Eye tracker data quality: What it is and how to measure it

Article · March 2012

DOI: 10.1145/2168556.2168563

CITATIONS
172

READS
13,694

3 authors, including:

Kenneth Holmqvist
Nicolaus Copernicus University; Universität Regensburg; University of the Free State
179 PUBLICATIONS **6,760** CITATIONS

SEE PROFILE

Eye tracker data quality: What it is and how to measure it

Kenneth Holmqvist* Marcus Nyström† Fiona Mulvey‡

Lund University, Sweden

Abstract

Data quality is essential to the validity of research results and to the quality of gaze interaction. We argue that the lack of standard measures for eye data quality makes several aspects of manufacturing and using eye trackers, as well as researching eye movements and vision, more difficult than necessary. Uncertainty regarding the comparability of research results is a considerable impediment to progress in the field. In this paper, we illustrate why data quality matters and review previous work on how eye data quality has been measured and reported. The goal is to achieve a common understanding of what data quality is and how it can be defined, measured, evaluated, and reported.[1]

CR Categories: I.3.7 [Eye tracking]: Data quality—Standardization;

Keywords: data quality, eye tracker, eye movements, precision, accuracy, latency

1 Does data quality matter?

The validity of research results based on eye movement analysis are clearly dependent on the quality of eye movement data. The same is true of the performance of gaze based communication devices. Eye data contain noise and error which must be accounted for. There are currently no norms or standards for what researchers report about data quality in publications, or for what manufacturers report about their eye tracker's typical performance. What may be a serious impediment for one purpose may not be significant for other purposes, for example, a cheap eye tracker composed of off-the-shelf components may be sufficient for clicking large buttons in gaze interaction or for looking at larger AOIs with sufficient margin sizes, and may work as an assistive device mounted on a wheelchair, whereas a more expensive, high performance eye tracker may have better data quality and a greater number of valid eye movement measures necessary in much psychological, neurological and reading research. It is a case of matching the system to the purposes and also to the user or participant group, and this is a very difficult task without some standardized measures of data quality. If data quality is measured and characterised for the eye tracker, participant group and in terms of the specific experimental measures of interest, there are methods of dealing with low quality to maximise the validity of results: correcting or abandoning data [Holmqvist et al. 2011, p. 140 and 224]. However, these methods cannot be considered without first analysing the data and identifying what is and is not noise or error.

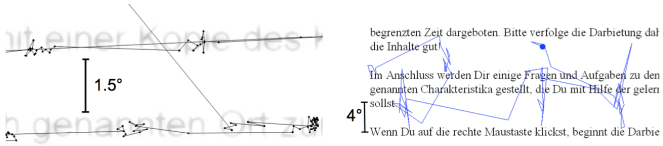
*kenneth.holmqvist@humlab.lu.se

†marcus.nystrom@humlab.lu.se

‡fiona.mulvey@humlab.lu.se

[1]We thank the members of the COGAIN Technical Committee (see www.cogain.org/EyeDataQualityTC) for the standardisation of eye data quality for their ongoing participation and comments to this text.

ETRA 2012, Santa Barbara, CA, March 28 – 30, 2012.
© 2012 ACM 978-1-4503-1225-7/12/0003 $10.00



(a) Raw samples at 50 Hz. (b) Raw samples at 50 Hz.

Figure 1: *Good and poor precision in two remote 50 Hz eye trackers as seen in an x-/y-visualisation (scanpath view). From [Holmqvist et al. 2011], page 149.*

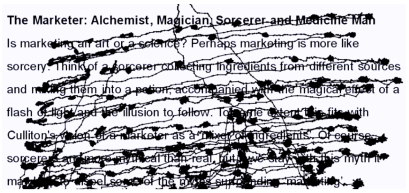


Figure 2: *Very inaccurate data in one corner. From [Holmqvist et al. 2011], page 132.*

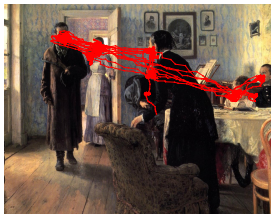
Since fixation analysis obscures the original data quality, most researchers estimate the quality of their own recordings from various plots of raw data samples. For instance, Figure 1 shows good versus poor precision, and Figure 2 a case of poor accuracy in the upper left corner. It may be obvious that eye tracker data quality affects the validity of results, but how large is the effect? Is it reasonable to assume valid results from a commercial eye tracker without measuring quality in a particular data set, or should all eye movements researchers check their data quality and report it as part of their results? To illustrate these issues, we begin with four examples.

1.1 Example 1: Effect of accuracy on dwell time measures

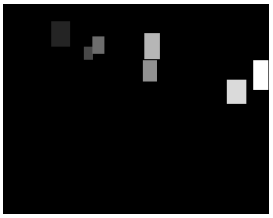
Accuracy (sometimes called offset) is one of the most highlighted aspect of data quality. Loosely speaking, it refers to the difference between the true and the measured gaze direction.

Figure 3(a) shows high quality data recorded from one participant looking at the stimulus image for 30 seconds, with the task of estimating the age of people in the scene. Binocular data were recorded with a tower-mounted eye tracker sampling at 500 Hz, but only data from the left eye are shown and analysed. The eye tracker reports an average accuracy of 0.30° horizontally and 0.14° vertically after calibration and a four-point validation procedure.

