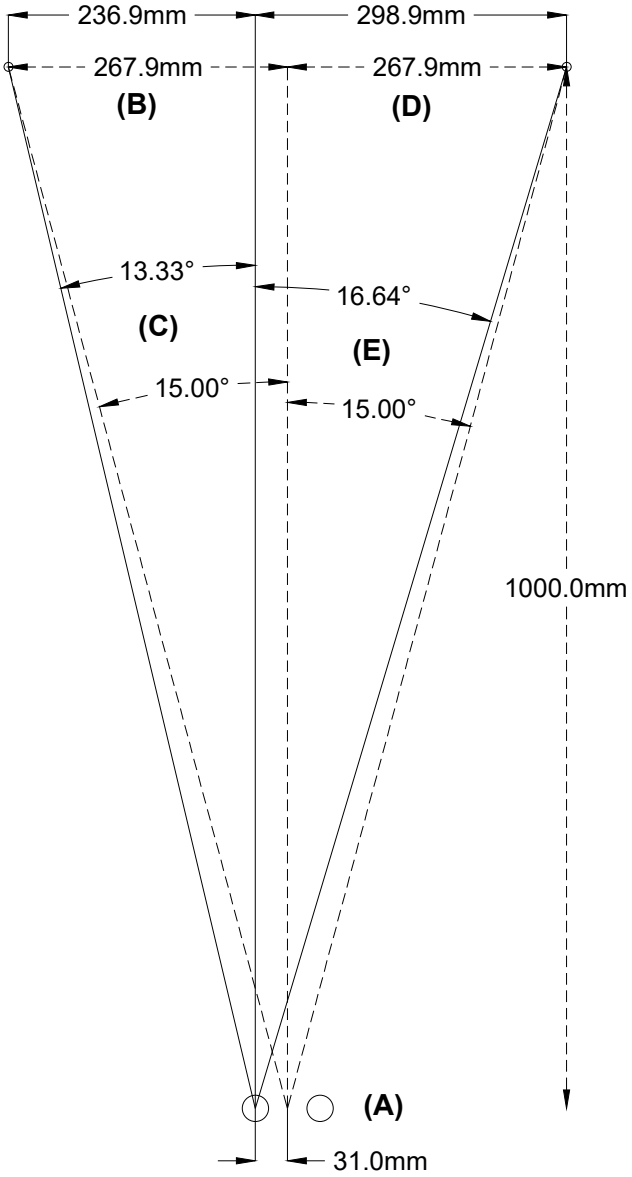

Figure 4: An illustration of the actual vs expected gaze angles for a stimulus positioned 1000 mm away from the eyes. We will consider the case of the left eye, but the same kind of problem occurs for the right eye, as well. (A) The two circles near the bottom of the figure represent the eyes, and the small circles near the top of the figure represent the stimulus. We assumed an interpupillary distance of 62 mm (31 mm from the nasal bridge to either eye). Dashed lines show the angle from the nasal bridge, while solid lines show the angle from the left eye. (B) When looking to the left, the stimulus is closer to the left eye (236.9 mm) than the nasal bridge (267.9 mm). (C) Therefore, a stimulus positioned 15° left of the nasal bridge is only 13.33° left of the left eye. (D) When looking to the right, the stimulus is further from the left eye (298.9 mm) than the nasal bridge (267.9 mm). (E) Therefore, a stimulus positioned 15° right of the nasal bridge is actually 16.64° right of the left eye.

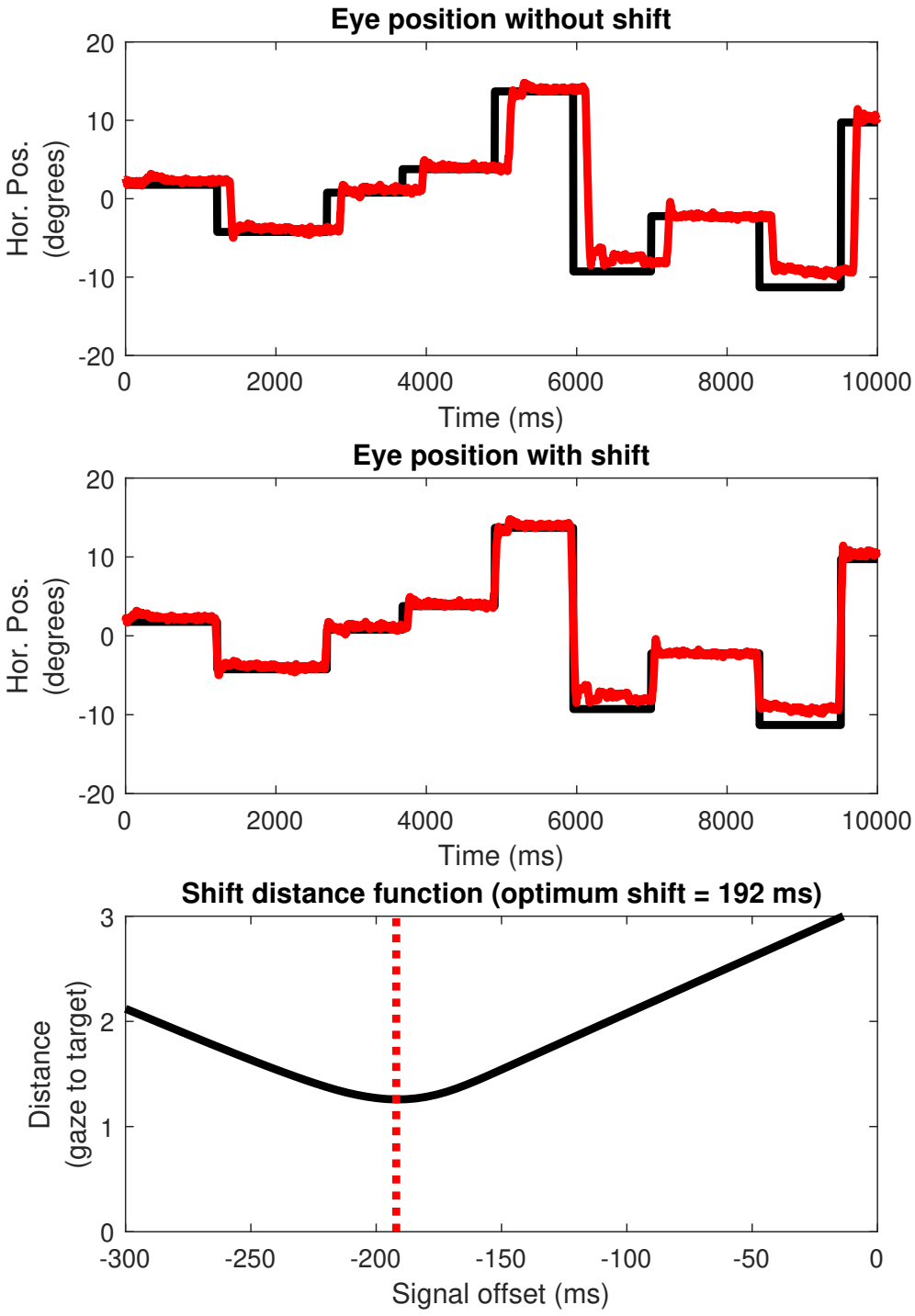

Figure 5: Illustration of our calculation of saccade latency. The top panel shows a typical saccade latency. The bottom panel illustrates the method we employed to determine the optimum shift to minimize the mean euclidean distance between the gaze position signal and the target position signal. We computed the mean euclidean distance between the two signals at various shifts and found the shift with the lowest mean distance (marked here with a red dashed line). The middle panel shows the data after the gaze position signal has been shifted by the optimum shift (in this case, 192 ms).

Table 2: Outliers Removed per Subject

Subject	Outliers removed (%)*	
	Screening step 1	Screening step 2†
1	4.03 ± 6.73	0.00 ± 0.00
2	4.58 ± 6.14	0.00 ± 0.00
3	1.60 ± 2.42	0.00 ± 0.00
4	2.75 ± 4.59	0.00 ± 0.00
5	4.64 ± 7.02	2.43 ± 13.33
6	3.89 ± 5.30	0.00 ± 0.00
7	3.55 ± 6.59	1.97 ± 10.81
8	5.86 ± 8.23	1.70 ± 9.30
9	3.44 ± 4.91	0.00 ± 0.00
10	3.87 ± 7.31	2.40 ± 13.15
11	3.44 ± 4.89	0.00 ± 0.00
12	2.83 ± 3.80	0.00 ± 0.00
Mean‡	3.71 ± 1.08	0.71 ± 1.06

* Presented values are for the ET-HMD's binocular position signal.

† These values do not include samples that were also classified as an outlier during step 1.

‡ Mean and SD values across subjects.

2. We removed all samples that were more than 2° away from the centroid of the measured gaze (using the Euclidean distances calculated above).

Finally, we calculated spatial accuracy and spatial precision for each fixation. Linearity and crosstalk were measured across all fixations. We repeated this process for all subjects and kept track of each fixation's data quality measures for later analysis.

2.6 Description of recalibration

We tested if our spatial accuracy measurements could be improved by employing a USC of some kind. This entailed prepending calibration targets prior to the random saccade task, and computing and applying recalibration coefficients. Targets were displayed at one of 13 positions from a 13-point grid, ranging from top-left at $(-15°, +10°)$ to the bottom right at $(+15°, -10°)$. The order of presentation of these positions was random. Recalibration occurred after reducing saccade latency (see Sect. 2.5).

We wanted to select samples from stable fixations on recalibration targets, with the idea that the eye is probably fixating on the target when it is moving least. To identify stable fixations without classifying the signal, we split each fixation period (based on the target signal) into bins of 20 samples (80 ms). Bins containing any invalid samples were dropped. We computed the interquartile range (IQR) of the remaining bins as a measure of signal stability. To combine the IQRs from the horizontal and vertical bins in a way that favors lower IQRs, we calculated a radial IQR as

$$IQR_{radial} = \sqrt{IQR_x{}^2 + IQR_y{}^2}. \tag{5}$$

The three bins with the lowest radial IQR for each fixation were used for recalibration. We used the full fixation period for binning. Figure 6 shows one of the signals obtained during the recalibration task. Figure 7 shows the selected bins for the horizontal and vertical position signals.

Finally, the horizontal and vertical gaze positions from the most stable bins were calibrated separately using the following equations from Kasprowski et al. [27]:

Linear:

