Behav Res (2013) 45:272–288
DOI 10.3758/s13428-012-0247-4

# The influence of calibration method and eye physiology on eyetracking data quality

**Marcus Nyström · Richard Andersson · Kenneth Holmqvist · Joost van de Weijer**

Published online: 7 September 2012


**Abstract** Recording eye movement data with high quality is often a prerequisite for producing valid and replicable results and for drawing well-founded conclusions about the oculomotor system. Today, many aspects of data quality are often informally discussed among researchers but are very seldom measured, quantified, and reported. Here we systematically investigated how the calibration method, aspects of participants' eye physiologies, the influences of recording time and gaze direction, and the experience of operators affect the quality of data recorded with a common tower-mounted, video-based eyetracker. We quantified accuracy, precision, and the amount of valid data, and found an increase in data quality when the participant indicated that he or she was looking at a calibration target, as compared to leaving this decision to the operator or the eyetracker software. Moreover, our results provide statistical evidence of how factors such as glasses, contact lenses, eye color, eyelashes, and mascara influence data quality. This method and the results provide eye movement researchers with an understanding of what is required to record high-quality data, as well as providing manufacturers with the knowledge to build better eyetrackers.

**Keywords** Eyetracking · Data quality · Calibration · Eye physiology

## Why do we need eye movement data with high quality?

Holmqvist et al. (2011, p. 29) defined data quality as a "property of the sequence of raw data samples produced by the eye-tracker." Data quality is influenced by the eyetracker and the experimental setup, the participant, the operator setting up the eye image and providing instructions to the participant, and the physical recording environment, in terms of, for instance, lighting conditions. In this article, we focus on three of the most highlighted properties of data quality, which are central to obtaining valid and replicable results in oculomotor research: accuracy, precision, and the proportion of valid data samples during fixation.

Accuracy (or offset) is one of the most important properties of data quality in eyetrackers (Holmqvist et al., 2011, pp. 41–43). It refers to the distance between the actual (reference) gaze location and the recorded ($x$, $y$) position in the eyetracker data. Since the true gaze direction can only be estimated by observing external features of the eye (cf. Putnam et al., 2005), the location of a target that participants are asked to fixate can be used as the reference point. Using such a definition of accuracy includes both inaccuracy from the visual system and inaccuracy from the eyetracker, and it coheres with how accuracy is used and reported by the majority of researchers and manufacturers (Holmqvist et al., 2011; SensoMotoric Instruments, 2009; SR Research, 2007; Tobii Technology, 2011). Accuracy is of great importance in studies with small stimuli, such as reading research in which the areas of interest are close to one another, neurological research (Minshew, Luna, & Sweeney, 1999), and in gaze-input systems (Kumar, Klingner, Puranik, Winograd, & Paepcke, 2008). As an example, Rayner, Pollatsek, Drieghe, Slattery, and Reichle (2007, p. 522) stated that "there can be a discrepancy between the word that is attended to even at the beginning of a fixation and the word that is recorded as the fixated word. Such discrepancies can occur for two reasons: (a) inaccuracy in the eye tracker and (b) inaccuracy in the eye movement system." Minimizing the inaccuracy in the eyetracker maximizes the possibility of making clear scientific claims.

While an arbitrarily small accuracy can be recorded for specific participants under optimal conditions with tower-

M. Nyström (✉) · R. Andersson · K. Holmqvist · J. van de Weijer
Lund University,
Lund, Sweden
e-mail: marcus.nystrom@humlab.lu.se

mounted high-end systems operated by skilled operators, the best averages over many nonprescreened participants are around 0.3° (e.g., Jarodzka et al., 2010). Manufacturer flyers have for a long time stated that their systems have average accuracies better than 0.5°, although remote systems, in particular, are often reported as less accurate. Komogortsev and Khan (2008), for example, used an eyetracker with an accuracy specification of 0.5° but found that, after removing all invalid recordings, the average accuracy over participants was 1.0°. Zhang and Hornof (2011), Hansen and Ji (2009), and others have reported similar results. Furthermore, accuracy depends strongly on the particular characteristics of the individual participant (Hornof & Halverson, 2002), with head movements, astigmatism, and eyelid closure being particularly troublesome factors that can cause inaccuracies of several degrees of visual angle and can be detrimental to position-based data analysis and interaction. Stating the manufacturer's specifications only could therefore be directly misleading.

Precision is another important property of eyetracking data, and is defined as the ability to reliably reproduce a measurement given a fixating eye (see Fig. 1). Measured with an artificial eye, the precision of video-based eyetrackers stretches from around 0.001°–1.03° (Holmqvist et al., 2011, p. 40), where the lower values indicate that microsaccades can be reliably detected and the higher end of the scale makes the detection of fixations difficult.

The detection of any event, be it fixations, saccades, microsaccades, or smooth pursuit, is easier in data with high precision, but accuracy is largely irrelevant for event detection. In clinical applications, high precision is critical to investigate imperfections in the oculomotor system—for instance, when measuring fixation stability (see, e.g., Crossland, Culham, & Rubin, 2004; Tarita-Nistor, González, Mandelcom, Lillakas, & Steinbach, 2009) or the prevalence of square-wave jerks (Rascol et al., 1991).

Poor precision can be caused by a multitude of technical factors, mostly relating to the quality of the eye camera and of the algorithms for calculating the position of pupil and corneal reflection. Participant-specific factors such as eye color are also assumed to influence precision (Holmqvist et al., 2011, p. 43). Remedies to poor precision include filtering, but also improving the eyetracker hardware (e.g., by using a tower- or head-mounted system instead of a remote one or by increasing the resolution of the eye camera) and the recording setup (e.g., through the use of a bite board or chinrest).

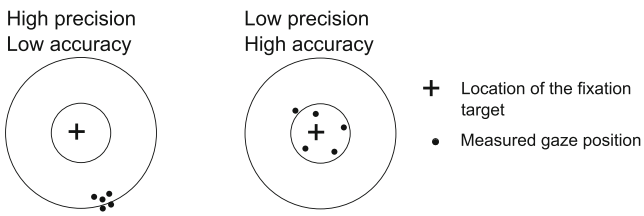


**Fig. 1** While the accuracy of an eyetracker is the (average) difference between the location of the fixation target (the position that the participant is asked to look at) and the recorded gaze position, precision is defined as the ability of the eyetracker to reliably reproduce a measurement of gaze position. Both precision and accuracy are properties of the data samples exiting the eyetracker

In the ideal situation, the eyetracker should generate valid data samples—that is, those that are captured within the tracking range of the eyetracker and have physiologically plausible values as long as the eyes are visible in the view of the camera and the eyelids are open. However, a variety of situations can cause invalid samples to be generated, such as objects occluding the pupil or corneal reflection(s), poor image analysis algorithms to detect features in the eye image, or additional features or reflections that resemble the pupil or the corneal reflection(s). Data loss can occur in long bursts or during short intervals. Irrespective of why this occurs, data loss forces the researcher to decide how to treat the gaps in the eye movement signal. For example, should they be filled with values estimated from adjacent samples, or should we accept that a fixation or saccade is split in two halves, separated by the gap? In general, it could be questioned whether an eye movement signal should be used at all if the proportion of lost data is large. Nyström and Holmqvist (2010), for instance, disregarded trials with more than 20 % of lost data samples. It should be noticed that the causes of poor data quality include imperfections from, for instance, biological, environmental, and eyetracker-related sources. Our primary interest for this article is not in the absolute values of data quality, but in how they change in relation to calibration methods, over time and space, and with respect to participants' eye physiologies.