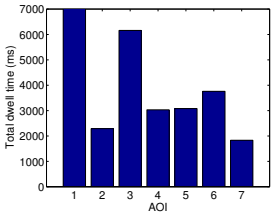
Figure 3(b) displays areas of interest (AOIs) for faces in the stimulus image. Because this is a real image, there is no whitespace—i.e. an area not covered by any AOIs—between the faces that could be used as AOI margins. AOIs with small margins are common in reading research, web studies, and studies that use videos or real world stimuli. They are also common in gaze interaction scenarios, e.g. when typing on an onscreen keyboard. When there is no room for margins, data with poor accuracy will sometimes move to another AOI than the one intended. We can simulate degrees of poor quality by adding 0.5° offset to the recorded data, moving them a bit in space. Even with this additional offset, the accuracy



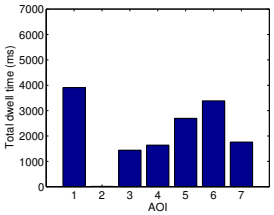
(a) Original data in high quality.



(b) AOI positions.

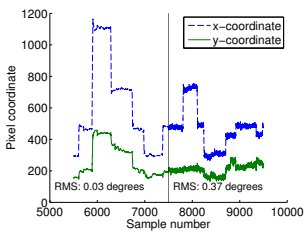


(c) Total dwell time in each AOI; original data.

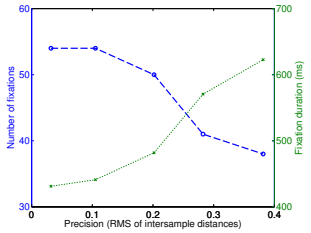


(d) Total dwell time after 0.5° inaccuracy (offset) has been added to the data.

Figure 3: *Figure (c) and (d) compare total dwell times with accurate vs slightly inaccurate data. The inaccuracy was added to the original data. Note that* 0.5° *is considered a very small error.*



(a) Illustration of data with high (left) and low (right) precision.



(b) Influence of precision on the number of fixations and the average fixation duration.

Figure 4: *How a decrease in precision affects the number and duration of detected fixations. The precision was decreased by adding Gaussian noise with an increasing variance. Fixations were detected with the algorithm by [Nyström and Holmqvist 2010], using default settings.*

is still considered rather high in comparison to what is commonly reported in the literature, in fact several manufacturers report 0.5° offset as their standard or even best possible accuracy. If a system's inaccuracy is not taken into account when designing test stimuli and analysing data from a study, what kind of effect may it have on results?

Dwell time ('gaze duration', 'glance time', ...) is the time gazed at an AOI, from entry to exit, whereas total dwell time is the sum of all dwell times to a specific AOI over a trial [Holmqvist et al. 2011, pp. 190 and 389]. It is a very common measure in eye-movement research. Figure 3(c) shows dwell times for seven AOIs based on the original, and Figure 3(d) shows dwell time for the same AOIs after a 0.5° offset has been added. Note that for some AOIs, total dwell time is reduced, for others it is significantly reduced or even totally removed from one AOI, and for some AOIs, dwell time is hardly affected at all. The effect is not uniform across AOIs and so can't be corrected or controlled for. The purpose of this example study was to analyse AOIs for dwell time and number of fixations, which is typical of many studies. Adding 0.5° degree imprecision to the data simulates many common recording scenarios. The point to note is that even when precision is relatively good, the small amount of imprecision present can lead to significant differences in the results.

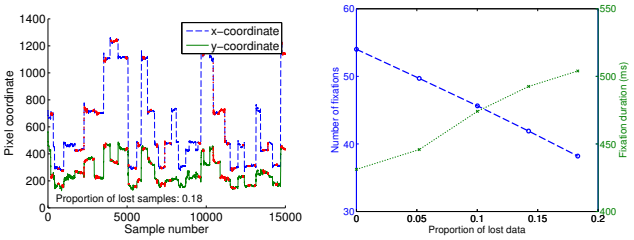
Often noise in data can be counteracted by increasing the amount of data; as, for instance, with the effect of low sampling frequency on fixation duration [Andersson et al. 2010]. In contrast, more data does not remedy the effect of poor accuracy on AOI measures such as dwell time, because the different data are likely often to be distributed in the same direction, out of the AOI.

Apart from the effect of accuracy on research results, accuracy also affects gaze based communication technologies. In gaze based interaction, interactive on screen targets are in fact AOIs with clear margins. Dwell time select is a common method of 'clicking' a button with gaze. Having several buttons side by side in an array, for example in an on-screen keyboard, will produce error in selection if data moves to the neighbouring button. When the selection method involves dwell time, this may cause an almost complete dwell-time based selection to restart, or if very inaccurate with no space between targets, may mean selection is very difficult or can only be made on very large (or magnified) targets.

1.2 Example 2: Effects of precision on the number and duration of fixations

Inaccuracy is not the only data quality issue affecting the viability of research results. While accuracy refers to the difference between true and recorded gaze direction, precision refers to how consistent calculated gaze points are, when the true gaze direction is constant. It is often tested with an artificial eye, which does not move at all. Precision measures are commonly conducted to test a particular eye tracker, and when using an artificial eye, this measure gives an idea of system noise or error, which varies with the quality of the eye tracking system. In essence, this enables us to investigate the effect of collecting data with or without a bite-bar or chin rest, or with a tower-mounted eye tracker compared to a remote one. It is also one aspect of testing eye tracker quality. By adding Gaussian noise with an increasing standard deviation to the eye movement data in Figure 3(a), we can simulate poor precision in an eye tracker. Figure 4(a) shows an example of the original data (left part of figure) and the data after noise has been added. The range of added noise has been chosen to conform to recorded precision values for current eye trackers, which according to [Holmqvist et al. 2011] is $0.01-0.05^\circ$ for tower-mounted systems and $0.03 - 1.03^\circ$ for remote ones. The larger values in the latter range, however, are likely to reflect eye trackers with exceptionally poor precision, and are therefore not included in the data presented. Precision values are calculated as the root mean square (RMS) of intersample distances in the data.