$$x' = A_x x + B_x y + C_x \tag{6a}$$

$$y' = A_y x + B_y y + C_y \tag{6b}$$

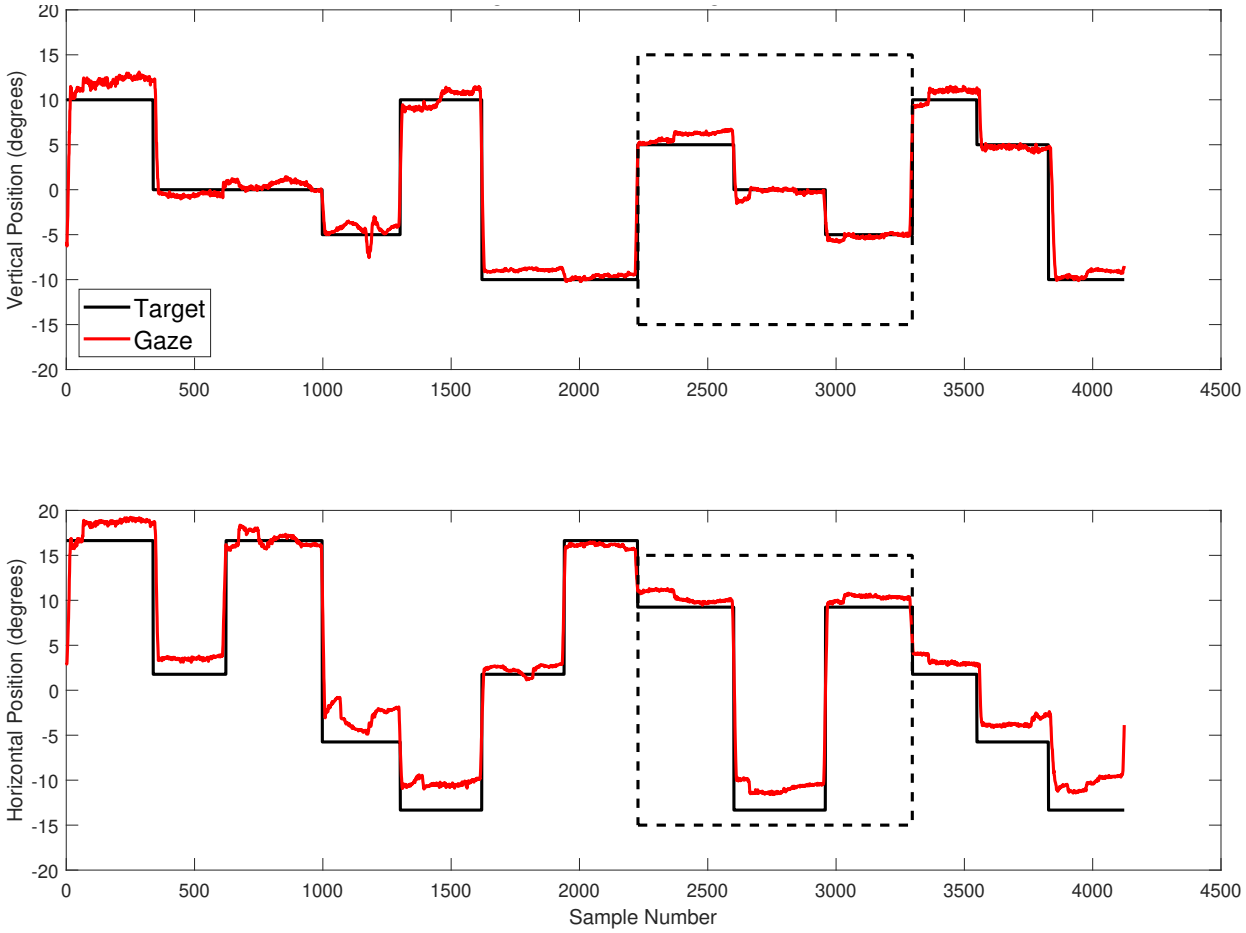

Figure 6: The full signal used for recalibration for one of the subjects. The boxed portion of the signal is enlarged and used in Figure 7.

Quadratic:

$$x' = A_x x^2 + B_x y^2 + C_x x + D_x y + E_x \tag{7a}$$

$$y' = A_y x^2 + B_y y^2 + C_y x + D_y y + E_y \tag{7b}$$

where A, B, C, D, E are the weights calculated by the regression procedure. Note that all of these equations account for potential crosstalk between horizontal and vertical signals.

All of the data in this report were calibrated with a manufacturer-supplied calibration routine (for both the ET-HMD and the EyeLink). In what follows, we use "NO-USC" to mean no additional user-supplied calibration was applied beyond that provided by the device. We refer to the linear calibration as USC-1 and the quadratic calibration as USC-2. Each recalibration regression function was tested on the data obtained during the random saccade task.

2.7 Statistical analyses

When evaluating the data quality across eyes within the ET-HMD data, the statistical procedures will be most sensible if we imagine a particular dependent variable—say, for example, horizontal accuracy. As noted above, each subject has one estimate of horizontal accuracy for each recording. To compare horizontal accuracy across eyes, we have one fixed effect with three levels of "eye," and we have 12 subjects. Because the subjects are the same across eyes, the "eye" effect is a repeated-measures (or within-subjects) effect. This eye fixed effect was evaluated with a mixed model with subjects modeled as a random effect (R package `lmerTest`, Kuznetsova et al. [31]). Degrees of freedom were estimated using the Satterthwaite method (the default method for `lmerTest`). A statistically significant F-value for

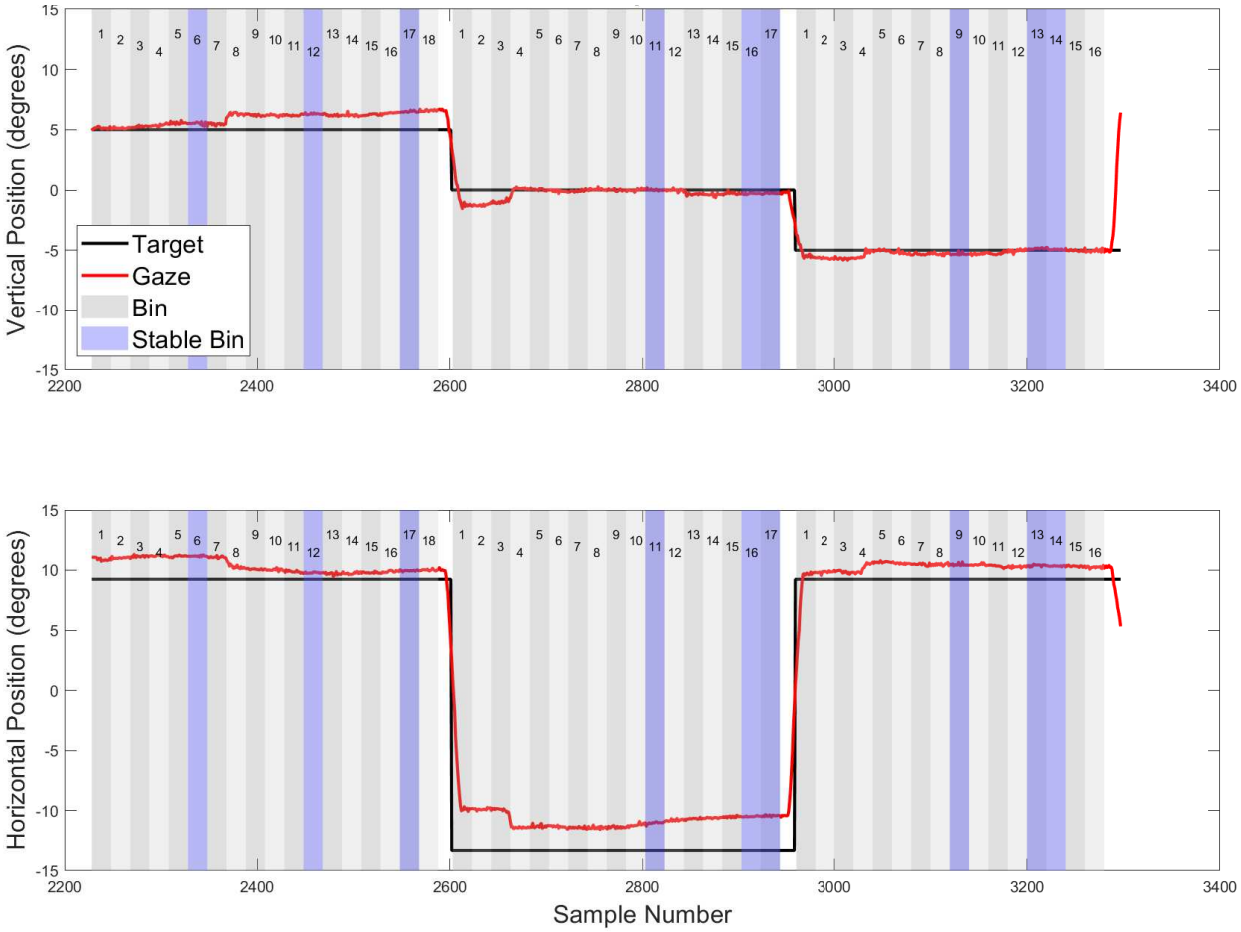

Figure 7: The most stable shared horizontal and vertical bins for a handful of fixations. Notice that each calibration target position has three "stable bins." The data in each of these stable bins are used for recalibration; the rest are ignored.

the effect of "eye" was followed up with Dunnett-style contrasts with "binocular" as the reference level (R package `multcomp`, Hothorn et al. [25]), and the Holm-Bonferroni method [20] was used to control for multiple comparisons. All p-values were two-tailed.

When evaluating the effect of calibration method on data quality within the ET-HMD data, let's keep horizontal accuracy in mind as the dependent variable. We had two statistical questions related to calibration: (1) Is calibration effective in improving data quality? (2) Are the two calibration methods different in effectiveness? All of the calibration effects were tested using mixed models with subjects treated as a random effect (again using the `lmerTest` package [31]). We employed Helmert-style contrasts (again using the `multcomp` package [25]) with a planned comparison approach. Therefore, each statistical test for calibration gave us two p-values: one for the effect of calibration, and one for the difference between the two calibration methods. The Holm-Bonferroni method [20] was again employed to control for the presence of multiple tests. All p-values were two-tailed.

To compare data quality from the ET-HMD to the EyeLink 1000, we note that the data from these two devices come from completely independent subjects, and so a between-subjects design is appropriate. We conducted several one-way analyses of means (R package `stats`, R Core Team [41]), not assuming equal variances. (These are the exact statistical equivalent of a Welch's t-test.) A statistically significant F-test indicates that the two devices have different horizontal accuracy (for example). Inspection of the means indicated the direction of the effect. All p-values were two-tailed.

Regarding linearity slopes, the above comparisons only test whether the slopes are significantly different between groups. In order to test the significance of the slopes themselves, we also performed a one sample t-test comparing slopes to the ideal slope of 1.0.

2.7.1 Data transformations

Mathematical transformations were employed to enhance the normality of the underlying distributions of dependent measures prior to statistical analysis. Both spatial accuracy and spatial precision were, as a general matter, right-skewed, and a cube-root transformation improved normality for these measures. Linearity R^2_{adj} are proportions, so a logit-transform was effective. No transformation was necessary for linearity slope.

2.8 Comparing left eye, right eye, and binocular signals using Fourier analysis

It was important to characterize the frequency content of our signals. To this end, we employed a Fast Fourier Analysis (FFT). For each of 12 subjects, we chose 3 stable fixation periods manually that were all 256 samples in length. Data from each fixation period was mean-centered and detrended using a second-order polynomial fit. Next, a Hann window was applied. A FFT was performed on each fixation period using MATLAB FFT functions. The single-sided magnitude spectra were extracted for each fixation period. Average magnitude spectra were created by averaging across all fixation periods (M=3 per subject) and all subjects (N=12) for a total of 36 fixation periods. This was done for the left eye, the right eye, and the binocular signal. As per Hooge et al. [23], we also created a "version" signal which is the literal average of the left and right eye signals. Since the version signal appeared even more similar to the binocular signal than either the left or right eye alone, and since the version signal should, in theory, also be a good approximation of a binocular signal, we also analyzed the version signal in the same way. Examination of the magnitude spectrum of the version signal and the binocular signal suggested to us that the binocular signal may simply represent a low-pass filtered form of the version signal. We wanted to characterize the filter which might have transformed the version signal into the binocular signal, and we posted a question on the website StackExchange[4] where user MBaz posted the following method to obtain the filter.

The goal is to find a filter $h(t)$ such that:

$$y(t) = x(t) * h(t), \tag{8}$$

where $x(t)$ is the version signal and $y(t)$ is the binocular signal. To find the filter, note that:

$$Y(f) = X(f)H(f), \tag{9}$$

where f indicates the Fourier domain. So,

$$h(t) = \text{Inverse FFT}\{Y(f)/X(f)\}. \tag{10}$$

We computed h(t) for each of the three 256-sample fixation periods from each subject, and then averaged all 36 h(t) filter estimates to get an average filter. We then used the MATLAB tool `freqz` to plot the frequency response of the filter.

3 Results

Sect. 3.1 assesses the data quality of the ET-HMD at different levels of Eye and Calibration Method. Sect. 3.2 assesses the data quality of the ET-HMD compared to the EyeLink. All of the raw (untransformed) means and SDs of all of our quantification results are presented in Table 3, which will be further discussed as the results section progresses.

3.1 Analysis of ET-HMD data quality

Our results with respect to the ET-HMD will be focused on two questions: (1) How do the data from the left eye, the right and and the binocular signal compare on measures of data quality? (2) What is the impact of our two different methods of recalibration on data quality measures?

In Sect. 3.1.1, we separately analyze the effect of Eye at each level of Calibration Method (NO-USC, USC-1, and USC-2). Similarly, in Sect. 3.1.2, we separately analyze the effect of Calibration Method at each level of Eye (left, right, and binocular).

[4] `https://dsp.stackexchange.com/questions/60098/reverse-engineering-a-digital-filter/60137`

Table 3: Results for spatial accuracy, spatial precision, and linearity for both the ET-HMD (with each recalibration method) and the EyeLink. Values are presented as the mean ± 1 SD across subjects.