In this article, we will address several issues related to the quality of eyetracking data and will focus on questions that are often informally discussed but seldom systematically measured and reported. The first of these issues concerns calibration, based on a comparison of three methods to ensure that the eye is still and precisely directed to a specific target during calibration. More specifically, we investigated whether the operator, an automatic procedure controlled by the system, or the participants themselves know best when their eyes are still and fixating a target. We further investigated how accuracy, precision, and the proportion of valid fixation samples vary over time and stimulus area. Finally, the influences of eye physiology, visual aids, and mascara on data quality were investigated.

## Practical aspects of calibration related to data quality

Calibration in video-based eyetracking is required to establish a mapping between features detected in the eye image and the physical orientation of the eye and/or the position of

gaze in the stimulus space (Hammoud, 2008). Calibration is typically performed by asking participants to look at a number of predefined positions in the stimulus space. At each calibration target, the eyetracker detects a number of eye image features and associates their positions in the eye image with the position of the target (Fig. 2).

Calibration is associated with several challenges, both theoretical and practical. Theoretical challenges include finding a good mathematical model for the eye and then a mapping function from eye features to gaze direction (Hammoud, 2008). In conjunction with the mathematical modeling, a number of important practical issues surround the procedure of calibration. For instance, how should the targets be presented visually on the monitor so as to guide participants' gazes efficiently and unambiguously? During monocular recordings, does it matter whether you record the dominant or nondominant eye? How can the poorer calibration accuracy in the corners of the calibration area be counteracted? Do more calibration points give a higher accuracy? Does the experience of an operator have an effect on the quality of the calibration? And do the instructions and task given to the participants before calibration influence data quality? Few of these questions have been investigated in detail.

Figure 3 provides an overview of the steps required to perform a calibration. Before calibration starts, the operator sets up the eye camera for an optimal recording of the participant's eye(s). The operator typically checks for correct pupil tracking, making certain that the entire pupil is involved when finding the center of the pupil, as well as that neither mascara-covered lashes nor glass rims are likely to become targets of the pupil-finding algorithm. The second thing that the operator looks for is the tracking of corneal reflection. Potential problems include competing reflections in glasses, a split reflection due to air bubbles under contact lenses, and potentially lost reflection if the stimulus area involves very wide visual angles that place the infrared reflection on the sclera of the eye (Holmqvist et al., 2011, chap. 4). Then the operator provides instructions for the participant: for instance, "look as precisely as you can at the points that you will see on the monitor, and do not move the eye until the point disappears." Directly before the calibration starts, the operator instructs the presentation software to show the first calibration target on the screen. After calibration is done and the operator has evaluated its result—sometimes aided by numerical results from the manufacturer software—the calibration is either accepted and the data collection can begin, or it is rejected and has to be performed all over again. The critical decision of when the eye is fixating a calibration target can be left to the system, the operator, or the participant (as was discussed by Goldberg & Wichansky, 2003).

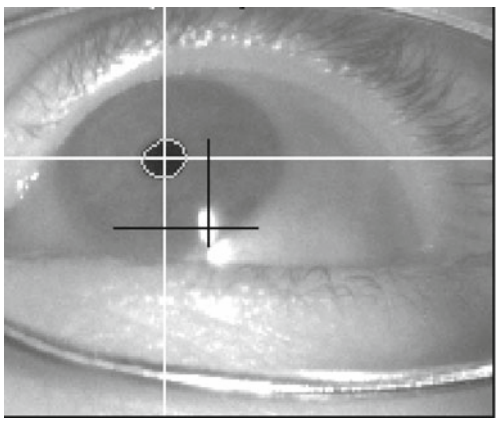


**Fig. 2** *Calibration* means establishing a mathematical mapping from features in the eye image—such as the positions of the pupil and any corneal reflection—and the position of the calibration target looked at. Because the eye is not completely still before, during, or after looking at the point, a crucial problem in calibration is to choose the right period in time to sample the coordinates of the eye image features

During system-controlled calibration, the eyetracker software automatically decides whether or not the eye is stably directed toward a calibration target. In such automatic calibration—which often relies on closed, manufacturer-specific algorithms—verification from the operator or participant is not given.

*Operator-controlled calibration* means that the operator accepts a calibration target when he or she has the impression that the participant is fixating the target. This decision can be further aided by verifying that the participant's eye is stable in the video feed of the eye image and by asking the participant for verbal confirmation.

Finally, *participant-controlled calibration* means that participants themselves decide when they are fixating a calibration target and confirm this decision by pressing a mouse or keyboard button. Leaving the control over calibration to participants may seem very natural—after all, they should know best where they are looking.

The current trend is that increasingly more control over the calibration procedure goes to the system; for instance, three of the largest eyetracker manufacturers all use system-controlled calibration by default (SensoMotoric Instruments, 2010; SR Research, 2007; Tobii Technology, 2010). Consequently, the majority of calibrations in eye-tracking research are system-controlled, even though their superiority to operator- and participant-controlled calibration remains to be proven. Below, we will examine each of these three calibration methods in detail.

*System-controlled calibration* In the system-controlled calibration method, the decisions of when the eyes are fixating and directed toward a target are made algorithmically, possibly with the exception of the first calibration point, which

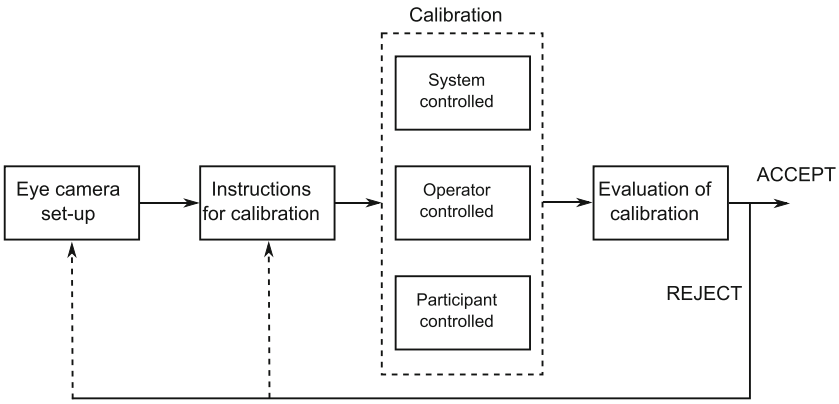


**Fig. 3** Overview of the steps required to prepare, execute, and evaluate a calibration. Note that setting up the eye camera, instructing the participants, and evaluating the results of the calibration are always the responsibilities of the operator.

may need to be accepted manually by the operator or the participant.

The benefits of a fully automatic system-controlled calibration are primarily the ease and speed with which the calibration is handled. A quick and unobtrusive calibration phase reduces the risk of directing too much attention to the calibration process and in turn of making the participant aware of the fact that eye movements are being monitored. In many experiments, the reason for calibrating is something best explained after the experiment during the debriefing session.

The decision to accept a calibration target is based on the assumption that the eye is more or less stationary over a minimum period of time when it is fixating the target. The same assumption is used when detecting fixations from raw data samples. Therefore, system-controlled decisions during calibration are associated with the same problems of detecting fixation samples as are standard fixation detection algorithms. Such problems include selecting appropriate methods and thresholds that decide when the eye is still enough to be considered a fixation, and these problems are well documented in the eyetracking literature (Nyström & Holmqvist, 2010; Salvucci & Goldberg, 2000; Shic, Scassellati, & Chawarska, 2008). The problems inherently associated with fixation detection therefore introduce uncertainty as to whether the eye-to-gaze mapping was calculated with information obtained during steady fixation.

A quick calibration procedure also makes it hard for the operator to keep up with the pace of the calibration, difficult to detect any less-than-perfect moments in the calibration, and troublesome to anticipate future problems in the recording phase. This makes it more difficult to decide whether the calibration is good enough to accept or whether the system should be recalibrated.