Figure 4(b) illustrates how precision influences the number and duration of fixations, as detected by the adaptive velocity algorithm developed by [Nyström and Holmqvist 2010]. According to this algorithm, fixations become fewer and longer as precision decreases. This is most likely due to the saccade detection threshold increasing as a direct consequence of the higher noise level, which prevents small saccades from being detected. These small saccades then become part of adjacent fixations, merged into one longer fixation. The effect in Figure 4(b) is dramatic; even though the data should represent exactly the same eye movement behaviour, the number of fixations decreases by more than 30%, whereas the average fixation duration increases by about 10%. The size of this effect will change with the method of event detection used. For example, it is likely that systems using dispersion based fixation detection algorithms produce a different result.



(a) Data with missing samples (indicated with red dots). 18% of the samples were lost in this example.

(b) Influence of data loss on the number of fixations and the average fixation duration.

Figure 5: *How data loss affects the number and duration of detected fixations. Data loss was simulated by randomly inserting burst losses with a length uniformly drawn from the interval* $[10, 100]$ *pixels. Fixations were detected with the algorithm by [Nyström and Holmqvist 2010], using default settings.*

1.3 Example 3: Effect of data loss on the number and duration of fixations

Lost data refers to samples that are reported as invalid by the eye tracker. Typically, this correspond to $(0, 0)$-coordinates or samples that are flagged with a certain validity code in the data file. Data losses derive from periods when critical features in the eye image—often the pupil and the corneal reflection(s)—cannot be reliably detected and tracked. This can occur when, for example, glasses, contact lenses, eyelashes, or blinks prevent the video camera from capturing a clear image of the eye.

Sometimes, it may be desirable to differentiate blinks from other sources of data loss. This may be because blinks are used as a behavioural measure (e.g. [Holland and Tarlow 1972], [Tecce 1992]) or because they are used for gaze based interaction, for example as a 'click' select input. In such cases, simply removing raw data samples with $(0, 0)$ coordinates is not possible, and blinks need to be modeled and differentiated from other causes of loss of signal. Many eye trackers do not output blinks as an event.

Figure 5(a) shows how losses have been introduced into the eye movement signal, where red dots represent lost or invalid samples. To simulate short, local losses of data, invalid data are inserted as burst losses, which occur with probability P_l and last for N_l samples, where N_l is drawn uniformly from $\mathcal{A} = \{10, 11, \ldots, 100\}$.

Figure 5(b) reveals the same trend for data loss as Figure 4(b) did for decreased precision: a reduction in the number of fixations and an increase in fixation duration.

1.4 Example 4: Effect of screen position on pupil size

Pupil size reacts primarily to changes in illumination, but it is often used as a measure of mental workload, emotional valence, or as an indication of drug use [Holmqvist et al. 2011, pp. 393–394]. A prerequisite for such investigations (apart from controlled light conditions) is that the recorded change in pupil size reflects the true change in pupil size, and therefore that the eye tracker does not add any systematic or variable error to the data. Pupil size measures will include systematic error if the apparent change in pupil size with viewing angle is not controlled for by the eye tracking system, or corrected in the recorded data subsequently. The effect of viewing angle is that pupil size is larger when the eye is on-axis with the eye camera. This typically means that the pupil is largest when looking in the centre of the screen compared to the edges. Without knowing this relationship between pupil size and screen position for the particular system being used, the difference in pupil size may be attributed to differences in cognitive processing or emotional responses to the objects. [Gagl et al. 2011] reported a similar effect and also propose a method to correct the errors. Such problems can be corrected, but only if the error is first measured for the particular set up used.

2 Factors influencing data quality

Many factors influence data quality, including:

1. *Participants* have different eye physiologies, varying neurology and psychology, and differing ability to follow instructions. Some participants may wear glasses, contact lenses, or mascara, or may have long eyelashes or droopy eyelids which all interfere with the eye image and may or may not be accounted for in a system's eye model [Nyström et al. submitted].
2. *Operators* have differing levels of skill, and more experienced operators should be able to record data with higher quality [Nyström et al. submitted]. Operator skills include adjusting eye to camera angles and mirrors, monitoring the data quality in order to decide whether to recalibrate, as well as providing clear instructions to the participants.
3. A *task* that requires participants to move around a lot, for example, could affect data quality. A task that causes participants blink more often leads to more data loss, unless blinks are modeled as eye events.
4. The *recording environment* has a strong influence on data quality. Was the data collected outdoors in sunlight or indoors in a controlled laboratory environment, for instance? Were there any vibrations in the room that reduced the stability of the eye movement signal? These factors should be considered and reported.
5. The *geometry*, that is the relative positions of eye camera, participant, and stimulus affects data quality, as does the position of the head in what is known as the head box [Holmqvist et al. 2011, p. 58]. This may be of particular importance when using eye trackers as a communication aide for the disabled, who may be constrained in their movement or sitting/lying position.
6. The *eye tracker design* does of course have a large impact on the quality of the recorded data. Simply put, an eye tracker consists of a camera, illumination, and a collection of software that detects relevant features in the eye, and map these to positions on the screen. The resolution of the video camera and the sharpness of the eye image are important factors that are directly related to some aspects of data quality. Equally important are the image analysis algorithms, the eye model, the eye illumination and the calibration procedure. Eye tracker system specifications will also have an influence on data quality. The most quoted system specification is sample rate, or sampling frequency. Sample frequency will dictate the system's ability to record brief events and to produce accurate velocity profiles. Other system specifications which influence data quality are whether the system is bright or dark pupil based (i.e. whether the eye illumination is on or off axis, producing a bright or dark image of the pupil, for a review of the various set-ups currently in use see [Hansen et al. 2011]). This may interact with eye colour or other factors to effect data quality. Finally, whether the eye tracker records monocularly or binocularly is of interest. Accuracy and precision of *fixation* data may improve if data from two eyes is combined, particularly if using a dispersion based fixation detection method, but, if data from two eyes are not separable, saccade velocity profiles, microsaccades, drift, and saccade amplitude measures will lose validity.