Device	Dim*	Eye†	Spatial accuracy	Spatial precision	Linearity	
					Slope	R^2_{adj}
ET-HMD	H	L	0.892 ± 0.366	0.102 ± 0.022	0.965 ± 0.047	0.993 ± 0.012
		R	1.041 ± 0.236	0.084 ± 0.015	0.976 ± 0.033	0.998 ± 0.001
		B	0.375 ± 0.173	0.052 ± 0.010	0.990 ± 0.037	0.998 ± 0.004
	V	L	0.690 ± 0.518	0.109 ± 0.044	0.992 ± 0.055	0.991 ± 0.012
		R	0.528 ± 0.272	0.092 ± 0.023	0.979 ± 0.040	0.995 ± 0.006
		B	0.474 ± 0.301	0.059 ± 0.012	1.003 ± 0.042	0.996 ± 0.005
	C	L	1.242 ± 0.651	0.173 ± 0.053	-	-
		R	1.251 ± 0.312	0.148 ± 0.027	-	-
		B	0.671 ± 0.338	0.107 ± 0.023	-	-
ET-HMD + USC-1	H	L	0.566 ± 0.379	0.105 ± 0.025	0.991 ± 0.029	0.993 ± 0.012
		R	0.358 ± 0.138	0.086 ± 0.016	1.000 ± 0.018	0.998 ± 0.002
		B	0.382 ± 0.264	0.054 ± 0.011	0.999 ± 0.020	0.996 ± 0.006
	V	L	0.418 ± 0.210	0.110 ± 0.038	1.005 ± 0.028	0.992 ± 0.011
		R	0.376 ± 0.150	0.093 ± 0.023	0.996 ± 0.029	0.995 ± 0.005
		B	0.300 ± 0.118	0.059 ± 0.011	1.005 ± 0.016	0.996 ± 0.004
	C	L	0.776 ± 0.467	0.176 ± 0.051	-	-
		R	0.575 ± 0.193	0.151 ± 0.027	-	-
		B	0.542 ± 0.295	0.111 ± 0.029	-	-
ET-HMD + USC-2	H	L	0.548 ± 0.359	0.105 ± 0.024	0.989 ± 0.022	0.992 ± 0.013
		R	0.366 ± 0.128	0.086 ± 0.016	1.000 ± 0.019	0.997 ± 0.004
		B	0.376 ± 0.239	0.053 ± 0.010	0.996 ± 0.016	0.996 ± 0.008
	V	L	0.450 ± 0.242	0.109 ± 0.038	1.005 ± 0.033	0.991 ± 0.011
		R	0.384 ± 0.154	0.093 ± 0.024	0.995 ± 0.034	0.994 ± 0.005
		B	0.300 ± 0.122	0.059 ± 0.011	1.004 ± 0.018	0.996 ± 0.004
	C	L	0.784 ± 0.470	0.176 ± 0.050	-	-
		R	0.588 ± 0.200	0.151 ± 0.028	-	-
		B	0.536 ± 0.274	0.109 ± 0.025	-	-
EyeLink	H	L	0.677 ± 0.297	0.059 ± 0.023	1.009 ± 0.025	0.982 ± 0.044
	V	L	0.773 ± 0.480	0.074 ± 0.027	1.002 ± 0.068	0.932 ± 0.156
	C	L	1.137 ± 0.528	0.141 ± 0.066	-	-

* Denotes the measurement dimension: horizontal (H), vertical (V), or combined (C).
† Denotes the measured gaze signal: left eye (L), right eye (R), or binocular (B).

Table 4: Analysis of the effect of eye on data quality at different levels of calibration method.

Calibration method	Dependent variable		DF* Num	DF* Den	F†	p	Sig.	L - B‡ Est.	L - B‡ p	R - B§ Est.	R - B§ p
NO-USC	Accuracy	H	2	22.00	35.93	<.001	✓	0.24	<.001	0.30	<.001
		V	2	22.00	7.25	.004	✓	0.09	<.001	0.03	.173
		C	2	22.00	28.69	<.001	✓	0.19	<.001	0.21	<.001
	Precision	H	2	22.00	67.47	<.001	✓	0.09	<.001	0.06	<.001
		V	2	22.00	58.69	<.001	✓	0.09	<.001	0.06	<.001
		C	2	22.00	45.57	<.001	✓	0.08	<.001	0.05	<.001
	Lin. slope	H	2	22.00	4.40	.025	✓	-0.02	.006	-0.01	.109
		V	2	22.00	3.58	.045	✓	-0.01	.209	-0.02	.015
	Lin. fit	H	2	22.00	11.95	<.001	✓	-0.90	<.001	-0.31	.093
		V	2	22.00	19.31	<.001	✓	-0.69	<.001	-0.23	.039
USC-1	Accuracy	H	2	22.00	7.24	.004	✓	0.10	.002	0.00	.883
		V	2	22.00	22.47	<.001	✓	0.07	<.001	0.05	<.001
		C	2	22.00	10.04	<.001	✓	0.10	<.001	0.03	.216
	Precision	H	2	22.00	63.03	<.001	✓	0.09	<.001	0.06	<.001
		V	2	22.00	69.59	<.001	✓	0.09	<.001	0.06	<.001
		C	2	22.00	42.38	<.001	✓	0.08	<.001	0.05	<.001
	Lin. slope	H	2	22.00	0.95	.403					
		V	2	22.00	0.60	.560					
	Lin. fit	H	2	22.00	8.06	.002	✓	-0.85	<.001	-0.21	.348
		V	2	22.00	11.62	<.001	✓	-0.61	<.001	-0.39	.003
USC-2	Accuracy	H	2	22.00	6.37	.007	✓	0.10	.002	0.01	.717
		V	2	22.00	15.04	<.001	✓	0.09	<.001	0.06	<.001
		C	2	22.00	12.19	<.001	✓	0.10	<.001	0.03	.105
	Precision	H	2	22.00	72.48	<.001	✓	0.09	<.001	0.07	<.001
		V	2	22.00	69.90	<.001	✓	0.09	<.001	0.06	<.001
		C	2	22.00	50.31	<.001	✓	0.08	<.001	0.06	<.001
	Lin. slope	H	2	22.00	1.66	.213					
		V	2	22.00	0.67	.522					
	Lin. fit	H	2	22.00	7.84	.003	✓	-0.89	<.001	-0.17	.490
		V	2	22.00	16.26	<.001	✓	-0.67	<.001	-0.38	.001

* Numerator and denominator degrees of freedom, estimated using the Satterthwaite method.

† F-values from the mixed model ANOVAs.

‡ Contrasts between the left and binocular signals. Omitted when the ANOVA is not significant. A positive estimate indicates the value from the left eye is higher (but not necessarily *better*).

§ Contrasts between the right and binocular signals. Omitted when the ANOVA is not significant. A positive estimate indicates the value from the right eye is higher (but not necessarily *better*).

3.1.1 Effect of eye on data quality

First, we will consider the data when Calibration Method is NO-USC. The means and SDs for the uncalibrated and untransformed data for all quality measures is in the top panel of Table 3, labeled "ET-HMD." The means for accuracy and precision of the binocular signal were consistently superior to the left and right eye. The linearity slopes for the binocular signal were closer to ideal than that for the left and right eye. For linearity fit, the binocular signal was either superior to or tied with the best performance from the left or right eye.

Statistical tests for these comparison are in the section labelled "NO-USC" in Table 4. All of the F-values are statistically significant, indicating that the data quality depends on eye. Seventeen of 20 tests comparing either the left eye or the right eye to the binocular signal were statistically significant, indicating the superiority of the binocular signal for data quality in uncalibrated data.

Table 5: Results from the t-tests comparing linearity slopes to the ideal of 1.0.

Device	Eye	Dim	DF	t	p	Sig.*	Slope estimate	95% conf. int.	
								Lower	Upper
ET-HMD	L	H	11	-2.60	.025	✓	0.965	0.935	0.995
		V	11	-0.52	.613		0.992	0.957	1.027
	R	H	11	-2.50	.029	✓	0.976	0.956	0.997
		V	11	-1.82	.096		0.979	0.953	1.004
	B	H	11	-0.94	.366		0.990	0.966	1.014
		V	11	0.27	.796		1.003	0.977	1.030
ET-HMD + USC-1	L	H	11	-1.11	.291		0.991	0.972	1.009
		V	11	0.60	.561		1.005	0.987	1.023
	R	H	11	0.05	.961		1.000	0.989	1.012
		V	11	-0.43	.678		0.996	0.978	1.015
	B	H	11	-0.14	.888		0.999	0.986	1.012
		V	11	1.06	.311		1.005	0.995	1.015
ET-HMD + USC-2	L	H	11	-1.70	.118		0.989	0.975	1.003
		V	11	0.54	.602		1.005	0.984	1.026
	R	H	11	0.00	1		1.000	0.988	1.012
		V	11	-0.50	.630		0.995	0.973	1.017
	B	H	11	-0.96	.360		0.996	0.985	1.006
		V	11	0.85	.412		1.004	0.993	1.016
EyeLink	L	H	9	1.08	.309		1.009	0.991	1.027
		V	9	0.09	.927		1.002	0.954	1.050

* A significant result indicates that the linearity slope is significantly different from the ideal slope of 1.0.

Next, we will consider the data after applying the USC-1 method of recalibration. The means and SDs for the untransformed quality measures after USC-1 recalibration are in the second panel of Table 3, labeled "ET-HMD + USC-1." The means for vertical and combined spatial accuracy of the binocular signal are better than the left and right eye. The means for spatial precision of the binocular signal are better than the left and right eye. The linearity slopes are not much different between the binocular, left, and right eye signals. For linearity fit, the binocular signal is either superior to or tied with the best performance from the left and right eye.

Statistical tests for these comparisons are in the section labeled "USC-1" in Table 4. The F-values for accuracy, precision, and linearity fit are all statistically significant, indicating that the data quality depends on eye. Thirteen of 16 tests comparing either the left eye or the right eye to the binocular signal were statistically significant, indicating the superiority of the binocular signal for data quality even after recalibration with USC-1.

Lastly, we will consider the data after applying the USC-2 method of recalibration. The means and SDs for the untransformed quality measures after USC-2 recalibration are in the third panel of Table 3, labeled "ET-HMD + USC-2." Again, the means for vertical and combined spatial accuracy of the binocular signal are better than the left and right eye. The means for spatial precision of the binocular signal are better than the left and right eye. The linearity slopes are again not much different between the binocular, left, and right eye signals. For linearity fit, the binocular signal is only better in the vertical direction.

Statistical tests for these comparisons are in the section labeled "USC-2" in Table 4. As with USC-1, the F-values for accuracy, precision, and linearity fit are all statistically significant, indicating that the data quality depends on eye. Again, 13 of 16 post-hoc tests were statistically significant, indicating the superiority of the binocular signal for data quality even after recalibration with USC-2.

In Table 5, we present the results of the t-tests comparing each linearity slope to the ideal slope of 1.0. As a general matter, the empirical slopes for linearity were not statistically different from the ideal slope. However, for the horizontal slopes of the uncalibrated left and right eyes, the slopes were significantly lower than the ideal.

Table 6: Analysis of the effect of calibration method on data quality for each eye.