Moreover, an automatic calibration may not suit all participants. For instance, the EyeLink manual mentions that manually accepting calibration targets may "be useful for subjects showing difficulty fixating targets" (SR Research, 2007, p. 25).

*Operator-controlled calibration* For each target in an operator-controlled calibration, the operator visually verifies that the target is presented on the participant's monitor. Then the operator checks the eye image to verify that the participant's eye is being robustly tracked and that the gaze seems to be directed toward the target.

Manufacturers give some advice as to when to accept a target, such as "The pupil tends to come to rest gradually and to make small vergence movements at the start of the fixation, so do not respond too quickly. However, do not wait too long before accepting the fixation, as subjects soon begin to make involuntary saccades" (SR Research, 2007, p. 70).

The benefits of an operator-controlled calibration include the ability to halt the calibration upon detection of potential sources of inaccuracies, remedy them, and recalibrate. Even if the calibration works well, the careful visual inspection of the eye image during the calibration phase allows the operator to predict potential problems that may manifest later in the experiment when the participant's head or glasses shift in position. The operator is also able to accept an eye that is actually, and not just assumed to be, visually stable.

A drawback of this calibration method is the response latency due to visually inspecting the eye, determining whether to accept the calibration target, and executing the final motor action to click the "Accept" button. We know from standard reaction time tests—for instance, lexical decision tasks—that a decision takes from around 250 ms to perform (Nebes, 1978). This time may be long enough for the participant to shift his or her eyes or to make anticipatory eye movements to the next probable calibration target. If the operator blinks or makes a saccade during this buttonpress, data deriving from periods when the participant is not fixating may be recorded, and the calibration will be inaccurate.

Another drawback, which we have experienced ourselves, is the phenomenon that participants become habituated to the mouse clicks and the rate of progress of the operators through the calibration points, so that participants begin to move the eyes ahead to expected positions as soon

as a mouse click is heard or the expected acceptance time has passed. In some cases, this may lead the operator to accept a position just as the participant moves his or her eyes to a new, expected position. This is typically solved by showing calibration targets in a random order, avoiding an even acceptance rate, and providing clear instructions to the participants.

Finally, manually inspecting and verifying the eye image makes the calibration phase slow. The method also requires experienced operators to fully utilize its benefits.

*Participant-controlled calibration* In the participant-controlled calibration method, more responsibility is transferred to the participant. The participant is instructed to look at a target and to click a button whenever he or she is confident that the eye is stable and locked on the target. The click triggers the presentation of the next target, and the participant proceeds until all calibration targets have been accepted.

The benefits of the participant-controlled calibration is access to any phenomenological insight that participants have on the stability and direction of their own gaze. If they have this insight, they should be able to click at a moment when the eye is open, maximally stable, and directed at the center of the calibration target. This should lead to higher data quality. Also, placing the participant in control makes the participant more likely to be motivated to perform well during the calibration phase.

A drawback of this calibration method is that it depends on the participant's phenomenological insight, which, at least in this domain, has not been investigated for validity and reliability. Even if the average participant is reliable and cooperative, it may be that a relatively large subset of participants are unreliable and inappropriate for this method.

As with the system-controlled calibration, the operator here can only passively watch the calibration phase, with the option to recalibrate if he or she spots any potential problem during the calibration. The operator is also required to detect problems in the short window between the appearance of a calibration target and its acceptance by the participant, which, if the participant is quick, may be very short.

## The role of the operator

Setting up the eye image, instructing the participant how to behave during calibration and later recording, and deciding whether to accept or reject the calibration are always the tasks of the operator. Even though the current trend among eyetracker manufacturers is to leave increasingly more of the decisions during calibration to the system, there are reasons to believe that more experienced operators can generate data with better overall quality than can novice operators, in particular when participants with difficult eye physiologies are being recorded. Whether due to using a difficult eyetracker, a challenging task, a difficult participant population, or inexperienced operators, several examples illustrate the problem of having to discard substantial amounts of data before analysis. For example, in experiments reported by Schnipke and Todd (2000), Mullin, Anderson, Smallwood, Jackson, and Katsavras (2001), and Pernice and Nielsen (2009), 20 %–60 % of the participants or trials were excluded, whereas only 2 %–5 % of participants/trials should be expected, according to Holmqvist et al. (2011).

## The role of the participant physiology

In addition to how steadily a participant gazes at the target when sampling eye feature coordinates, the very physiology of the eye affects the stability of the eye feature data at the moment of sampling, and therefore also the quality of the calibration. In addition, some eye physiologies are more easily modeled to obtain accurate mapping functions from eye features to gaze positions. Well-known issues that make feature extraction unstable are droopy eyelids, contact lenses, bifocal glasses (Holmqvist et al., 2011, p. 118), and certain eye colors (Kammerer, 2009). It is very likely that these factors have a stronger effect on accuracy and precision than does the choice of calibration method, and the effects that one was looking for would then be concealed by much greater sources of variation. For example, a participant with mascara may have offsets far beyond normally expected levels, and it should come as no surprise that it is difficult to find subtle effects, given the presence of such a large error source.

## Data quality in space and time

The accuracy of eyetracking data is best directly after calibration, which is why many eyetrackers have built-in support for on-demand recalibration or drift correction[1] (SensoMotoric Instruments, 2009; SR Research, 2007). Only few articles, however, have described systematic investigations of the effect of data quality over time and across the visual field. A notable exception is the work by van der Geest and Frens (2002), who compared the performance of a video-based system with scleral search coils, which long has been considered the gold standard in eyetracking research. They found no systematic difference in gaze position from simultaneous recordings with the two systems, and concluded that there is "high stability in the output of the video recording system over a time course of several minutes" (p. 188).

[1] A one-point calibration that linearly shifts the data in the calibration plane.

### Research questions

The previous subsections have shown that several issues could influence the quality of data recorded with a video-based eyetracker. However, such issues are mostly discussed informally and are seldom measured and quantified. To address the lack of empirical data to back up these informal claims, we conducted an experiment to address the questions of how different calibration methods, positions of calibration targets, visual aids, and eye physiologies affect data quality in terms of accuracy, precision, and data loss. We predicted that several of these factors would have significant effects on data quality.

## Method

### Participants

A group of 149 students from the Department of Business Administration at Lund University participated in the experiment as an obligatory part of a course in English business communication. Their average age was 22.5 years ($SD$ = 2.2). To be able to investigate a larger range of possible recording difficulties, no prescreening for difficult tracking conditions (glasses, contact lenses, mascara, or eye physiology) was made while selecting the participants. Instead, data were taken prior to calibration on a variety of participant-specific properties: visual aids (including type of correction and dioptrics), eye color (brownish, bluish, or other) and brightness (bright, mixed, or dark), eye dominance[2] and hand dominance (left or right), direction of eyelashes (up, forward, or down), and presence of mascara. To limit the number of predictors, some of them have been merged (e.g., those for mascara). Table 1 gives an overview of the factor coding that was used in the statistical models.

### Operators

Six operators participated in the data recording. They all had at least three years of previous experience from running eyetracking experiments. Five of the six operators had long-term experience with the type of tower-mounted eyetracker used, while the sixth operator had only made recordings using head-mounted eyetrackers from the same manufacturer. Furthermore, participants who were judged to be difficult to calibrate[3] were assigned to the two most experienced operators for this type of system.

### Stimuli

Stimuli—the circles shown in Fig. 4—were presented with black color on a bright background. The outer diameter of each target spanned 0.5° of visual angle. These targets were shown at the same position as those used during calibration, to minimize the influence of the underlying (and unknown) method for calculating the positions of intermediate values on the calibration surface.