3 Terminology for data quality

First, let us make clear that we cannot know where a human is looking. Even when a participant says she looks at a point, the centre of the fovea can be slightly misaligned. When we talk about 'actual gaze' we refer to this subjective but reportable impression, which is what the vast majority of eye trackers are designed to measure.

Thus, in general terms, data quality can be defined as the spatial and temporal deviation between the actual and the measured gaze direction and the nature of this deviation, on a sample to sample basis. In the very simplest case, we consider these deviations in the presence of only one data sample $\hat{x}_i$. This sample can either be reported as valid or invalid by the eye tracker, where an invalid sample usually means that relevant eye features could not be detected from the video feed of the eye, for instance due to loss of the eye image. Clearly, with the exception of blinks, it does not make much sense to characterize the quality of missing data other than to classify it as invalid. When the eye tracker reports a valid sample, data quality can be defined as the distance θ_i (in visual degrees) between the actual x_i and the measured $\hat{x}_i$ gaze position, known as the *spatial accuracy*, or just accuracy, as well as the difference between the time of the actual movement of the eye t_i and the time reported by the eye tracker $\hat{t}_i$, known as *latency* or *temporal accuracy*. If both accuracy and precision differences are zero, the data quality for this single sample is optimal.

The example with only one sample is however mainly of academic interest. Typically, one needs to consider several samples recorded from a whole experiment, a trial, or a single event such as a fixation. Given n recorded samples, accuracy can be calculated as

$$\theta_{\text{Offset}} = \frac{1}{n}\sum_{i=1}^{n}\theta_i, \tag{1}$$

The variance in accuracy is often referred to as *spatial precision* and the variance in latency is typically called *temporal precision*. Two common ways to estimate the spatial precision in the eye movement signal are the standard deviation of the samples and the root mean square (RMS) of inter-sample angular distances, but a whole range of other dispersion measures exist that could be alternatives [Holmqvist et al. 2011, p. 359-369]. The standard deviation for a set of n data samples $\hat{x}_i$ is calculated as

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - \hat{x}_{\text{avg}})^2} \tag{2}$$

where $\hat{x}_{\text{avg}}$ denotes the sample average. Letting θ denote the angular distances between samples, precision can be expressed as

$$\theta_{\text{RMS}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}{\theta_i}^2} = \sqrt{\frac{\theta_1^2 + \theta_2^2 + \cdots + \theta_n^2}{n}} \tag{3}$$

These two precision calculations reflect different factors. Precision in particular reacts to vibrations in the environment when calculated as standard deviation, but not so much when calculated as root mean square (RMS). Figure 6 illustrates this important difference. It is likely that a full standard needs several precision calculations that each measure an aspect of data.

Both accuracy and precision can be computed separately for horizontal and vertical dimensions. This may be of particular significance for persons with physical disability. [Cotmore and Donegan 2011] outlines the development of a gaze controlled interface for a user who only has good control of movements in one dimension, for example. Moreover, the proportion of valid data samples recorded is often a good indication of whether the system has problems tracking a particular individual or in a particular environment.

Figure 6: *The set of raw data samples on the left have large sample-to-sample distances, and therefore RMS will be high. They are not so dispersed, so standard deviation will be low. The data set on the right, typical of a vibration in the eye tracker, has short sample-to-sample distances, which gives a low RMS, but it is fairly dispersed, so standard deviation will higher.*

The spatial accuracy and precision of pupil size can be defined in a similar manner. The unit of measurement is either pixels in the eye camera, or the perhaps more intuitive unit millimeters. Since pupil size values are recorded at the same rate as gaze samples, temporal quality values for pupil size are shared with those calculated for gaze samples.

Closely related to spatial precision is a measure termed as *spatial resolution*, which refers to the smallest eye movement that can be detected in the data. If such small eye movements are oscillating quickly, they can only be represented in data with high *temporal resolution* or *sampling frequency*, according to the Nyquist–Shannon sampling theorem [Shannon 1948].

4 Measuring data quality using an artificial eye

The artificial eye is an important and versatile tool in the assessment of data quality. However, eye trackers vary in terms of their eye models, therefore, finding an artificial eye which will 'trick' all eye trackers is difficult. When deciding which eye tracker to buy or use for a particular study, artificial eyes provide a way of comparing inherent system noise and error, and can be used to check system latency. Artificial eyes are usually available from the manufacturer, at least for systems intended for research purposes. While it is relatively simple to produce artificial eyes for systems which are dark pupil (i.e. the eye illumination is off-axis with the eye) based, it is trickier for bright pupil based systems (i.e. where the eye illumination is on-axis). Battery-equipped eyes with actively luminous pupils would be one solution.