Eye	Dependent variable		NO-USC vs mean(USC-1 and USC-2)					USC-1 vs USC-2				
			Est.	Std. Err.	z*	p	Sig.	Est.	Std. Err.	z*	p	Sig.
Left	Accuracy	H	0.15	0.03	4.71	<.001	✓	0.01	0.04	0.20	.841	
		V	0.11	0.03	4.04	<.001	✓	-0.02	0.03	-0.49	.626	
		C	0.16	0.03	6.12	<.001	✓	-0.00	0.03	-0.06	.953	
	Precision	H	-0.00	0.00	-2.41	.031	✓	0.00	0.00	0.27	.785	
		V	-0.00	0.00	-0.68	.993		0.00	0.00	0.19	.993	
		C	-0.00	0.00	-2.20	.055		0.00	0.00	0.02	.982	
	Lin. slope	H	-0.02	0.01	-3.28	.002	✓	0.00	0.01	0.19	.846	
		V	-0.01	0.01	-1.12	.522		-0.00	0.01	-0.01	.990	
	Lin. fit	H	0.13	0.06	1.99	.094		0.09	0.07	1.24	.215	
		V	-0.06	0.07	-0.84	.401		0.13	0.08	1.64	.201	
Right	Accuracy	H	0.30	0.02	14.43	<.001	✓	-0.01	0.02	-0.23	.815	
		V	0.08	0.02	3.98	<.001	✓	-0.01	0.02	-0.24	.807	
		C	0.24	0.01	17.16	<.001	✓	-0.01	0.02	-0.31	.754	
	Precision	H	-0.00	0.00	-3.48	.001	✓	-0.00	0.00	-0.25	.802	
		V	-0.00	0.00	-2.45	.029	✓	0.00	0.00	0.15	.877	
		C	-0.00	0.00	-3.96	<.001	✓	0.00	0.00	0.32	.747	
	Lin. slope	H	-0.02	0.01	-3.50	<.001	✓	0.00	0.01	0.03	.973	
		V	-0.02	0.01	-3.37	.002	✓	0.00	0.01	0.22	.826	
	Lin. fit	H	0.03	0.11	0.29	1		0.01	0.12	0.05	1	
		V	0.14	0.06	2.18	.059		0.06	0.07	0.85	.395	
Binocular	Accuracy	H	0.01	0.02	0.61	1		0.00	0.02	0.03	1	
		V	0.10	0.02	4.24	<.001	✓	-0.00	0.03	-0.01	.994	
		C	0.06	0.02	3.71	<.001	✓	0.00	0.02	0.03	.978	
	Precision	H	-0.00	0.00	-1.23	.440		0.00	0.00	0.83	.440	
		V	-0.00	0.00	-0.30	1		0.00	0.00	0.21	1	
		C	-0.00	0.00	-1.35	.353		0.00	0.00	0.88	.380	
	Lin. slope	H	-0.01	0.01	-1.04	.598		0.00	0.01	0.43	.668	
		V	-0.00	0.01	-0.20	1		0.00	0.01	0.05	1	
	Lin. fit	H	0.16	0.10	1.68	.187		0.05	0.11	0.43	.667	
		V	-0.01	0.06	-0.11	.915		0.07	0.06	1.15	.504	

* z-statistics from the Helmert-style planned comparisons.

Based on these results, if one had to choose only one signal based on accuracy, precision, and linearity, the binocular signal would have to be that choice.

3.1.2 Effect of calibration method on data quality

The means and SDs for the untransformed data are in Table 3, and the statistical analysis of the effect of calibration is presented in Table 6. We also present a visualization of the comparisons in Figure 8.

For the left eye, right eye, and binocular signal, as a general matter, the means for spatial accuracy for NO-USC were consistently worse than USC-1 and USC-2 in the horizontal, vertical, and combined directions (see the row of subplots labeled "Accuracy" in Figure 8). These observations were overwhelmingly supported statistically (Table 6). The one exception was for the horizontal spatial accuracy of the binocular signal, where there was no significant difference between the uncalibrated and recalibrated data. The median percent improvement in spatial accuracy from uncalibrated to recalibrated across dimensions and eyes was 12.8%. Another consistent finding was that there were not statistically significant differences between calibration methods for accuracy in any case. This is evident in the row of subplots labeled "Accuracy" in Figure 8.

For all eyes, spatial precision either did not improve or actually worsened with calibration (see the row of subplots labeled "Precision" in Figure 8). However, spatial precision never significantly worsened with calibration for the binocular signal. These observations were supported statistically (Table 6). Although the effect of calibration on precision was occasionally significant, the median percent change of spatial precision for these statistically significant models was only 0.7% compared to 12.8% for accuracy. So, although some comparisons of spatial precision were statistically significant, the percent change in these cases was trivial. Another consistent finding was that there were not statistically significant differences between calibration methods for precision in any case. This is evident in the row of subplots labeled "Precision" in Figure 8.

Generally, for monocular data (left and right eyes), the linearity slopes were closer to ideal after recalibration for the horizontal and vertical directions (see the row of subplots labeled "Linearity slope" in Figure 8). These observations are supported statistically, except in one case for the vertical linearity slope of the left eye, where there was no significant difference between uncalibrated and recalibrated data (Table 6). For binocular data, calibration did not improve linearity slopes. This is also evident in the row of subplots labeled "Linearity slope" in Figure 8. The median percent change of linearity slope from uncalibrated to recalibrated across dimensions and eyes was 1.5%. Another consistent finding was that there were not statistically significant differences between calibration methods for linearity slope in any case.

Regarding linearity fit, there was no statistically significant difference between uncalibrated and recalibrated data for any of the eyes (see the row of subplots labeled "Linearity fit" in Figure 8). The median percent change of linearity fit from uncalibrated to recalibrated across dimensions and eyes was -1.3%. Another consistent finding was that there were not statistically significant differences between calibration methods for linearity fit in any case.

On balance, calibration improved performance, but there was no evidence that either calibration method was superior. Summarizing Sect. 3.1.1 and Sect. 3.1.2, the best overall performance was from the binocular signal calibrated by any means.

3.2 Analysis of ET-HMD versus EyeLink

Our goal here is to compare data quality on the ET-HMD and the EyeLink (see Table 7). In the previous section, we found that the binocular signal from the ET-HMD always performed at least as well as—and often better than—the left and right eye signals. Additionally, we found that recalibration improved the vertical and combined spatial accuracy of the binocular signal, but there was no significant difference between the performance of USC-1 and USC-2. Given these findings, for the purpose of simplification, in this section, we will only present analyses using the binocular signal from the ET-HMD for NO-USC and USC-1. We did, however, assess the effect of Device for each pair of eye and calibration method. The statistical analysis revealed that the ET-HMD had better accuracy and linearity fit than the EyeLink before and after recalibration.

3.3 Temporal precision

Figure 9 shows the distribution of ISIs with the ET-HMD across subjects. This distribution has a mean of 4.000 ms and a SD of 0.071 ms. The SD of the distribution represents the temporal precision as we define it. If the goal is to compute variability irrespective of nominal sampling rate, then one can always compare devices by dividing the SD by the nominal rate.

We also investigated how many samples were dropped with the ET-HMD. We defined a dropped sample as any sample with an ISI over 6 ms (50% more than the ideal ISI of 4 ms). In this way, we found that only 4 samples were dropped (1 for one subject, 3 for another) out of 113,264 total samples (approximately 9,400 samples per subject). We also found that there were 6 extremely short ISIs ($< .04$ ms), each occurring for different subjects. The cause of these short ISIs is unknown to us.

According to information provided by technical staff at SR-Research, the ISI for the EyeLink 1000 is precisely 1 ms. Also, the Eyelink 1000 does not provide timestamp information with sub-millisecond precision. Therefore, we are forced to assume that the EyeLink has perfect temporal precision.

3.4 Crosstalk

Crosstalk analyses for the ET-HMD are presented in Figure 10. In 7 of 18 test conditions, the best model was the intercept-only model (neither linear nor quadratic components). In 7 of 18 test conditions, the quadratic-only fit was the best model. In 3 cases, both a linear component and a quadratic component were needed to optimally fit the data. In only one case was a linear-only fit optimal. This is despite the fact that in the literature, only linear fits are tested for and we are not aware of any prior reports of quadratic fits for crosstalk.

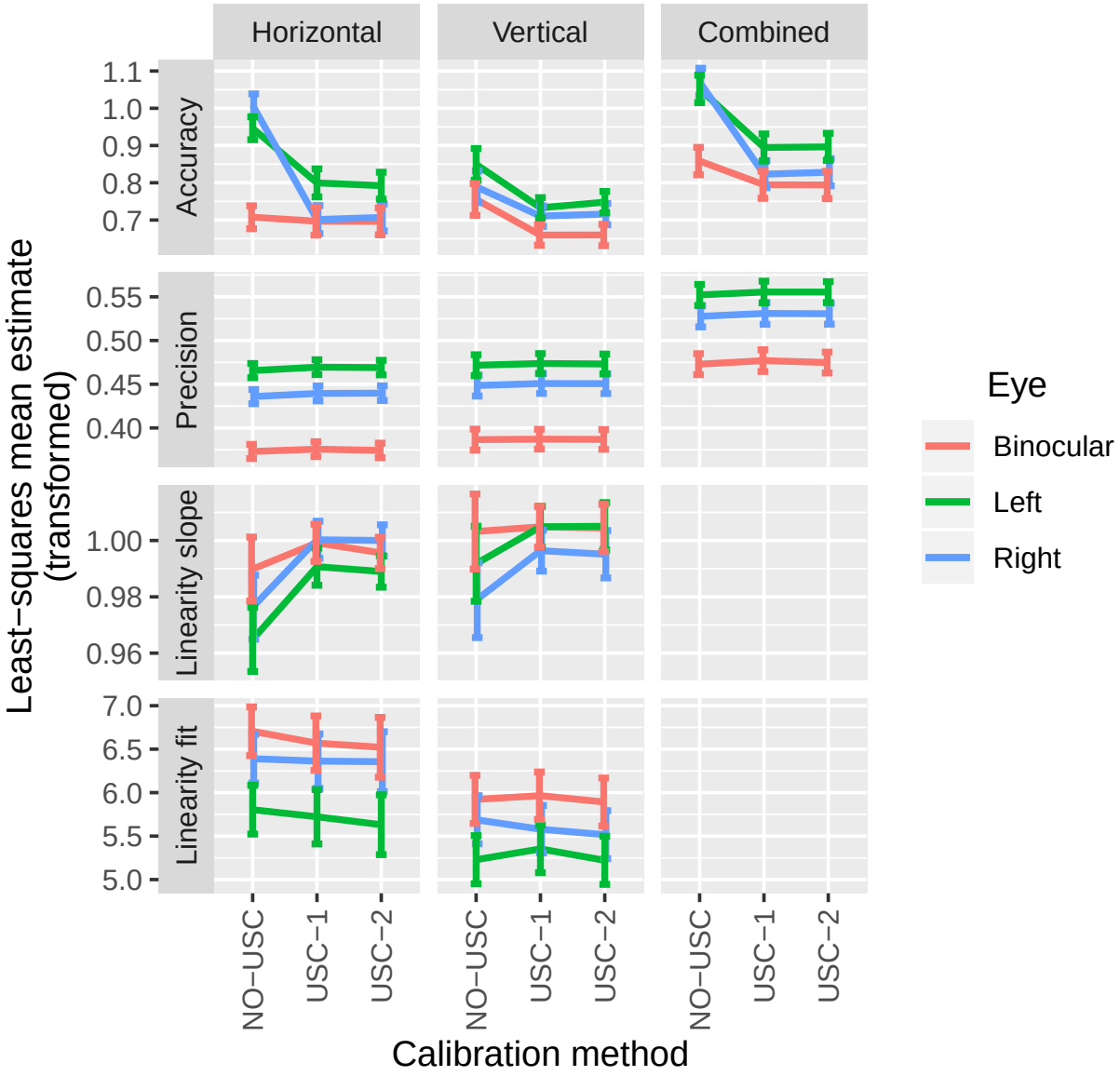

Figure 8: Least-squares means of data quality measures across eyes for each calibration method and dimension.

Table 7: Comparison of the data quality between the ET-HMD (binocular signal) and the EyeLink (left eye signal).

Calibration method*	Dependent variable		DF† Num	DF† Den	F‡	p	Sig.	Means§ ET-HMD	Means§ EyeLink
NO-USC	Accuracy	H	1	16.95	9.00	.008	✓	0.707	0.861
		V	1	17.79	4.14	.057		0.755	0.889
		C	1	17.66	7.81	.012	✓	0.858	1.024
	Precision	H	1	12.75	0.56	.469		0.373	0.385
		V	1	13.21	2.57	.132		0.387	0.415
		C	1	11.94	1.94	.189		0.473	0.509
	Lin. slope	H	1	19.33	1.96	.177		0.990	1.009
		V	1	14.42	0.00	.962		1.003	1.002
	Lin. fit	H	1	16.80	6.08	.025	✓	6.705	5.385
		V	1	14.76	16.89	<.001	✓	5.922	3.726
USC-1	Accuracy	H	1	19.84	7.69	.012	✓	0.697	0.861
		V	1	12.70	16.05	.002	✓	0.660	0.889
		C	1	18.22	14.36	.001	✓	0.795	1.024
	Precision	H	1	13.16	0.33	.575		0.376	0.385
		V	1	12.91	2.52	.137		0.387	0.415
		C	1	13.11	1.45	.251		0.477	0.509
	Lin. slope	H	1	17.03	0.92	.351		0.999	1.009
		V	1	9.83	0.02	.899		1.005	1.002
	Lin. fit	H	1	18.73	4.25	.054		6.569	5.385
		V	1	14.71	17.60	<.001	✓	5.965	3.726

* Recalibration was only done for the ET-HMD.