### Apparatus

Two computers were used to run the experiment: One, henceforth the *stimulus computer* (SC), was used to present stimuli and to interact with the participants, and the other, the *control computer* (CC), to receive and process the information from the eyetracker. The SC had a 2.2-GHz dual-core processor, 2 GB of RAM, and an ATI Radeon HD 2400 XT graphics card. The SC was connected to a Samsung Syncmaster 931c TFT LCD 19-in. monitor operating at 60 Hz with a resolution of 1,024 × 768 pixels (380 × 300 mm [31.6° × 24.0°]). Stimuli were presented with MATLAB R2009b and the Psychophysics Toolbox (Version 3.0.8, Rev. 1591; Brainard, 1997). The CC was running iView X (Version 2.4.19; SensoMotoric Instruments, 2009) adjusted for binocular pupil and corneal reflection (CR) recordings at 500 Hz, and otherwise using the default settings. Viewing and recording were binocular, with monocular gaze estimation made separately and simultaneously for each eye.

The monitor was placed $d$ = 670 mm in front of the position that the eye had when a participant's head was correctly positioned in the eyetracker's chinrest and forehead rest. See Fig. 5 for details about the setup. The parameters $h$ and $\alpha$ were chosen to mimic a typical situation[4] in which participants read text on a computer screen.

Data were recorded with four SMI HiSpeed 500-Hz systems located in different windowless basement rooms, which were prepared so as to keep the recording environments as consistent and similar as possible. Illumination came from overhead fluorescent lighting, which is known to minimize the infrared noise in the eye image (Holmqvist et al., 2011, p. 17).

[2] "The participant extends both arms and brings both hands together to create a small opening while looking at an object with both eyes simultaneously. The observer then alternately closes the left and the right eye to determine which eye is really viewing the object. That is the eye that should be measured" (Holmqvist et al., 2011, p. 199). See also Miles (1929).

[3] Difficulty was estimated subjectively, by visual inspection of the participant's eye physiology and visual aids when arriving at the experiment.

[4] In the authors' experience.

**Table 1** Overview of the predictors for statistical analysis

| Predictor | Values |
|---|---|
| (Intercept) | Not a predictor, represents the mean value of the dependent variable |
| Calibration method | **operator-controlled** (62), participant-controlled (43), system-controlled (44) |
| Off-center target | numerical variable in steps of 100 screen pixels |
| Rightward target | numerical variable in steps of 100 screen pixels |
| Downward target | numerical variable in steps of 100 screen pixels |
| Visual aids | **none** (102), contact lenses (35), glasses (12) |
| Eyelash direction | downward (8), **upward** (141) |
| Eye color | **nonblue** (35), bluish (114) |
| Mascara | yes (38), **no** (111) |
| Pupil diameter | numerical variable in camera pixels |
| Recording number | **first** (beginning of experiment), second (end of experiment) |
| Measured eye | **left**, right (left or right eye measured?) |
| Dominant eye | **left** (64), right (85) (which eye is dominant?) |
| Measured dominant | yes, **no** (is the measured eye dominant?) |

The number of participants with a certain characteristic is given in parentheses. Factor levels used as reference groups are marked with bold font. The intercept, which will be reported in the result tables that follow, represents the unweighted mean (mean of the means) of the different groups—that is, the different levels of the factors (Cohen et al., 2003; cf. p. 333, last sentence).

## Procedure

*Overview* The participants were welcomed into the experiment room and introduced to the experimental procedure and equipment. First, an online questionnaire was filled in by the operator as a means to collect the participant-specific visual and eye data. Participants were then seated in the eyetracker, and the image of the eye was optimized by the operator. The participants were then instructed to look closely at calibration targets as the targets appeared, after which they were calibrated with a method randomly selected from the three that we were testing. Then followed a short measurement of the accuracy and precision achieved, about 15 min of self-paced reading, and finally another measurement at the end of the recording session.

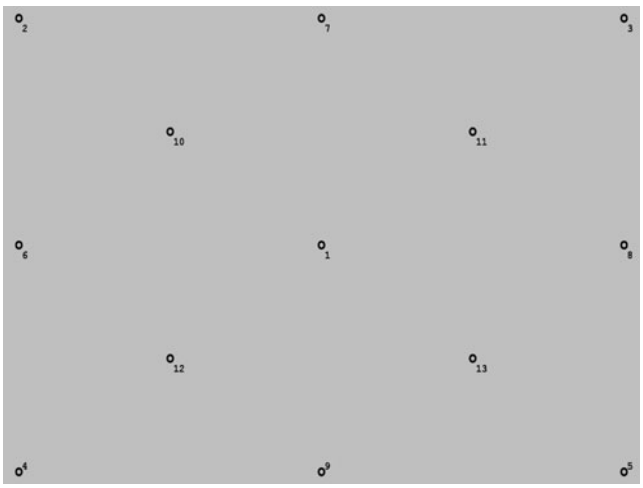


**Fig. 4** Manufacturer default order of calibration points

*Calibration* After verifying that both pupil and corneal reflection features were robustly detected in all four corners of the monitor, a 13-target binocular calibration was performed using iView X, in which calibration targets were shown one at the time in a predefined order, as is shown in Fig. 4. Both eyes were calibrated simultaneously.

In the default running mode of iView X, a calibration point can be accepted only when "valid data" have been recorded for a minimum amount of time. Data validity is controlled by a check level (strong, medium, or weak) in iView X (SensoMotoric Instruments, 2009, p. 316), with medium set as the default.

For a given participant, one calibration method was chosen at random:

1. *System-controlled calibration*, in which each calibration target was automatically accepted when a fixation had been detected.
2. *Operator-controlled calibration*, in which, as described above, the operator accepted each of the 13 calibration targets after convincing him- or herself that the participant's eye was still and directed toward the target, and that its features were correctly detected.
3. *Participant-controlled calibration*, in which the participants accepted each calibration target themselves by clicking the mouse when (they believed that) they were looking at the center of a calibration target.

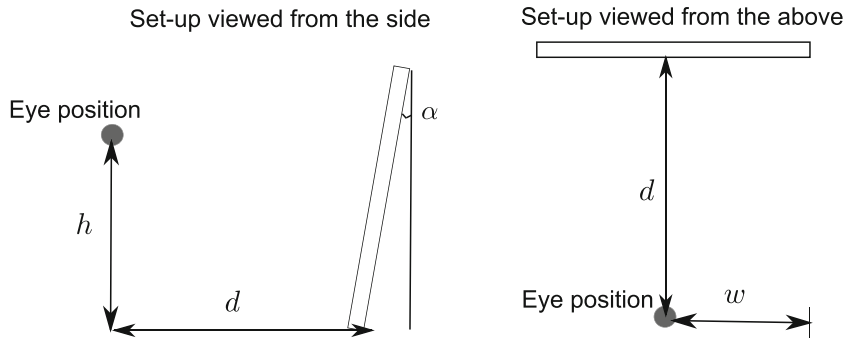


**Fig. 5** Experimental setup ($d$ = 670 mm, $\alpha = 5°$, $w = 1/2 \cdot W$ mm, $h = 3/4 \cdot H$mm). $W$ and $H$ represent the width and height of the computer screen

Since each period of stillness could automatically be accepted as a valid calibration target, it was crucial that the eyes go directly to the correct calibration target without having to search for it. Therefore, to make targets more salient and to prevent eye movements other than those directed to the correct calibration target, each dot blinked before the steady onset; it appeared for 200 ms, disappeared for 200 ms, and then reappeared for 1,500 ms or until the target was accepted and the next calibration target was shown. The instructions were the same across conditions, except for details that explained how participants were expected to click in the self-controlled version. Irrespective of the method, all participants received the instruction to look at the center of each target and to keep looking at it until it changed position.