4.1 Precision measurements with an artificial eye

Optimal precision for an eye tracker should be calculated with samples originating from a period when the eye is fixating. The only way to completely eliminate biological eye movement from the eye movement signal is to use a completely stationary eye [Holmqvist et al. 2011, p. 35-40]. Since this is not possible with actual participants, an artificial eye, which produces the corneal reflections required by the eye tracker, is usually employed. This is also how many manufacturers measure precision [SR Research 2007; Sadeghnia 2011; Johnsson and Matos 2011]. When assessing precision in real data, it is useful to know what the maximum possible precision of the system is. If system noise means that baseline precision is low, many eye movement measures may not be validly recorded. For example, the measurement of velocity profiles will be far more effected by low precision than by low accuracy, if the offset in accuracy is uniform across the screen. Likewise, low precision may effect which kind of event detection is preferable for the data. The experimental procedure is simple: first of all, calibrate

with a human eye in the normal way so that you can start recording coordinate data. Calibration of a human eye *may* introduce some small noise, so if you have a system where you can get data without first calibrating, you may do that, but be aware that the precision value will not be comparable to systems that require calibration before data recording. Then, put one or a pair of artificial eyes where the human eye(s) would have been, and make sure the artificial eyes are securely attached. Beware of vibration movements from the environment, which should not be part of your precision measurement. See to it that the gaze position of the artificial eye(s) is somewhere in the middle of the calibration area, and then start the recording. Export the raw data samples, use trigonometry and the eye-monitor distance with the physical size and resolution of the monitor to calculate sample-to-sample movement in visual degrees. Then select a few hundred samples or more where the gaze position appears to be still, and calculate the RMS or standard deviation of these samples.

Different artificial eyes tend to give slightly different RMS values for the same eye tracker. [Holmqvist et al. 2011] found RMS values of 0.021° and 0.032° on the same eye tracker when using two different artificial eyes from two manufacturers. The variance for real eyes will be even greater. Part of standardization work might be to build the specifications for a single or a set of artificial eyes that can be used on all eye trackers. This may include a variation in the colour of the artificial iris, as well as the possibility of having a reflective artificial retina (to test bright-pupil detection based eye trackers, i.e. those systems where the infra red light source is placed on-axis).

Testing only with an artificial eye may be misleading, however. The artificial eyes do not have the same iris, pupil, and corneal reflection features as human eyes, and may be easier or more difficult for the image analysis algorithms in the eye tracker to process. Also, in actual eye-tracking research, real eyes tend to vary greatly in terms of image features that cannot be simulated with artificial eyes. Therefore, some manufacturers compliment the artificial eye test with a precision test on a human population with a large variation in eye colour, glasses, and contact lenses, as well as ethnic background, having them fixate several measurement points across the stimulus monitor. The full distribution of precision values from such a test across many measurement points and participants is an important indicator of what precision you can expect in actual recordings, and its average defines the *typical precision*. The drawback is that this data includes oculomotor noise, and therefore both human and artificial eyes are needed.

4.2 Pupil diameter quality measurements with artificial eyes

The quality of pupil diameter data is also typically measured using artificial eyes. There are three such data quality measures. First, pupil *precision* is calculated as the RMS on a sequence of pupil diameter samples recorded from an artificial eye.

Pupil *accuracy* can be measured by presenting the eye tracker with artificial eyes that have known pupil diameters (such as 2, 3, and 4 mm). If all eyes are presented at the same distance, the eye tracker should output a line of diameters proportional to the input. For instance, the pupil dilation value for the 2 mm artificial pupil should be 50 % of the diameter recorded for the 4 mm pupil. Figure 7 shows data from such a measurement.

Pupil *resolution* is the smallest detectable change in pupil dilation. It can be measured by showing artificial eyes with small differences in diameter to the eye tracker. In Figure 7, the four values around 4 mm differ with 0.1 mm. The clear proportional output shows that the eye tracker is capable of distinguishing between these dilations.

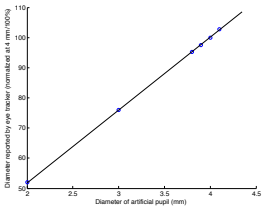


Figure 7: Pupil accuracy *means that the diameter recorded by the eye tracker should be directly proportional to the diameter of the pupil in the artificial eye. This data—from a tower-mounted eye tracker—shows a good accuracy with only minor inaccuracies.* Pupil resolution *refers to the smallest change in diameter of the artificial pupil that can be distinguished by the eye tracker. The four values around 4 mm show that this eye tracker has a pupil resolution at least on the 0.1 mm level. Data from personal communication with one manufacturer.*

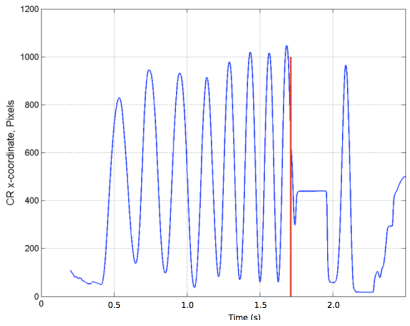


Figure 8: *Measurement of* maximum head movement speed *in a remote eye tracker. Reproduced from manufacturer document.*

4.3 Controlled motion of artificial eyes

If the artificial eye could be made to move as a real eye during saccades, fixation and smooth pursuit, it would be possible to measure optimal data quality during movements. At least one such prototype has been built, which mimics human eye movements well, except at a slightly slower speed.

However, motion of an artificial eye can be used in other ways, also. For instance, *maximum head movement speed* in a remote eye tracker can be measured by setting artificial eyes in front of the eye tracker in an increasingly rapid sinusoidal movement across the so-called head box. At some speed, tracking will be lost, which can be seen in gaze coordinates, or as in Figure 8 the x-coordinate of the corneal reflection. The maximum speed at the last oscillation before tracking is lost is the maximum head movement speed.

4.4 Switching corneal reflection

There are a number of ways that *latency* can been measured. To control the exact onset of an 'actual movement', one possibility is to turn off the infrared diode and at the same time turn on another, identical infrared diode at a different position (Figure 9). Since the time it takes to turn on (and off) the diode can be made arbitrarily small in comparison to the sampling rate of the eye tracker, the latency can be reliably measured as the time between the off- and onset of the illumination from the diodes and the corresponding change in coordinate data from the eye tracker.