† Estimated numerator and denominator degrees of freedom.

‡ F-values from the one-way analyses of means.

§ The means shown have undergone transformations to normality, but these transformations are monotonic and do not affect whether higher or lower values are better.

For the EyeLink, we found no evidence of significant horizontal or vertical crosstalk. In other words, in both cases, the model with only the intercept was the best model.

3.5 Fourier analysis of signals

In the upper panel of Figure 11, we can see the magnitude spectra for version signal and the binocular signal. The binocular signal appears to be a low pass filtered form of the version signal. The low pass filter is not very sharp (ergo, low-order) and has substantial ringing in the passband. Using the method suggested by MBaz (see Methods), the lower panel is the frequency response of a filter that would transform our version signal into our binocular signal. The -3db point is approximately 11 Hz. The -3db point is the point where the power of the signal (magnitude^2) is reduced by 50% or the magnitude is reduced by approximately 30%.

To get a sense of the effect of such a filter on saccade dynamics, Figure 12 illustrates several saccades unfiltered (A and D), lightly filtered with a Savitzky-Golay filter (B and E), and the binocular signal (C and F).

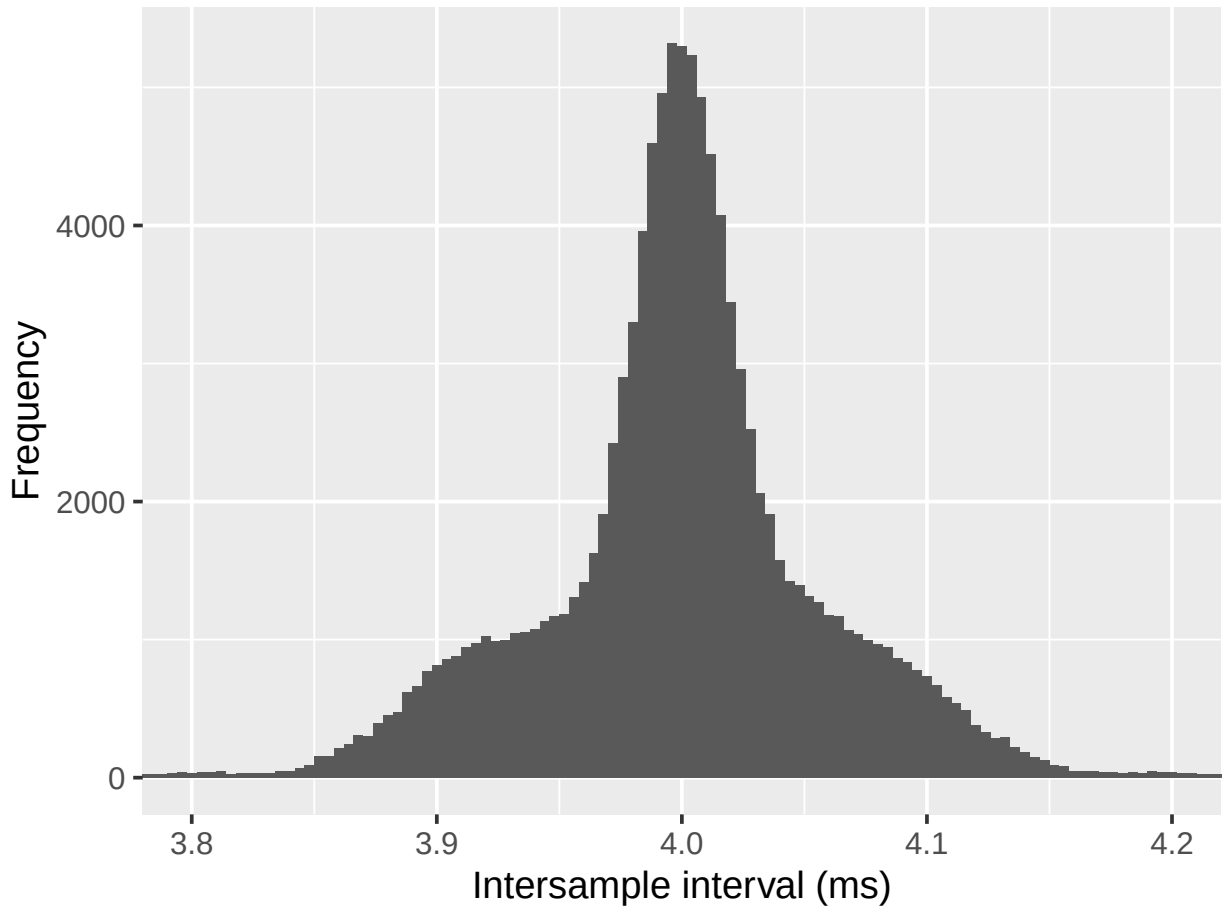

Figure 9: A visualization of the temporal precision of the ET-HMD, combined across subjects. For the purposes of visualization, ISIs that are 5% above or below the expected value are not shown. 99.4% of ISIs obtained by the ET-HMD fit within this range. This means that 99.4% of the time, the ET-HMD has an ISI between 3.8 and 4.2 ms.

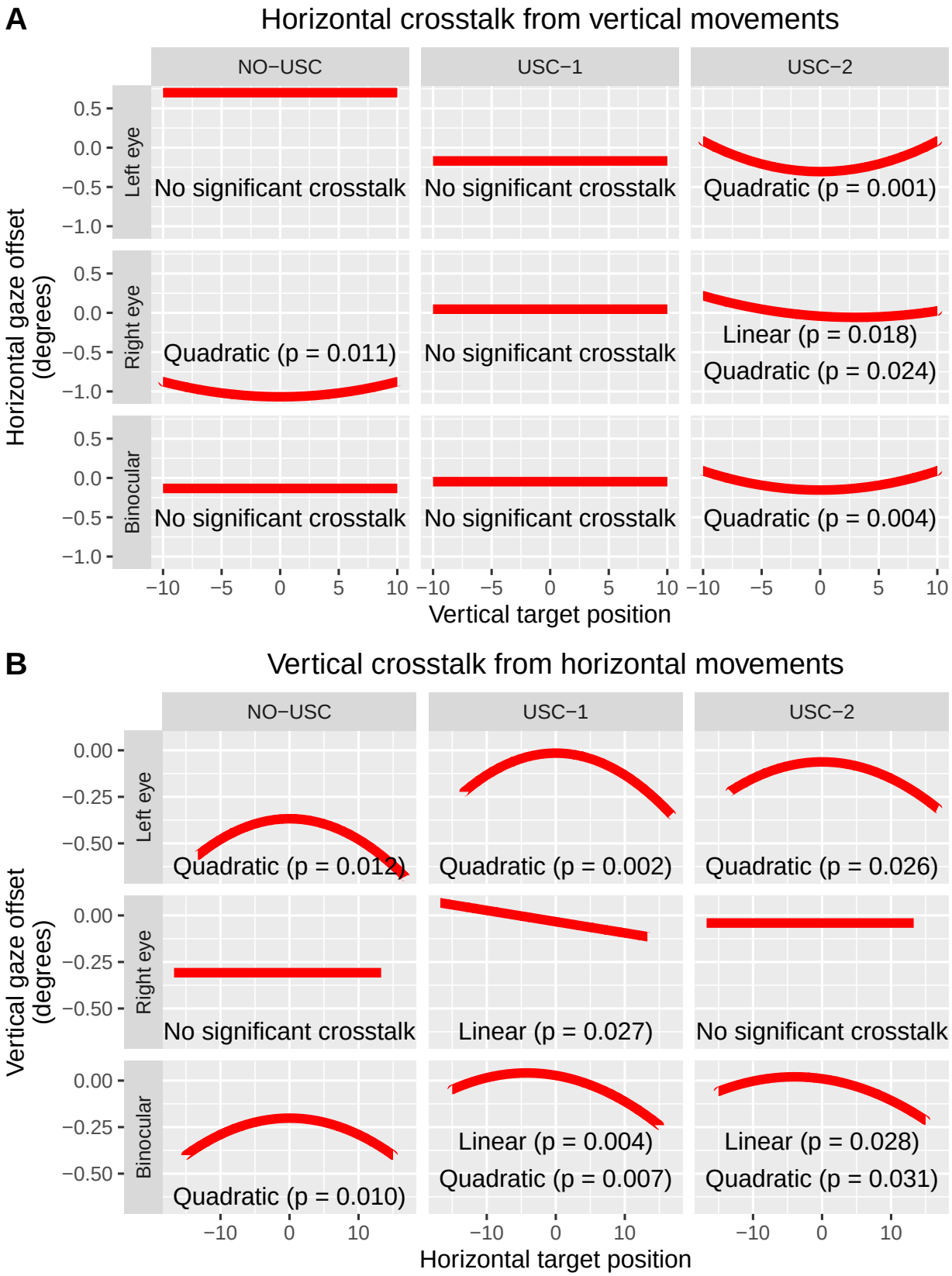

Figure 10: Best-fitting crosstalk models across subjects for each pair of eye and calibration method. The p-values shown are the significance of each component according to the stepwise regression.

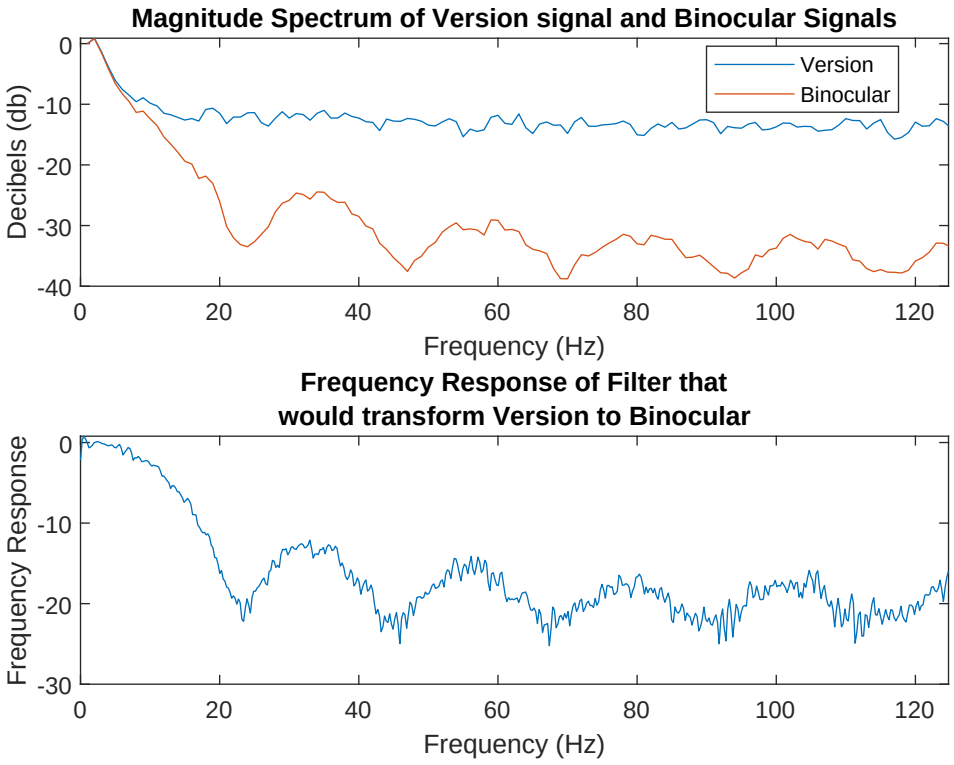

Figure 11: An illustration of the Fourier analysis of signals. The top plot is the single-sided magnitude spectra for the version signal and the binocular signal. The bottom plot is the frequency response of the filter that would transform the version signal into the binocular signal.