After the calibration was done, all calibration targets were shown simultaneously, to allow the operator to inspect the calibration results visually while the participant looked at the targets and to decide whether the quality of the calibration was good enough to then record data that could be used to analyze the participant's reading behavior on a word level. The operator chose to accept the calibration and start recordings, to recalibrate using the same method, or to switch to operator-controlled calibration (in cases in which another method had initially been chosen). This last option was used when automatic or participant-controlled calibration was not possible—for example, when manual adjustment of eye feature detection thresholds was required to successfully complete the calibration. The two recalibration options were used very sparingly, however, to minimize their influence on the results. The option to override the randomly chosen calibration method in favor of manual calibration was used a total of nine times for the 149 calibrated participants.

*Data collection* Calibration was directly followed by the recording procedure, in which participants looked at the targets, one after the other in a random order, while eye movement data were collected. As during the automatic calibration, each target appeared for 200 ms, disappeared for 200 ms, and was finally displayed for 1,500 ms. Consequently, 950 data samples were recorded for each target.

Participants then read 16 text screens at their own pace, which took between 6 and 19 min. No recalibrations or drift corrections were performed during the reading phase. Finally, a second recording of data—identical to the first—was performed. All of the reading data were omitted from the analysis.

## Data analysis

*Preprocessing of the data* As measures of data quality, we calculated the proportion of valid fixation samples, accuracy, and precision, with the data from each target and eye treated separately. Fixation samples were identified as the samples that fulfilled all of the following criteria:

- They were recorded 400 ms after target onset and for as long as the target was displayed. This threshold was selected to ensure that participants had sufficient time to program a saccade and move the eye to a new location, and to reduce periods of instability in the eye-movement data due to saccadic overshoot, undershoot, or postsaccadic vergence. Moreover, this ensured that no data were included from the 200-ms period when the target was absent.
- The samples were available from the eyetracker [i.e., were not registered as $(x, y) = (0, 0)$ coordinates, which would indicate that no pupil was found—for instance, during blinks].
- They resided in the Voronoi cell belonging to the target being looked at.
- They exceeded the border of the stimulus monitor by no more than 1.5°.
- They were not saccade samples, according to the saccade detection algorithm of Engbert and Kliegl (2003), using $\lambda = 6$ and a minimum saccade duration of one sample. Conceptually, $\lambda$ controls the degree of confidence at which a velocity sample can be considered to exceed the velocity of a saccade-free period. Note that this excludes any sample with sufficiently high velocity, regardless of whether it belongs to a saccade or is a recording artifact.
- They were part of at least a 100-ms period of samples contiguous in time—that is, in which no sample had been lost.

Samples that met these criteria were included in the analyses, and henceforth will be labeled "valid fixation samples." The remaining samples were considered to be "lost" data.

Consequently, the proportion of valid fixation samples ($P_v$) was defined as

$$P_v = \frac{N_{\text{valid}}}{N_{\text{all}}}, P_v \in [0, 1] \quad (1)$$

where $N_{\text{valid}}$ and $N_{\text{all}}$ represent, respectively, the numbers of valid samples and of all samples recorded from 400 ms after target onset and for as long as the target was displayed, regardless of whether the samples were valid or not. Figure 6 illustrates the results of applying the steps above to data recorded for one target; after lost data have been removed, three fixations are detected.

*Offset* was defined as the angular distance from a recorded fixation location to the position of a target that the participant was asked to look at. If $\theta_i$ represents the offset for target $i = 1, 2, \ldots, n$, where $n$ is the number of targets, then the overall accuracy in terms of offset can formally be defined as

$$\theta_{\text{Offset}} = \frac{1}{n}\sum_{i=1}^{n}\theta_i \quad (2)$$

If more than one fixation is detected for a certain target, the one closest to the measurement target is selected. This is motivated by the fact that the eye frequently overshoots or undershoots the target, resulting in small, corrective saccades following the main saccade to the target. As a consequence, short fixations occur between these saccades that should be excluded from the offset calculations. Figure 7 shows all detected fixations during the first measurement.

*Precision* is calculated as the root-mean square (RMS) of the angular distances $\theta_i$ (in degrees of visual angle) between $m + 1$ successive data samples $(x_i, y_i)$ to $(x_{i+1}, y_{i+1})$:

$$\theta_{\text{RMS}} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\theta_i^2} \quad (3)$$

Only samples belonging to the previously selected fixation were included in the precision calculations.

Finally, median values of the vertical and horizontal pupil diameters were calculated for all samples labeled as valid, again for each eye and target separately.

*Statistical analysis* Offset, precision, and proportions of valid fixation samples were analyzed using the same approach: Linear mixed-effects models, using the lme4 package in the statistical software R (Bates & Maechler, 2010; R Development Core Team, 2009), were fit to the data. Using a mixed-effects model is superior to most traditional methods with regard to repeated measurements of participants, handling missing data, and using any combination of categorical and continuous variables (Baayen, Davidson, & Bates, 2008).

The model was fit to the data using participants and operators as random effects with random intercepts. The data were transformed in order to acquire Gaussian-looking distributions; the offset and precision were log-transformed, whereas log-odds/logit transformation was applied to the proportions of valid fixation samples. The calibration method, participant-specific properties such as eye dominance, mascara, and eyelash direction, and target placement and recording number (first or second data collection) were used as fixed effects in the model. See Table 1 for a full list. The $p$ values were calculated using a Markov-chain Monte Carlo method (MCMC) from the package languageR (Baayen, 2010), and .05 was selected as the $\alpha$ level.

Due to the nonlinear nature of log- or logit-transformed data, the magnitude of an individual factor alone will depend on what

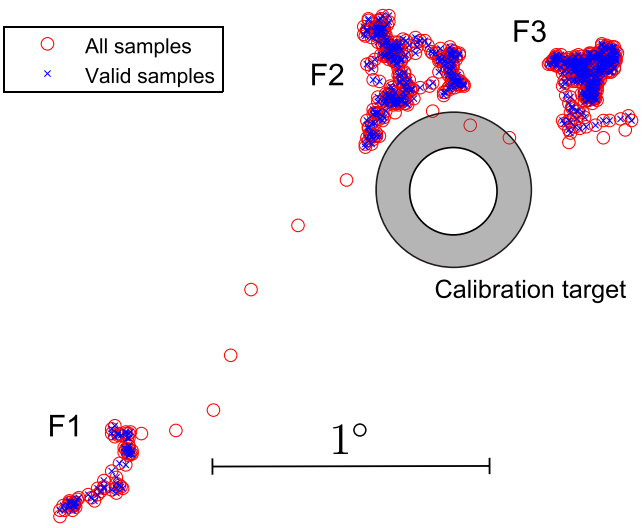


**Fig. 6** Three fixations containing valid fixation samples (×). Invalid samples are represented with empty circles (○). The fixations have been numbered (F1, F2, and F3) according to their temporal order

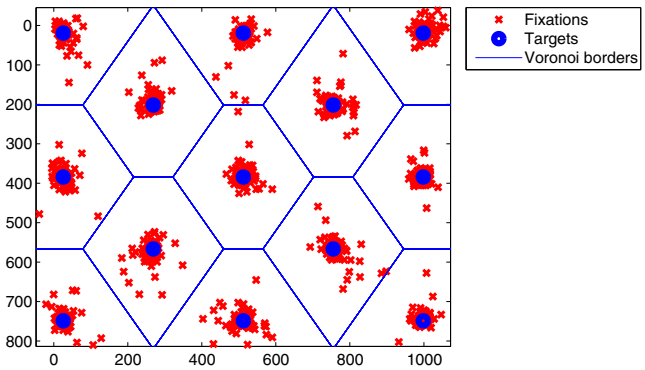


**Fig. 7** Offsets for all fixations from both eyes detected from the first measurement (directly after calibration)

other factors it is grouped with, and therefore we leave it up to the reader to back-transform[5] the particular combination of factors that are relevant to him or her. In that case, note that categorical predictors were zero-sum contrast-coded [−0.5, 0.5] and, consequently, that the intercepts represent unweighted means (Cohen, Cohen, West, & Aiken, 2003, chap. 8).