The same setup for measuring latency can be used for simulating loss of tracking. The infrared illuminators are turned off, so that the eye tracker cannot detect any corneal reflections. An illuminator is

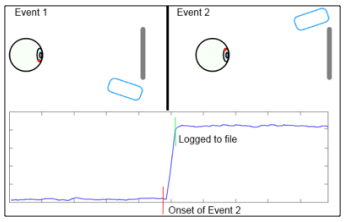


Figure 9: *Example measure of* eye-tracker latency*: an artificial eye is positioned so that gaze coordinates can be measured. A single infrared light on one side of the eye tracker is used to to create a corneal reflection. This light is turned off, and another one on the other side, immediately turned on. This will cause a immediate change in position of the corneal reflection at a time that is known by software, the time until a change in gaze coordinates has been registered is the latency. Reproduced from manufacturer document.*

then turned on again after a fixed period. The *recovery time* can then be defined as the time it takes from turning on the infrared illuminator, until the change is recorded in gaze coordinates.

5 Measuring data quality using real participants looking at targets

The factors in Section 2 must be an integral part of any design to measure data quality from human. For instance, if the purpose is to compare data quality across different set-ups or different eye trackers, characteristics of the sample group are important, and those factors listed above should be measured to compare robustness of the system to, for example, changes in eye colour or eye shape. Within an experiment, such factors may be important in terms of the relative data quality across participants, for example, is data quality significantly lower for those wearing glasses? and if so, were these participants more prevalent in one comparison group?

5.1 Calibration validation procedures

When assessing data quality for the data collected in an experiment, the issue is not to test the system performance but to assess the quality of data for each individual, for exclusion criteria, or for a particular experimental group. If standardized test reports were available for the eye-tracking system in question, the data from an experimental group could be compared to normative data for the same system. Such comparisons require independent testing across a large sample group. This work is underway but not yet complete. In the absence of such standardized measures, data quality could be assessed across experimental and control groups, to check if quality may be a confounding factor for results. Calibration procedures are proprietary to the system in question, hence testing the accuracy of calibration will mean running a subsequent calibration validation procedure. For this, targets (points) should be included between trials, at known positions, so that the data can later be assessed for accuracy and precision over the duration of recording.

Having participants look at points presented at known locations on screen is by far the most common data quality evaluation method and serves to validate the system calibration. It is essentially a repeat of the system calibration to check if inferred gaze direction matches the actual gaze direction using targets at known screen coordinates. The size, colour and shape of these targets effects resulting measures; for example, very large calibration targets will be 'hit' even when the (x, y) point at the centre of the target is quite far from the recorded gaze coordinate. The colour of the background is also important; bright backgrounds will cause the pupil to close down, which may affect accuracy [Jan Drewes 2011].

These measurement points should be placed across the stimulus presentation screen or area which the data quality values refer to. This area is typically the whole monitor, and for standardisation purposes or when testing an eye tracker (as opposed to the data recorded for a particular study), it seems reasonable to assume the monitor provided with the system is the relevant presentation area.

In many eye trackers, accuracy tends to be best in the middle of the monitor/recording area, and worst in the corners [Hornof and Halverson 2002]. If the purpose is to give a realistic account of data quality across varying stimulus presentations in future experiments or for future interfaces, then we should select measurement points at positions between calibration points, across the whole area of the monitor, varying gaze angle and position across the whole range possible when looking at the screen. Hence, the target points presented should cover the entire area used to display the experimental stimuli.

5.2 Selecting data samples to include in the calculation of data quality

As artificial eyes do not move, any samples from the recording can be used. With humans, accuracy and precision values are calculated from samples recorded when the participant is assumed to fixate a stationary target. The decision of when the eye is still is typically made by an algorithm under the assumption that a fixating eye remains relatively stable over a minimum period of time. As a consequence, the data quality values calculated from the fixation samples are directly related to the performance of the fixation detection algorithm. To date, it has been well documented that given the same set of eye movement data, fixation detection algorithms can output very different results [Karsh and Breitenbach 1983; Salvucci and Goldberg 2000; Shic et al. 2008; Nyström and Holmqvist 2010].

Even when fixations are correctly detected, one target can be associated with several fixations. This can happen due to saccadic undershoot, overshoot, or small corrective saccades and microsaccades required to align the gaze direction with the target. The researcher must then decide which fixation(s) should be included in the calculations for a given target. Figure 10 illustrates a situation where the the eye first undershoots the target (bottom left), then continues towards the target, to finally shift its position a little to the right. Three fixations are detected in this case. Including the fixation closest to the target would give the highest accuracy, but what motivates this choice of fixation over a different one? Should perhaps all be included? Could we even omit the fixation detection stage and include all samples recorded during the period when the target was shown, even though saccade samples are present? How might we account for the detrimental effects of latency in calculating precision, if latency values not reported by the manufacturer? There is a strong argument that if the researcher or interface designer will not have access to system latency values for their recorded data, that a standard measure should also assume no latency at all, and calculate precision values in the same way as the consumer will be forced to. The means of selecting which data points (raw samples) are included for calibration validation purposes should be stated as part of the research report, or manufacturer specification sheet, and the exclusion criteria should eventually be standardized for comparability across studies.

A related problem concerns how deviating fixation samples or 'outliers' due to various recording imperfections should be handled. A single outlier can significantly affect the calculated data quality values, particularly if the sampling frequency of the eye tracker is low. Perhaps the samples included for the measurement of data quality should be the same ones chosen by the system for the calculation of fixation position, since this reflects the end user situation, but fixation accuracy and sample accuracy are two different things; fixation accuracy is affected by the raw data plus event detection methods.