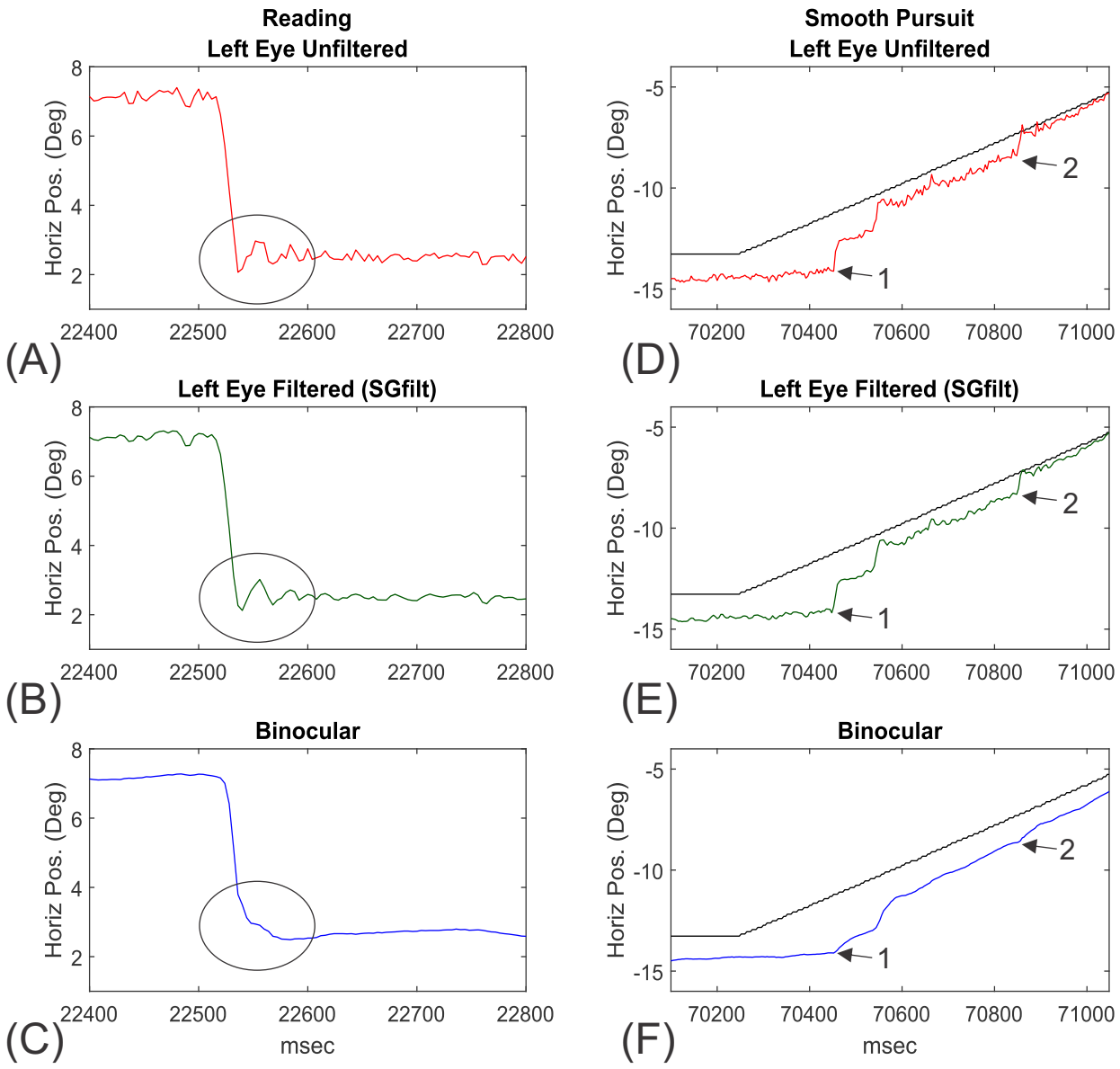

Figure 12: Illustration of saccade features in left eye and binocular signals. (A) A saccade of about 4° to the left that is followed by a marked post-saccadic oscillation. (B) The same saccade, but filtered with a Savitzky-Golay filter (order 2, window 5). (C) The binocular signal. Note the complete distortion of the end of the saccade and the eliminiation of the post-saccadic oscillation. (D) A section during the initiation of smooth pursuit ramp moving at 10° per second. Note the sharp and large initial saccade at the very start of tracking (arrow 1). Also note the catch-up saccade around 70850 ms (arrow 2). (E) Both saccade features are preserved after applying a Savitzky-Golay filter (order 2, window 5). (F) In the binocular signal, both indicated saccades are undetectable as saccades. The signals for this figure are from a new data collection.

4 Discussion

We have reported on the eye-tracking quality of SMI's ET-HMD in terms of spatial accuracy, spatial precision, temporal precision, linearity, and crosstalk. We have noted that the binocular signal is a low-pass filtered version of the individual eye signals. This result affects the choice as to which Vive signal is best for what purpose. Considering the high accuracy and precision of the binocular signal, it would appear to be the best choice if the goal was simply to identify the position of a fixation (i.e., point of regard (POR) study, or a study emphasizing human-computer interaction). However, if the goal of the study is to evaluate saccade dynamics, catch-up saccades during smooth pursuit, or the measurement of vergence, the binocular signal simply cannot be used. For a POR study, across all metrics taken together, the binocular signal with USC-1 recalibration generally would provide optimal performance. With USC-1, binocular spatial accuracy was 0.382° (H), 0.300° (V), and 0.542° (C). Binocular spatial precision was 0.054° (H), 0.059° (V), and 0.111° (C). The temporal precision of the ET-HMD, measured as the SD of ISIs, was 0.071 ms. Binocular linearity slope was 0.999 (H) and 1.005 (V). Binocular linearity fit was 0.996 (H) and 0.996 (V). We found statistically significant evidence for crosstalk often having a non-linear form (either quadratic-only or having both linear and quadratic components).

The manufacturer-supplied specifications for the ET-HMD indicated a typical spatial accuracy of 0.2°, but the details of this calculation are not available. Without recalibration, we measured the combined spatial accuracy of the binocular signal to be 0.671° on average. Similarly, the manufacturer-supplied specifications for the EyeLink 1000 indicated a typical spatial accuracy between 0.25° and 0.50° for human observers. We measured the combined spatial accuracy of the left eye signal from the EyeLink to be 1.137° on average. These large differences compared to the manufacturers' claims are part of the reason why data quality analysis is so important.

We made a number of enhancements to a traditional analysis that were effective in providing improved measures of quality. For our USC methods, the recalibration task was prepended to our random saccade task to minimize the time between recalibration and the collection of experimental data. Recalibration did generally improve spatial accuracy for all three gaze signals (left, right, and binocular) and linearity slopes for the left and right eyes. Our method for adjusting for saccade latency increased the time that each fixation overlaps with its target, increasing the amount of data available for recalibration and data quality assessment. Our novel binning procedure appeared to be successful for selecting samples to employ in the recalibration. Our non-parametric estimate of spatial precision produced measures which were robust to deviations from normality (non-Gaussianity). Our linearity assessment, including our use of the R^2_{adj} to assess the degree of overall linearity and our comparisons of the slope (and confidence limits) to the ideal slope of 1.0, provided new and useful information about the performance of our system. As far as we are aware, we are the first to assess the fit of crosstalk to both linear and quadratic functions, and we found evidence of a generally non-linear nature of crosstalk inherent in the ET-HMD. Our Fourier analysis of the signals was novel and revealed the binocular signal to be a low-pass filtered version of the individual eye signals. We were the first to precisely characterize the frequency response of such a filter.

For studies of saccade dynamics, saccades during smooth pursuit, and vergence, monocular signals will need to be employed. Our recalibration procedure produced marked enhancements in the accuracy of these monocular signals. Perfect linearity would mean that the relationship between the measured eye position and the target positions would have a slope of 1.0 and an R^2_{adj} of 1.0. Our recalibration generally brought the slopes of this relationship statistically significantly closer to the ideal slope for the monocular data, but linearity fit was generally unaffected by recalibration.

We are not aware of prior reports on the data quality of other eye-tracking VR HMDs, so there is no direct comparison to a similar device we can provide at this time. We can benchmark the performance of the ET-HMD against the much more commonly employed EyeLink 1000. Spatial accuracy is generally much better with the ET-HMD than the EyeLink, especially after recalibration. Neither spatial precision nor linearity slope are significantly different between the two devices. Linearity fit is generally significantly better for the ET-HMD than the EyeLink. We found no evidence of significant horizontal or vertical crosstalk in the EyeLink, whereas the ET-HMD generally exhibits quadratic crosstalk.

In Table 8, we compare our measures of spatial accuracy with similar measures from 5 different eye trackers as reported by Blignaut et al. [7]. Our EyeLink 1000 performs on par with these other machines. However, the ET-HMD performs much better than the other devices without recalibration. Since our measures of spatial precision, temporal precision, linearity, and crosstalk are unique to the present study, direct comparisons are problematic. Furthermore, the different recalibration routines applied in the Blignaut study [7] are different from those employed here, so an accurate direct comparison of calibrated results is not possible.

Table 8: Spatial accuracy across various devices, before and after some USC.

Dim	Device	Spatial accuracy*	
		Before USC	After USC
Horizontal	SMI RED 250 (250 Hz)†	0.62	0.49
	SMI RED 500 (250 Hz)†	0.69	0.44
	SMI RED 500 (500 Hz)†	0.73	0.52
	SMI Hi-Speed (500 Hz)†	0.50	**0.30**
	Tobii TX300 (300 Hz)†	0.79	0.31
	EyeLink 1000 (1000 Hz)‡	0.68	-
	ET-HMD (250 Hz)‡	**0.38**	0.38
Vertical	SMI RED 250 (250 Hz)†	0.85	0.63
	SMI RED 500 (250 Hz)†	0.62	0.47
	SMI RED 500 (500 Hz)†	0.75	0.58
	SMI Hi-Speed (500 Hz)†	0.51	0.31
	Tobii TX300 (300 Hz)†	0.64	0.35
	EyeLink 1000 (1000 Hz)‡	0.77	-
	ET-HMD (250 Hz)‡	**0.47**	**0.30**
Combined	SMI RED 250 (250 Hz)†	1.15	0.88
	SMI RED 500 (250 Hz)†	1.03	0.72
	SMI RED 500 (500 Hz)†	1.17	0.87
	SMI Hi-Speed (500 Hz)†	0.80	**0.49**
	Tobii TX300 (300 Hz)†	1.13	0.53
	EyeLink 1000 (1000 Hz)‡	1.14	-
	ET-HMD (250 Hz)‡	**0.67**	0.54

* A bold value indicates the best value in each group.

† The data from these devices come from the study by Blignaut et al. [7]. Their participant-controlled USC was used for the "After USC" values.

‡ The data from these devices come from the present study. Our USC-1 recalibration method was used for the "After USC" values.

4.1 How our spatial accuracy method compares to the literature

We employed the standard definition of spatial accuracy described by Holmqvist et al. [22], but our method of choosing which samples to include in our calculation of spatial accuracy differs from that employed in the literature. We computed spatial accuracy during a random saccade task that spanned 30° horizontally and 20° vertically, where 30 fixations occurred with random durations between 1000-1500 ms. This choice provided us with 30 fixations, distributed randomly across the field of view, with which to base our accuracy estimate on. Other researchers typically evaluate accuracy on a smaller subset of calibrations points. We minimized the impact of saccade latency (see Sect. 2.5) and used samples in the 400-900 ms interval after each target movement for our calculations. Gaze samples were screened based on their Euclidean distance to the gaze position centroid in 2 steps. In the first step, samples with a distance-to-centroid outside Tukey's fences [53] were ignored. In the second step, samples that were more than 2° away from the centroid were ignored in our calculations.

For comparison, Tobii [52] used a 9-point grid, presented each target for 2 s, and used data in the 800-1800 ms interval after each target movement. Tobii required that 80% of samples must be valid (non-missing), the SD of the sample positions must not exceed 1.5°, and the gaze centroid must be no more than 5° away from the target position. If the set of samples collected for any of the 9 points did not meet these criteria, they recollected the data for those point(s) up to 3 times until the criteria were met.

Blignaut et al. [7] computed accuracy across 40 targets (8 columns and 5 rows), but used very little outlier screening. This may have been justified based on their method of determining when fixation occurred. Fixations were timed based on a mouse click by the subject, which some have concluded results in improved estimation of fixation position [39].