## Results

For 145 of the 149 participants we collected 52 measurements (13 targets and two eyes, at the beginning and end of experiment). For the remaining four participants, however, only data from the beginning of the experiment were available. Consequently, the total data set consisted of 7,644 measurements. A total of 650 data values (8.5 %) were excluded because no valid fixation samples could be identified for these targets. This reduced the data file to 6,994 observations.

As a check for correlation among the predictors, we calculated[6] a variance-inflation factor (VIF) for each predictor. The VIF provides an indication of collinearity among the predictors, which may lead to uninterpretable results. As a rule of thumb, the VIF should not be larger than 10 (cf. Cohen et al., 2003, p. 424). In our study, the highest value was 2.59 (for the visual-aid predictor), and the median VIF was 1.09, suggesting that collinearity was not a problem for the analysis.

A total of 146 participants were successfully calibrated and recorded. Of these, 60 participants had operator-controlled calibration, 42 had participant-controlled calibration, and 44 had automatic calibration. In all, 57.5 % of the participants had right eye dominance, whereas the rest had left eye dominance. The distributions of accuracy (offset), precision, and the proportions of valid fixation samples are shown in Fig. 8.

### Accuracy

As is summarized in Table 2, participant-controlled calibration produced significantly better accuracy (lower offset) than did operator-controlled calibration, whereas system-controlled calibration produced results marginally significantly worse than those from operator-controlled calibration.

Targets placed off-center did not differ in offset as compared to those positioned centrally. However, we found that targets to the left (as compared to the right) and to the bottom (as compared to the top, or "upward") of the screen had significantly lower offsets. Offsets were greater in the second recording phase, after reading had commenced—on average, around 0.2° larger than in the first recording. A post-hoc test using total reading time did not indicate ($p$ = .303) that the second recording produced more offset as a function of time—at least not within the reading times of this experiment (confidence interval = 5.57–18.52 min).

Contact lenses increased offset significantly, but glasses did not. Downward-pointing eyelashes had a significantly negative effect on accuracy (i.e., it became worse). We found no effect of eye color on accuracy and, surprisingly, no effect of mascara on accuracy. Pupil size did have an effect on accuracy, in that larger pupils produced significantly smaller offsets. There were no group-level differences between the left and the right eyes (Fig. 9a), nor did participants with a particular dominance perform differently during the recording phase. However, we found that dominant eyes did produce significantly less offset (Fig. 9b).

Although the role of the operator was modeled as a random effect under which the other predictors were nested, Fig. 10 indicates that it may be motivated, given enough data, to model each operator individually. A post-hoc exploration of the operators, modeled as fixed effects, revealed that one operator produced recordings that had significantly ($p$ < .001) poorer accuracy than the other operators. This effect remained significant even after controlling for multiple comparisons (number of operators = 6); this operator had extensive experience with head-mounted but not with tower-mounted systems and had only been assigned participants who had been judged easy to record high-quality data from.[7]

### Precision

As Table 3 shows, the choice of calibration method affects the average precision. More precisely, the participant-controlled calibration method yielded significantly better precision (i.e., lower RMS) than operator-controlled calibration did, whereas system-controlled calibration produced less precise measurements than did operator-controlled calibration.

Targets located off-center or farther down on the screen produced significantly higher RMSs. No horizontal asymmetries

[5] If you want to calculate, for instance, the predicted mean accuracy of the dominant eye versus its reference level (the nondominant eye), the means from the fitted model would be $y_{\text{nondom}}$ exp[intercept+(−0.5 * −0.0798)] and $y_{\text{dom}}$=exp[intercept+(0.5 * −0.0798)]. *Intercept* is the intercept value from the fitted model, –0.5 and +0.5 are the contrast codes for "nondominant" and "dominant," respectively, and the value −0.0798 is the fitted coefficient for the "measured dominant" factor. To get the output back to visual degrees (as the data were log-transformed), these expressions should be used as exponents to the natural logarithm *e*.

[6] Functions courtesy of Austin Frank at Haskins Laboratories.

[7] Probably the poorer data quality was caused by suboptimal camera angles toward the eye, which is often difficult for beginners to set up.

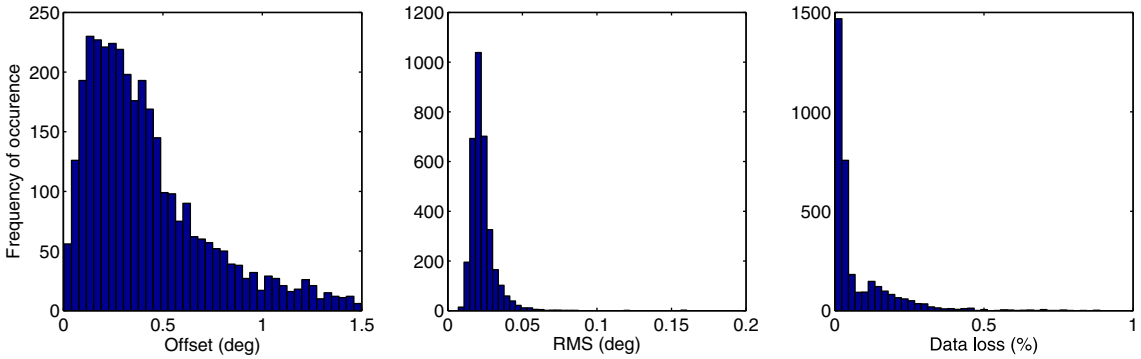


**Fig. 8** Histograms of accuracy (offset), precision (root-mean square [RMS]), and lost data (1 – proportion of valid fixation samples)

were found. Measurements from the first recording phase were significantly more precise than those from the second phase.

Concerning the participant effects, we found that participants with blue eyes had significantly worse precision than participants with brown (or "other") eyes. Glasses made measurements significantly less precise, and measurements with contact lenses actually produced significantly better precision than did using no visual aids at all. Eyelashes did not have an effect on precision, nor did mascara. The right eye—on average, over all participants and with eye dominance controlled for—did produce somewhat less precise data. Eye dominance did not have a significant effect on precision.

## Valid fixation samples

The results from the analysis of the probability to acquire valid fixation samples are summarized in Table 4. The data were analyzed as the log odds of acquiring a valid data sample.

The calibration method used did not influence the ability to record valid fixation samples from the participants, although there was a tendency for system-controlled calibrations to be less likely to generate valid data samples. Targets located off-center, as well as targets located in the lower part of the screen, were less likely to produce valid data. We found that the second recording phase was significantly less likely to capture valid data samples during target fixation.