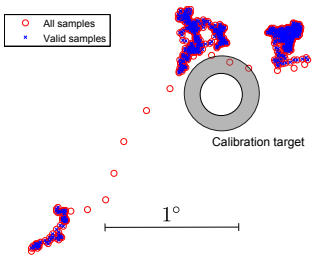


Figure 10: *Three fixations are detected (labelled 'valid samples') under the period when the participant is asked to look at the target (reproduced from [Nyström et al. submitted]). Which samples should be included in precision and accuracy calculations?*

Furthermore, removing samples raises the question of how the gaps should be treated. Whatever method is chosen, it should be fully described as part of the report on data quality.

6 How is data quality reported?

To date, researchers have rarely reported measured data quality values for their own data sets. The most common way to report accuracy is to refer to the manufacturer's specification sheets, for instance "This system is accurate to within 0.5° (citation to manufacturer)". A search on Google Scholar using "with an accuracy of 0.5" AND "eye tracker" returns 135 papers in all varieties of journals that have used this particular phrase for handing over responsibility for data quality to their particular manufacturer. The vast majority of researchers appear to treat the values on the specification sheet as a correct and objective characterization of the data they have collected, as if all data, from all participants, wherever they look at on the monitor, would have an accuracy better than 0.5° degrees of visual angle. This assumption of optimal accuracy across all recording conditions and participants is unlikely to be correct and may lead to invalid results even when data loss is accounted for and the data look reasonable.

Criteria used to exclude data which are cited in literature include, for instance, the percentage of zero values in the raw data samples, a high offset (poor accuracy) value during validation, a high number of events with a velocity above $800^\circ/\text{s}$, and an average fixation velocity above $15^\circ/\text{s}$ (indicative of low precision). For an example of accuracy and data loss criteria, see [Komogortsev et al. 2010]. [Holmqvist et al. 2011] conclude that around 2–5% of the data from a population of average non-pre-screened Europeans needs to be excluded due to participant-specific tracking difficulties. However, this number varies significantly: [Schnipke and Todd 2000], [Mullin et al. 2001] and [Pernice and Nielsen 2009] report data losses of 20–60% of participants/trials, and [Burmester and Mast 2010] excluded 12 out of 32 participants due to calibration (7) and tracking issues (5).

Manufacturer technical development groups need correct data quality values for internal benchmarking: to judge whether changes in hard- or software result in improved data quality. Therefore, several manufacturers develop data quality assessment methods for their own use. In fact, many of the methods developed by manufacturers can be expected to be essential parts in a standardization of data quality measures. This includes the artificial eye, the point-to-point measurement of latency and several suggestions for calculation that we will see below. Although there is as yet no consensus on the exact measures for data quality, the measures suggested below are nonetheless useful and informative, and could be included in a research report alongside a description of the measures chosen.

7 Reporting data quality from experiments

A standardized set of eye data quality measures could be automated for use in experimental research, as part of the software package and compared to an independent report for that eye tracking system or for a similar participant group tested on other systems. Automated data quality measures which are standardized across systems would mean that researchers can easily access them as part of running an ordinary study. They could also be made publicly available by an independent body, in a similar fashion to specifications for other computer based technologies. Table 1 shows what we propose such a report could look like in a publication.

Table 1: *Data quality report from a collection of data in an experiment. Precision values reflect the RMS of inter-sample distances.*

Data quality	Average	SD
Calibration accuracy	0.32°	0.11°
Accuracy just before end of recording	0.61°	0.27°
Calibration precision	0.14°	0.05°
Precision just before end of recording	0.21°	0.06°
Accuracy after post-recording processing	-	-
Precision after post-recording processing	-	-
Proportion of dismissed participants	9 %	-
Proportion of lost data samples in retained data	0.3 %	0.041%

In order to interpret these values in relation to the other parts of the scientific publication, it is important to specify the analysis: what event detection algorithm was used, with what settings, to detect fixations, saccades etc? What were the data exclusion criteria? What are the sizes of AOIs and their margins? Also, the type of eye tracker should be reported, alongside the recording, stimulus and analysis software with version numbers. If any values are unavailable or unknown, they should be stated as such. This is the basic level of information required in order to assess eye movement research for possible confounding variables in the data and compare research results.

7.1 Who benefits from eye data quality reports?

There are two major uses for a standardized method of testing and reporting eye data quality. We propose first, a test house tests eye trackers on the market using a battery of tests that result from a standardized set of measurements. They use both standardized artificial eyes and a large sample of real participants, standardized and selected according to the criteria known to influence data quality such as eye shape and colour. Because operator experience is a significant factor for data quality, experienced eye tracker operators should be used, or level of experience measured and controlled for. This activity results in a protocol that manufacturers can base their product documentation on. To make these results useful for gaze interaction as well as for replication in research results, the minimum target size and margin between targets viably selected by the eye can be calculated based on these values and reported alongside accuracy and precision values.

Second, authors will be able to calculate data quality values from their experimental data, and compare it to the known quality values for their eye tracker. In order to assist in the comparability of research results, journal reviewers could require these values in their papers. Such measures would greatly assist progress in the field by removing the large uncertainty in assessing the results of eye movement research using highly variant equipment and calculations. This approach would also benefit manufacturers, who need standardized measures to assess their systems and to compare performance to competitors or for internal benchmarking. Finally, it would make the task of deciding which eye tracker to buy for

particular purposes more transparent and straightforward for both researchers and users of gaze control systems.

8 Conclusion and future work

Clearly, standardization work for eye data quality would benefit eye movements technology and research in general. This work has already begun as a collaborative effort of the COGAIN Association, in the form of a technical committee for the standardisation of eye data quality. In the absence of agreed standard measures while this work is underway, there is an immediate benefit in promoting the testing and reporting of data quality as standard in eye movement research using the measures outlined above.