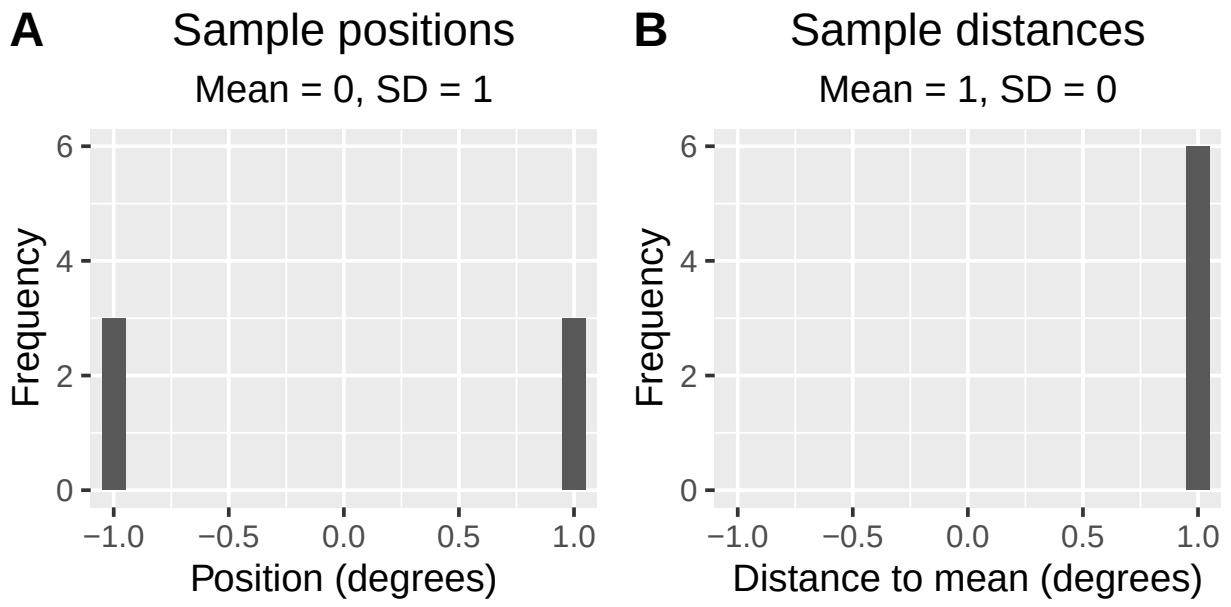

Figure 13: Consider the set of samples with positions $\{-1,-1,-1,1,1,1\}$. All samples are a distance of 1 unit away from their central tendency (i.e., 0). (A) The SD of sample positions is 1. (B) The SD of sample distances is 0.

Blignaut et al. [7] required that the gaze centroid must be within 5° from the target position, otherwise the mouse click was not accepted.

4.2 How our spatial precision method compares to the literature

First, we would like to highlight a certain confusion in the literature on precision. For example, Blignaut and Beelders [4] provide a confusing description of precision calculated using the SD approach. Consider a set of samples that are all a distance of 1 unit away from their central tendency (see Figure 13 for a simplified example). In one view, precision is the SD of a set of ***positions*** (see (A) in Figure 13). In the alternative point of view, precision is the SD of ***distances*** (see (B) in Figure 13). Blignaut and Beelders are not clear in terms of what they are taking the SD of, positions or distances. They discuss taking "the standard deviation of a set of points," suggesting that they are interested in the SD of positions. But they also provide a SD of 0 for points equidistant from their central tendency, which would only be true if they were referring to a SD of distances. The point is, contrary to the claim made by Blignaut and Beelders [4], using Euclidean distance as the distance measure in the SD formula does not produce near-zero precision values when the samples are equally spaced around the centroid, unless the SD of distances is (wrongfully) used instead of the SD of positions.

Many, if not most, investigators employ an artificial eye to get machine precision [33, 44, 52, 56], and we did not do this. Many also measure the precision of human eye tracking [6, 33, 52, 56], as we present here. Holmqvist et al. [22] suggest the preferred approach would be to use both artificial eyes and human subjects when assessing spatial precision. This would allow separate estimates of machine precision and human tracking precision.

The standard measures of spatial precision are RMS and the SD of positions [4, 22, 52]. RMS is often criticized for being heavily dependent on the sampling rate of the eye-tracking device [4, 52]. Blignaut and Beelders [4] advocate for the use of a different measure, the BCEA, claiming it is easily interpretable, independent of the sampling rate, and independent of the arrangement of samples during a fixation.

We prefer our non-parametric measure of spatial precision, i.e., the median absolute deviation of positions about the median (MAD). Like the BCEA, the MAD is easily interpretable, independent of the sampling rate, and independent of the arrangement of samples during a fixation. Unlike the BCEA which is based on a SD, however, the non-parametric MAD works well for non-normal distributions.

It should be noted that we could have also used a non-parametric approach to the measurement of spatial accuracy. However, since nearly all papers that measure spatial accuracy use the same final calculation, we felt that, for the purpose of comparability, we would employ the standard measure. On the other hand, there are already a number of

different methods to calculate spatial precision (RMS, SD, BCEA, and others [21]), so we deemed it more appropriate to use what made the most sense to us.

4.3 How our temporal precision method compares to the literature

Most investigators use the term "temporal precision" to refer to the time between the actual real-time occurrence of an eye movement and the time the movement is registered by the eye-tracking device [21]. In a real-time situation, this would include the exposure time for the digital camera, which is not trivial (on the order of 1 ms), the time to save the image, and the time to process the image to determine the position of the corneal reflection and the pupil. In the offline case, only the exposure time is relevant. We do not know the relevant times for the ET-HMD.

Herein, we used the term "temporal precision" in a different sense. We used this term to refer to the variation in sample timestamps. Some eye-tracking devices report a nominal sampling rate and provide timestamps at exactly that sampling rate (e.g., the EyeLink 1000 has a nominal sampling rate of 1000 Hz, and it produces integer timestamps in milliseconds). However, the ET-HMD provides timestamps with nanosecond precision, so we can determine the variability in timestamps to nanosecond precision. We are aware that the SMI Hi-Speed 1250 tracker also reports timestamps with microsecond precision (see the manually annotated eye-tracker data provided by Marcus Nyström).[5] So our sense of temporal precision would apply for that device as well.

4.4 How our linearity method compares to the literature

Typically, linearity is assessed by viewing a plot of spatial accuracy as a function of position [5, 7, 24]. Reulen et al. [46] employed linear regression to find the best-fitting line where gaze position was the dependent variable and target position was the independent variable. They found the best-fitting slope and intercept for this relationship, as well as the residuals for each point. Reulen et al. then expressed linearity as the maximum residual divided by the range of target positions. This approach seems to us to be overly influenced by potential outliers or extreme observations. We describe linearity both in terms of the slope of the best-fitting line (and how it compares to the ideal slope of 1.0), and in terms of the degree of fit of the line to the data using the R^2_{adj}.

4.5 How our crosstalk method compares to the literature

Typically, the assessment of crosstalk assumes linearity of the crosstalk artifact, and crosstalk is expressed as a percentage [36, 46, 47]. To assess crosstalk, we compare 4 models of crosstalk: (1) one with only a linear predictor, (2) one with only a quadratic predictor, (3) one with both linear and quadratic components, and (4) one with the intercept only. The best-fitting model, found using a stepwise regression, is chosen as the correct model. In the case of the ET-HMD, before our recalibration attempts, none of the best-fitting models were linear-only, and 3 of 6 best-fitting models were quadratic-only (see Figure 10). We think such an approach is clearly an advance and recommend that others consider it.

4.6 How our recalibration method compares to the literature

Some researchers are moving toward participant-controlled calibration routines [7, 39]. Nyström et al. [39] suggest that participant-controlled calibration leads to better estimates of fixation than algorithm-controlled or operator-controlled calibrations. In the present study, we developed our own sophisticated method for determining which samples are included from each fixation during our algorithm-controlled USCs. Future studies comparing different recalibration methods will help determine which approach is best, both in terms of the time needed to perform the USC and in terms of how well the spatial data are corrected.

4.7 Limitations

The ET-HMD and its accompanying software are no longer available (neither for purchase, technical support, nor repair), making it difficult for others to replicate our study.

Since we did not have access to an artificial eye, our measures of spatial precision and crosstalk are not measures of the characteristics of the device alone. Our spatial precision measurements include oculomotor noise, such as drift, tremor, and microsaccades. Our crosstalk measurements are sensitive to crosstalk at both the device and biological levels.

[5] `https://www.humlab.lu.se/en/person/MarcusNystrom/`.

If we had studied the same subjects on the ET-HMD and the EyeLink 1000, we would have had a much more powerful and meaningful test of the comparison between these devices. Although, for our analysis, a random saccade task was used on both devices, spanning 30° horizontally and 20° vertically, there were other technical differences which could be minimized in future studies. For example: (1) the targets, (2) the background color and target color, (3) viewing distance, (4) presence of a user-supplied calibration routine prior to the main task, and (5) whether subjects were financially and/or academically motivated to participate (yes for the EyeLink, no for the ET-HMD).

4.8 Future directions

In the future, we want to look into various options for an artificial eye, including a moving artificial eye. Mounting such an artificial eye within the ET-HMD would present a challenge. We would also like to record the same people on the ET-HMD, the EyeLink 1000, and even other devices to provide a more comprehensive evaluation across more devices. For quality control among eye-tracking VR HMDs, new measures of the quality of vergence signals and of the impact of the vergence-accommodation conflict (VAC) should be developed. As numerous studies have shown, the VAC negatively impacts oculomotor function (phoria) [40, 50], causes visual discomfort [50], leads to "simulator sickness" including headaches and nausea [17, 26], increases visual fatigue [18, 50], and reduces cognitive performance [11]. Therefore, it is reasonable to believe that the VAC would also negatively impact data quality. Various efforts are being made to reduce or outright eliminate the VAC in VR devices [26, 30]. But until these solutions become widely present in eye-tracking VR HMDs, quantifying the impact of the VAC could be useful for future studies.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144466. The study was also funded by 3 grants to Dr. Komogortsev: (1) National Science Foundation, CNS-1250718 and CNS-1714623, `www.NSF.gov`; (2) National Institute of Standards and Technology, 60NANB15D325, `www.NIST.gov`; (3) National Institute of Standards and Technology, 60NANB16D293. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Institute of Standards and Technology.

References

[1] Evgeniy Abdulin, Lee Friedman, and Oleg Komogortsev. Custom video-oculography device and its application to fourth Purkinje image detection during saccades. *arXiv e-prints*, art. arXiv:1904.07361, Apr 2019.

[2] Deepak Akkil, Poika Isokoski, Jari Kangas, Jussi Rantala, and Roope Raisamo. TraQuMe: a tool for measuring the gaze tracking quality. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '14*, pages 327–330, New York, New York, USA, 2014. ACM Press. ISBN 9781450327510. doi: 10.1145/2578153.2578192.

[3] Pieter Blignaut. Mapping the pupil-glint vector to gaze coordinates in a simple video-based eye tracker. *Journal of Eye Movement Research*, 7(1):1–11, 2014. ISSN 19958692. doi: 10.16910/jemr.7.1.4.

[4] Pieter Blignaut and Tanya Beelders. The precision of eye-trackers: a case for a new measure. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, page 289, New York, New York, USA, 2012. ACM Press. ISBN 9781450312219. doi: 10.1145/2168556.2168618.

[5] Pieter Blignaut and Tanya Beelders. TrackStick: a data quality measuring tool for Tobii eye trackers. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, page 293, New York, New York, USA, 2012. ACM Press. ISBN 9781450312219. doi: 10.1145/2168556.2168619.

[6] Pieter Blignaut and Daniël Wium. Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior Research Methods*, 46(1):67–80, mar 2014. ISSN 15543528. doi: 10.3758/s13428-013-0343-0.

[7] Pieter Blignaut, Kenneth Holmqvist, Marcus Nyström, and Richard Dewhurst. Improving the accuracy of video-based eye tracking in real time through post-calibration regression. In *Current Trends in Eye Tracking Research*, pages 77–100. Springer International Publishing, Cham, 2014. ISBN 9783319028682. doi: 10.1007/978-3-319-02868-2_5.

[8] Herve Cardot. *Gmedian: geometric median, k-median clustering and robust median PCA*, 2017.