Concerning the participant factors, we found that glasses did not differ from using no visual aids at all, but contact lenses predicted fewer valid samples acquired. Eyelash direction did not influence the ability to acquire valid data, nor did a participant's eye color or the presence of mascara. Eye dominance did not play a role, either, but we did find a main effect for the right

**Table 2** Results from the linear mixed-effects model for accuracy (offset in log-transformed degrees), shown for each predictor

| Predictor | Estimate | CI95 | *p* Value | VIF |
|---|---|---|---|---|
| Intercept | −0.4556 | (−0.7282, –0.1775) | .006** | N/A |
| Participant-controlled | −0.1681 | (−0.3292, –0.0118) | .041* | 1.55 |
| System controlled | 0.1407 | (−0.0090, 0.2913) | .068 | 1.57 |
| Off-center target | −0.0047 | (−0.0147, 0.0052) | .367 | 1.00 |
| Rightward target | 0.0065 | ( 0.0009, 0.0116) | .017* | 1.00 |
| Downward target | −0.0147 | (−0.0227, –0.0072) | <.001*** | 1.09 |
| Contact lenses | 0.3146 | ( 0.1156, 0.4977) | .002** | 2.59 |
| Glasses | 0.1975 | (−0.0633, 0.4681) | .138 | 2.57 |
| Downward eyelashes | 0.2671 | ( 0.0420, 0.4988) | .024* | 1.04 |
| Bluish eyes | −0.0430 | (−0.1670, 0.0832) | .500 | 1.06 |
| Mascara | 0.1030 | (−0.0223, 0.2417) | .121 | 1.16 |
| Pupil diameter | −0.0111 | (−0.0195, –0.0025) | .010** | 1.17 |
| Second recording | 0.4308 | ( 0.3893, 0.4700) | <.001*** | 1.07 |
| Right eye | −0.0142 | (−0.0545, 0.0240) | .482 | 1.03 |
| Right-dominant eye | 0.0408 | (−0.0733, 0.1437) | .461 | 1.16 |
| Measured dominant | −0.0798 | (−0.1572, 0.0001) | .047* | 1.03 |

Negative estimates indicate higher accuracy. The intercept represents the unweighted mean offset—that is, the mean of the means of the different groups. CI95, 95 % confidence interval; VIF, variance-inflation factor for the different predictors.

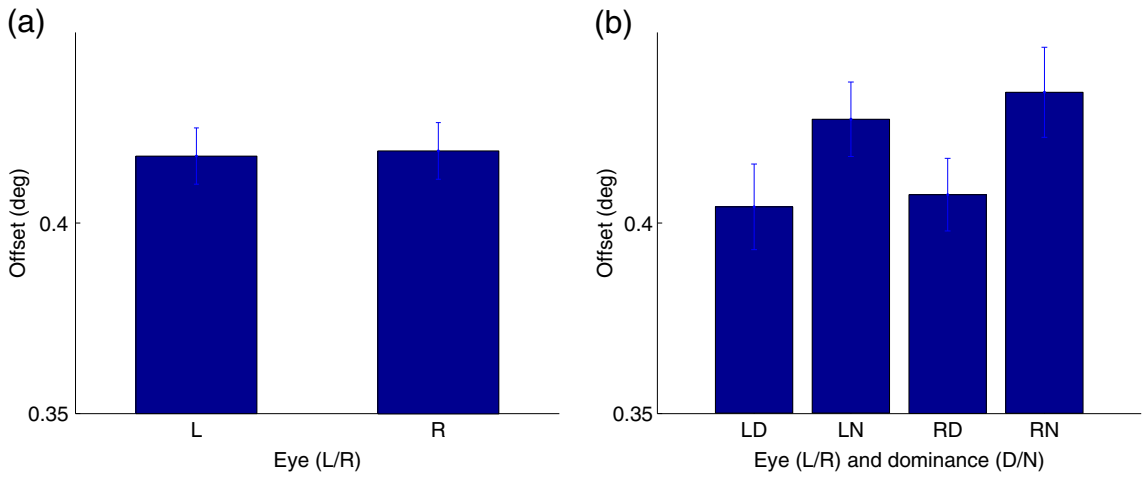


**Fig. 9** Accuracy for the left (L) and right (R) dominant (D) and nondominant (N) eyes. Error bars represent the standard errors

eye across all participants; it was slightly more likely to capture a valid fixation sample.

## Discussion

Recording eyetracking data with high quality is crucial to achieving accurate and replicable research results. While there is mostly informal knowledge of which factors influence the quality of eyetracking data, we systematically quantified how the calibration method, time, eye physiology, and operator skills affect the accuracy, precision, and proportions of valid fixation samples. These are practical issues of great concern to everyone who wants high-quality data from their eyetrackers.

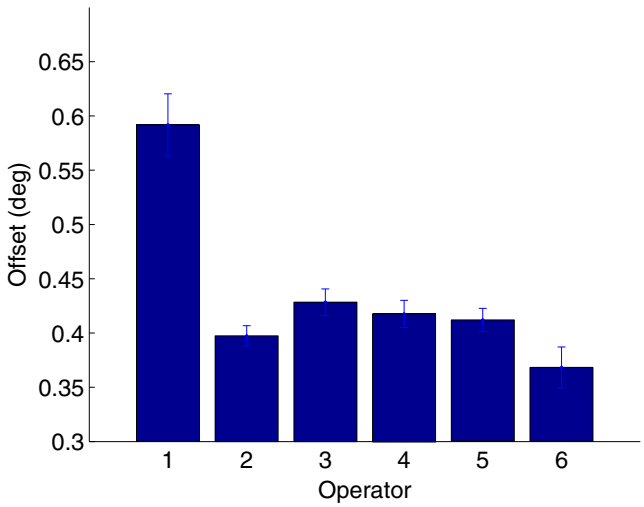


**Fig. 10** Accuracy for each operator ($M \pm SE$). The (experience of the) operator significantly influences accuracy

We found that both accuracy and precision are better for participant-controlled than for operator-controlled and, by extension, system-controlled calibration, which is currently the default calibration method in the majority of eyetracking software packages. It seems as if the participant's ability to judge the correct moment when the eye is stable and directed toward a target is better than those of the operator or of a fully automatic system, or, in other words, that the participant knows best when a specific target is being fixated.

While it can perhaps be intuitively understood why accuracy would be better for a participant-controlled calibration, the explanation with regard to an increase in precision is not quite as accessible. Why would the eye be more stable simply because the participant has been in charge of the clicking during the calibration phase? At least three explanations could be advanced: First, it is possible that the actual instruction to click a mouse button to accept a calibration target influences fixation stability, and that this effect is carried over the subsequent recording phase. Steinman, Cunitz, Timberlake, and Herman (1967), for example, suggested that an "appropriate instruction" can increase the fixation stability by reducing the microsaccadic rate. Note, however, that a change in the microsaccadic rate would not necessarily influence the precision, if the microsaccades could be detected and removed from further analysis. Second, the procedure of clicking may provide a better means of preparing, or may train, the participants for the forthcoming recording, in which they will be required to perform the same task, but without having to actively click on the targets. There is evidence that relevant training can improve the fixation stability of participants (Di Russo, Pitzalis, & Spinelli, 2003; Kosnik, Fikre, & Sekuler, 1986). Finally, the clicking procedure could make participants more aware of when and where they need to fixate in

**Table 3** Results from the linear mixed-effects model for precision (as root-mean squared [RMS] log degrees), shown for each predictor

| Predictor | Estimate | CI95 | *p* Value | VIF |
|---|---|---|---|---|
| Intercept | −3.7309 | (−3.8328, –3.6228) | <.001*** | N/A |
| Participant-controlled | −0.0844 | (−0.1559, –0.0148) | .018* | 1.55 |
| System-controlled | 0.0781 | (0.0125, 0.1460) | .019* | 1.56 |
| Off-center target | 0.0063 | (0.0038, 0.0088) | <.001*** | 1.00 |
| Rightward target | −0.0007 | (−0.0020, 0.0007) | .303 | 1.00 |
| Downward target | 0.0144 | (0.0125, 0.0163) | <.001*** | 1.11 |
| Contact lenses | −0.2457 | (−0.3266, –0.1609) | <.001*** | 2.58 |
| Glasses | 0.4094 | (0.2914, 0.5186) | <.001*** | 2.56 |
| Downward eyelashes | −0.0015 | (−0.0103, 0.0972) | .965 | 1.04 |
| Bluish eyes | 0.1315 | (0.0795, 0.1874) | <.001*** | 1.05 |
| Mascara | 0.0243 | (−0.0312, 0.0817) | .403 | 1.16 |
| Pupil diameter | −0.0114 | (−0.0137, –0.0091) | <.001*** | 1.21 |
| Second recording | 0.0165 | (0.0063, 0.0261) | <.001*** | 1.09 |
| Right eye | 0.0605 | (0.0506, 0.0697) | <.001*** | 1.03 |
| Right-dominant eye | 0.0375 | (−0.0116, 0.0841) | .121 | 1.15 |
| Measured dominant | 0.0142 | (−0.0046, 0.0337) | .144 | 1.03 |

Negative estimates indicate higher precision. The intercept represents the unweighted mean RMS—that is, the mean of the means of the different groups. CI95, 95 % confidence interval; VIF, variance-inflation factor for the different predictors.

relation to how the targets are presented and moved. In this way, the probability of including samples recorded from a fixation increases, and the samples are less likely to be contaminated by those originating from other types of eye movements or noise, which, by definition, decrease the precision in the data.