Not all aspects of data quality would benefit from standardization, however, there are a number of issues which might be better allowed to freely evolve, including: (a) How accuracy and precision is actually achieved, which is proprietary information and core business of manufacturers. (b) What the eye tracker can be used for? What conclusions can be drawn from the tests? This should be left up to the informed researcher or developer. (c) How can low accuracy or precision be accommodated or overcome? Standardising this would likely hold back research in the area. Magnifying windows in gaze interaction software or extra post processing of the data in research must be stated, but should not be standardized. (d) Event detection algorithms and filters used in them. Research is not mature enough.

Many researchers may be unaware of the magnitude of the effect of data quality on their research results or interface functionality and there are no guidelines on how to go about assessing their data. Likewise, manufacturers may be unsure if their in-house test methods compare to those of other manufacturers or end user's quality tests. We hope this paper sets a clear target which will have a positive impact on all aspects of eye movement research, eye tracker development and gaze based interaction.

References

ANDERSSON, R., NYSTRÖM, M., AND HOLMQVIST, K. 2010. Sampling frequency and eye-tracking measures: How speed affects durations, latencies, and more. *Journal of Eye Movement Research 3*, 6, 1–12.

BURMESTER, M., AND MAST, M. 2010. Repeated web page visits and the scanpath theory: A recurrent pattern detection approach. *Journal of Eye Movement Research 3*, 4, 1–20.

COTMORE, S., AND DONEGAN, M. 2011. ch. Participatory Design - The Story of Jayne and Other Complex Cases.

GAGL, B., HAWELKA, S., AND HUTZLER, F. 2011. Systematic influence of gaze position on pupil size measurement: analysis and correction. *Behavior Research Methods*, 1–11.

HANSEN, D. W., VILLANUEVA, A., MULVEY, F., AND MARDANBEGI, D. 2011. Introduction to Eye and Gaze Trackers. In *Gaze Interaction and Applications of Eye Tracking: Advances in Assistive Technologies*, P. Majaranta, H. Aoki, M. Donegan, D. W. Hansen, J. P. Hansen, A. Hyrskykari, and K.-J. Räihä, Eds., no. 2010. IGI Global: Medical Information Science Reference, Hershey PA, ch. 19, 288–295.

HOLLAND, M., AND TARLOW, G. 1972. Blinking and mental load. *Psychological Reports 31*, 119–127.

HOLMQVIST, K., NYSTRÖM, M., ANDERSSON, R., DEWHURST, R., JARODZKA, H., AND VAN DE WEIJER, J. 2011. *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

HORNOF, A., AND HALVERSON, T. 2002. Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods, Instruments, & Computers 34*, 4, 592–604.

JAN DREWES, ANNA MONTAGNINI, G. S. M. 2011. Effects of pupil size on recorded gaze position: a live comparison of two eyetracking systems. Talk presented at the 2011 Annual Meeting of the Vision Science Society.

JOHNSSON, J., AND MATOS, R. 2011. *Accuracy and precision test method for remote eye trackers*. Tobii Technology.

KARSH, R., AND BREITENBACH, F. W. 1983. Looking at looking: The amorphous fixation measure. In *Eye Movements and Psychological Functions: International Views*, R. Groner, C. Menz, D. F. Fisher, and R. A. Monty, Eds. Mahwah NJ: Lawrence Erlbaum Associates, 53–64.

KOMOGORTSEV, O. V., GOBERT, D., JAYARATHNA, S., KOH, D. H., AND GOWDA, S. 2010. Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Transactions on Biomedical Engineering 57*, 11, 2635–2645.

MULLIN, J., ANDERSON, A. H., SMALLWOOD, L., JACKSON, M., AND KATSAVRAS, E. 2001. Eye-tracking explorations in multimedia communications. In *Proceedings of IHM/HCI 2001: People and Computers XV – Interaction without Frontiers*, Cambridge: Cambridge University Press, A. Blandford, J. Vanderdonckt, and P. Gray, Eds., 367–382.

NYSTRÖM, M., AND HOLMQVIST, K. 2010. An adaptive algorithm for fixation, saccade, and glissade detection in eye-tracking data. *Behavior Research Methods 42*, 1, 188–204.

NYSTRÖM, M., ANDERSSON, R., HOLMQVIST, K., AND VAN DE WEIJER, J. submitted. Participants know best–influence of calibration method and eye physiology on eye-tracking data quality. *Journal of Neuroscience Methods*.

PERNICE, K., AND NIELSEN, J. 2009. *Eyetracking Methodology - How to Conduct and Evaluate Usability Studies Using Eyetracking*. Berkeley, CA: New Riders Press.

SADEGHNIA, G. R. 2011. *SMI Technical Report on Data Quality Measurement*. SensoMotoric Instruments.

SALVUCCI, D., AND GOLDBERG, J. H. 2000. Identifying fixations and saccades in eyetracking protocols. In *Proceedings of the 2002 Symposium on Eye-Tracking Research & Applications*, New York: ACM, 71–78.

SCHNIPKE, S. K., AND TODD, M. W. 2000. Trials and tribulations of using an eye-tracking system. In *CHI'00 Extended Abstracts on Human Factors in Computing Systems*, ACM, 273–274.

SHANNON, C. E. 1948. A mathematical theory of communication. *Bell System Technical Journal 27*, 379–423, 623–656.

SHIC, F., SCASSELLATI, B., AND CHAWARSKA, K. 2008. The incomplete fixation measure. In *Proceedings of the 2008 Symposium on Eye-Tracking Research & Applications*, New York: ACM, 111–114.

SR RESEARCH. 2007. *EyeLink User Manual 1.3.0*. Mississauga, Ontario, Canada.

TECCE, J. 1992. *McGraw-Hill Yearbook of Science & Technology*. New York: McGraw-Hill, ch. Psychology, physiological and experimental., 375–377.