[9] H. Collewijn, F. van der Mark, and T. C. Jansen. Precise recording of human eye movements. *Vision Research*, 15(3):447–IN5, mar 1975. ISSN 00426989. doi: 10.1016/0042-6989(75)90098-X.

[10] T N Cornsweet and H D Crane. Accurate two-dimensional eye tracker using first and fourth Purkinje images. *Journal of the Optical Society of America*, 63(8):921, 1973. ISSN 0030-3941. doi: 10.1364/JOSA.63.000921.

[11] François Daniel and Zoï Kapoula. Induced vergence-accommodation conflict reduces cognitive performance in the Stroop test. *Scientific Reports*, 9(1):1247, feb 2019. ISSN 20452322. doi: 10.1038/s41598-018-37778-y.

[12] Raymond Dodge and Thomas Sparks Cline. The angle velocity of eye movements. *Psychological Review*, 8(2): 145–157, 1901. ISSN 0033295X. doi: 10.1037/h0076100.

[13] Lee Friedman, Mark S. Nixon, and Oleg V. Komogortsev. Method to assess the temporal persistence of potential biometric features: Application to oculomotor, gait, face and brain structure databases. *PLoS ONE*, 12(6): e0178501, jun 2017. ISSN 19326203. doi: 10.1371/journal.pone.0178501.

[14] Joseph H. Goldberg and Anna M. Wichansky. Eye tracking in usability evaluation: a practitioner's guide. In *The Mind's Eye: Cognitive and Applied Aspects of Eye Movement Research*, pages 493–516. North-Holland, jan 2003. ISBN 9780080518923. doi: 10.1016/B978-044451020-4/50027-X.

[15] Stephen Gorard. Revisiting a 90-year-old debate: the advantages of the mean deviation. *British Journal of Educational Studies*, 53(4):417–430, dec 2005. ISSN 00071005. doi: 10.1111/j.1467-8527.2005.00304.x.

[16] Elias Daniel Guestrin and Moshe Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6):1124–1133, jun 2006. ISSN 00189294. doi: 10.1109/TBME.2005.863952.

[17] Jukka Häkkinen, Monika Pölönen, Jari Takatalo, and Göte Nyman. Simulator sickness in virtual display gaming. In *Proceedings of the 8th conference on Human-computer interaction with mobile devices and services - MobileHCI '06*, page 227, New York, New York, USA, 2006. ACM Press. ISBN 1595933905. doi: 10.1145/1152215.1152263.

[18] David M Hoffman, Ahna R Girshick, Kurt Akeley, and Martin S Banks. Vergence-accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, 8(3):33, mar 2008. ISSN 1534-7362. doi: 10.1167/8.3.33.

[19] Corey Holland and Oleg V. Komogortsev. Biometric identification via eye movement scanpaths in reading. In *2011 International Joint Conference on Biometrics, IJCB 2011*, pages 1–8. IEEE, oct 2011. ISBN 9781457713583. doi: 10.1109/IJCB.2011.6117536.

[20] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2): 65–70, 1979.

[21] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jardozka, and Joost van der Weijer. *Eye tracking: a comprehensive guide to methods and measures*. Oxford University Press, New York, New York, USA, 2011. ISBN 978-0-19-969708-3.

[22] Kenneth Holmqvist, Marcus Nyström, and Fiona Mulvey. Eye tracker data quality: what it is and how to measure it. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, page 45, New York, New York, USA, 2012. ACM Press. ISBN 9781450312219. doi: 10.1145/2168556.2168563.

[23] Ignace T.C. Hooge, Gijs A. Holleman, Nina C. Haukes, and Roy S. Hessels. Gaze tracking accuracy in humans: one eye is sometimes better than two. *Behavior Research Methods*, pages 1–10, oct 2018. ISSN 15543528. doi: 10.3758/s13428-018-1135-3.

[24] Anthony J. Hornof and Tim Halverson. Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods, Instruments, and Computers*, 34(4):592–604, nov 2002. ISSN 07433808. doi: 10.3758/BF03195487.

[25] Torsten Hothorn, Frank Bretz, and Peter Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, jun 2008. ISSN 03233847. doi: 10.1002/bimj.200810425.

[26] Fu-Chung Huang, Kevin Chen, and Gordon Wetzstein. The light field stereoscope: immersive computer graphics via factored near-eye light field displays with focus cues. *ACM Trans. Graph.*, 34(4):60:1—-60:12, jul 2015. ISSN 0730-0301. doi: 10.1145/2766922.

[27] Pawel Kasprowski, Katarzyna Harężlak, and Mateusz Stasch. Guidelines for the eye tracker calibration using points of regard. In *Information Technologies in Biomedicine, Volume 4*, pages 225–236. Springer, Cham, 2014. doi: 10.1007/978-3-319-06596-0_21.

[28] Frank Klefenz, Peter Husar, Daniel Krenzer, and Albrecht Hess. Real-time calibration-free autonomous eye tracker. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 762–765. IEEE, 2010. ISBN 9781424442966. doi: 10.1109/ICASSP.2010.5495004.

[29] Andrew J. Kolarik, Tom H. Margrain, and Tom C. A. Freeman. Precision and accuracy of ocular following: influence of age and type of eye movement. *Experimental Brain Research*, 201(2):271–282, mar 2010. ISSN 00144819. doi: 10.1007/s00221-009-2036-6.

[30] Gregory Kramida. Resolving the vergence-accommodation conflict in head-mounted displays. *IEEE Transactions on Visualization and Computer Graphics*, 22(7):1912–1931, jul 2016. ISSN 10772626. doi: 10.1109/TVCG.2015.2473855.

[31] Alexandra Kuznetsova, Per B. Brockhoff, and Rune H. B. Christensen. lmerTest package: tests in linear mixed effects models. *Journal of Statistical Software*, 82(13):1–26, dec 2017. ISSN 1548-7660. doi: 10.18637/jss.v082.i13.

[32] R John Leigh and David S Zee. *The neurology of eye movements*. Oxford University Press, New York, New York, USA, 4 edition, 2006.

[33] Dorion B. Liston, Sol Simpson, Lily R. Wong, Mark Rich, and Leland S. Stone. Design and validation of a simple eye-tracking system. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications - ETRA '16*, pages 221–224, New York, New York, USA, 2016. ACM Press. ISBN 9781450341257. doi: 10.1145/2857491.2857534.

[34] Michael B McCamy, Niamh Collins, Jorge Otero-Millan, Mohammed Al-Kalbani, Stephen L Macknik, Davis Coakley, Xoana G Troncoso, Gerard Boyle, Vinodh Narayanan, Thomas R. Wolf, and Susana Martinez-Conde. Simultaneous recordings of ocular microtremor and microsaccades with a piezoelectric sensor and a video-oculography system. *PeerJ*, 1:e14, 2013. ISSN 2167-8359. doi: 10.7717/peerj.14.

[35] George W. McConkie. Evaluating and reporting data quality in eye movement research. *Behavior Research Methods & Instrumentation*, 13(2):97–106, jan 1981. ISSN 1554351X. doi: 10.3758/BF03207916.

[36] John Merchant. The oculometer. Technical report, Honeywell, Inc., Boston, jul 1967.

[37] Takashi Nagamatsu, Junzo Kamahara, Takumi Iko, and Naoki Tanaka. One-point calibration gaze tracking based on eyeball kinematics using stereo cameras. In *Proceedings of the 2008 symposium on Eye tracking research & applications - ETRA '08*, page 95, New York, New York, USA, 2008. ACM Press. ISBN 9781595939821. doi: 10.1145/1344471.1344496.

[38] Takashi Nagamatsu, Junzo Kamahara, and Naoki Tanaka. Calibration-free gaze tracking using a binocular 3D eye model. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems - CHI EA '09*, page 3613, New York, New York, USA, 2009. ACM Press. ISBN 9781605582474. doi: 10.1145/1520340.1520543.

[39] Marcus Nyström, Richard Andersson, Kenneth Holmqvist, and Joost van de Weijer. The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods*, 45(1):272–288, mar 2013. ISSN 1554351X. doi: 10.3758/s13428-012-0247-4.

[40] Marius M. Paulus, Andreas Straube, and Thomas Eggert. Vergence-accommodation conflict in virtual reality displays induces phoria adaptation. *Journal of Neurology*, 264(S1):16–17, oct 2017. ISSN 14321459. doi: 10.1007/s00415-017-8425-z.

[41] R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019.

[42] Keith Rayner. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 124(3):372–422, 1998. ISSN 00332909. doi: 10.1037/0033-2909.124.3.372.

[43] Erik D Reichle, Keith Rayner, and Alexander Pollatsek. The E-Z reader model of eye-movement control in reading: comparisons to other models. *The Behavioral and brain sciences*, 26(4):445–76; discussion 477–526, 2003. ISSN 0140-525X.

[44] Eyal M Reingold. Eye tracking research and technology: towards objective measurement of data quality. *Visual Cognition*, 22(3):635–652, mar 2014. ISSN 14640716. doi: 10.1080/13506285.2013.876481.

[45] Daan R. van Renswoude, Maartje E.J. Raijmakers, Arnout Koornneef, Scott P. Johnson, Sabine Hunnius, and Ingmar Visser. Gazepath: An eye-tracking analysis tool that accounts for individual differences and data quality. *Behavior Research Methods*, 50(2):834–852, apr 2018. ISSN 15543528. doi: 10.3758/s13428-017-0909-3.

[46] J. P.H. Reulen, J. T. Marcus, D. Koops, F. R. de Vries, G. Tiesinga, K. Boshuizen, and J. E. Bos. Precise recording of eye movement: the IRIS technique part 1. *Medical & Biological Engineering & Computing*, 26(1):20–26, jan 1988. ISSN 01400118. doi: 10.1007/BF02441823.

[47] Ioannis Rigas, Hayes Raffle, and Oleg V. Komogortsev. Photosensor oculography: survey and parametric analysis of designs using model-based simulation. *IEEE Transactions on Human-Machine Systems*, 48(6):670–681, dec 2018. ISSN 21682291. doi: 10.1109/THMS.2018.2807244.

[48] David A. Robinson. A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Transactions on Bio-medical Electronics*, 10(4):137–145, oct 1963. ISSN 21681600. doi: 10.1109/TBMEL.1963.4322822.

[49] William Rosengren, Marcus Nyström, Björn Hammar, and Martin Stridh. A robust method for calibration of eye tracking data recorded during nystagmus. *Behavior Research Methods*, pages 1–15, mar 2019. ISSN 15543528. doi: 10.3758/s13428-019-01199-0.

[50] Takashi Shibata, Joohwan Kim, David M Hoffman, and Martin S Banks. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision*, 11(8):11–11, jul 2011. ISSN 1534-7362. doi: 10.1167/11.8.11.

[51] Takeshi Takegami, Toshiyuki Gotoh, and Ghen Ohyama. An algorithm for an eye tracking system with self-calibration. *Systems and Computers in Japan*, 33(10):10–20, sep 2002. ISSN 08821666. doi: 10.1002/scj.10125.

[52] Tobii. Accuracy and precision test method for remote eye trackers. Technical report, Tobii Technology AB, 2012.

[53] J. W. Tukey. *Exploratory data analysis*. Behavioral Science: Quantitative Methods. Addison-Wesley, Reading, Mass., 1977.

[54] Miguel A. Vadillo, Chris N.H. Street, Tom Beesley, and David R. Shanks. A simple algorithm for the offline recalibration of eye-tracking data through best-fitting linear transformation. *Behavior Research Methods*, 47(4):1365–1376, dec 2015. ISSN 15543528. doi: 10.3758/s13428-014-0544-1.

[55] W N Venables and B D Ripley. *Modern applied statistics with S*. Springer, New York, fourth edition, 2002.

[56] Dong Wang, Fiona B. Mulvey, Jeff B. Pelz, and Kenneth Holmqvist. A study of artificial eyes for the measurement of precision in eye-trackers. *Behavior Research Methods*, 49(3):947–959, jun 2017. ISSN 15543528. doi: 10.3758/s13428-016-0755-8.

[57] Laurence R. Young and David Sheena. Survey of eye movement recording methods. *Behavior Research Methods & Instrumentation*, 7(5):397–429, sep 1975. ISSN 1554351X. doi: 10.3758/BF03201553.