A reasonable hypothesis is that more experienced operators should be able to record data with higher quality than

**Table 4** Results from the linear mixed-effects model for the amount of valid data, shown for each predictor

| Predictor | Estimate | CI95 | *p* Value | VIF |
|---|---|---|---|---|
| Intercept | 3.0975 | (−3.3953, –2.8082) | <.001*** | N/A |
| Participant-controlled | 0.1336 | (−0.1404, 0.4050) | .341 | 1.55 |
| System-controlled | −0.2575 | (−0.5285, 0.0021) | .057 | 1.54 |
| Off-center target | −0.0713 | (−0.0869, –0.0557) | <.001*** | 1.00 |
| Rightward target | 0.0048 | (−0.0037, 0.0132) | .276 | 1.00 |
| Downward target | −0.0243 | (−0.0360, –0.0126) | <.001*** | 1.09 |
| Contact lenses | −0.4186 | (−0.7678, –0.0865) | .018* | 2.59 |
| Glasses | 0.2621 | (−0.1745, 0.7410) | .268 | 2.58 |
| Downward eyelashes | 0.1578 | (−0.2561, 0.5471) | .437 | 1.04 |
| Bluish eyes | 0.1274 | (−0.0843, 0.3517) | .248 | 1.05 |
| Mascara | −0.1281 | (−0.3629, 0.0944) | .274 | 1.15 |
| Pupil diameter | 0.0227 | (0.0091, 0.0366) | <.001*** | 1.18 |
| Second recording | −0.1049 | (−0.1670, –0.0414) | <.001*** | 1.08 |
| Right eye | 0.0938 | (0.0337, 0.1553) | .002** | 1.03 |
| Right-dominant eye | 0.1165 | (−0.0770, 0.3098) | .233 | 1.12 |
| Measured dominant | −0.0157 | (−0.1336, 0.1067) | .802 | 1.03 |

Negative estimates indicate a lower number of valid data samples. The scale is in log odds (logits) of acquiring a valid data sample, and the intercept represents the unweighted mean logit—that is, the mean of the means of the different groups. CI95, 95 % confidence interval; VIF, variance-inflation factor for the different predictors.

could novices, despite manufacturers' claims of "no operator experience needed."[8] With a high-speed tower-mounted eyetracker, in particular, there are many degrees of freedom to change mirrors and cameras in order to set up a good eye image, which is a skill that develops with time. We found that the least experienced operator for this particular system produced data with the poorest quality in terms of accuracy, even though all operators had significant experience of eyetracking in general. The poorer accuracy of the less experienced operator, although significant, stems from a single case study, and the operator was assigned participants who were judged as easy to record high-quality data from. This expected result is very relevant for training and competence development in eyetracking laboratories, but is in need of further investigation. We expect that the difference among operators with respect to data quality will become even larger when comparing experienced operators with first-time users. However, operator experience is problematic to define precisely, because there is no clear-cut path of development for each operator. Each person develops his or her own set of skills with his or her own bag of tricks to cope with problems such as mascara, eyelashes, and unwanted reflections. We would expect some interaction between operator experience and the problematic eye factors of the participants. However, the vast number of combinations prevented this kind of analysis in our present data set.

Data quality is directly related to the quality of the eye image and to how robustly features can be extracted from it, and everything that prevents information from being accurately captured by an eye camera is therefore a potential threat to recording high-quality data. Such factors can be both external—coming from, for instance, glasses, contact lenses, eyelashes, and mascara—and internal—stemming from the size of the pupil or the structure and color of the iris. We found that contact lenses produced data with significantly larger offsets as compared to the data from participants without visual aids. One explanation for this is that the lens may slip in relation to the eye, and thereby introduce air bubbles or other types of distortions that change the eye image in relation to what it looked like during calibration. Contrary to our expectations, we found that glasses did not produce less accurate data. This could reflect the fact that glasses remain stable in relation to the eye and to the eye camera; even if the glasses distort the shapes and locations of the eye features, similar distortions are present during calibration.

Glasses did produce less precise data. This was expected since, in our experience, some glasses—possibly ones with antireflection or antiscratch coatings—absorb some of the infrared light, making the eye appear slightly darker. This could, in turn, provide a less distinct border between the pupil and the iris, making the pupil more difficult to detect robustly. Contact lenses produced, to our initial surprise, a distinctly lower RMS. However, the reason is simply that the operators trained at our lab purposefully defocus the camera on the eyetracking system when recording participants with contact lenses. The reason is that ill-fitting lenses may have small air bubbles under the lens, and these may in certain positions end up right where the corneal reflection appears. The effect is that the corneal reflection is split into several smaller reflections, which in turn produces much larger problems when the system switches between the different competing reflections. The result is a very low precision ("jittery" fixations). Defocusing the camera merges the small reflections into a larger one, which can be used to more accurately calculate the center point of the corneal reflection. The fact that the center of a large object (pupil or corneal reflection) in the eye image can be computed with higher accuracy than the center of a small object is also reflected in the pupil size results; precision increases as the pupil becomes larger. In line with the investigation of Holmqvist et al. (2011, p. 42), we found an effect of eye color on precision; data recorded from participants with a bluish eye color had significantly lower precision than did data recorded from other participants. Supposedly, the contrast between iris and pupil becomes smaller when tracking blue-eyed participants with a dark-pupil eyetracker.

The effect of target location on data quality was somewhat inconclusive. Placing targets to the right seemed to produce higher offsets than did targets placed on the left part of the screen. Placing targets farther down on the screen seemed to improve accuracy, despite the fact that lower targets usually have more problems with eyelashes interfering with the eye video. Off-center targets as well as lower targets produced significantly worse precision. One explanation could be that this effect is driven by the decrease in pupil size that a larger gaze angle gives (cf. Gagl, Hawelka, & Hutzler, 2011). We know from the results in this article that such a decrease in pupil size reduces precision. The decrease in precision for lower targets could also be explained by an interaction with eyelashes; if the eyelashes are superimposed on the border between the pupil and the iris in the eye image, this may decrease the stability of pupil contour detection. Off-center and low targets produced significantly fewer valid data samples. Again, this could be related to eyelashes making feature detection less robust for targets positioned in the lower part of the screen, or to large gaze angles placing the corneal reflection on the sclera, and therefore rendering it difficult to track.

Accuracy deteriorated with time; it was on average 0.23° better directly after calibration than after 6–19 min of reading. A large part of the increase in offset likely occurred because of movements of the participants' heads and

[8] The quote is taken from the SMI RED250 flier, downloaded August 6, 2011, from www.smivision.com/.

postures over the course of the experiment. Similar effects were found for precision and for the amount of valid data acquired, where the second measurement phase produced significantly lower-quality values. A separate post-hoc analysis using reading time as a predictor did not show a connection between the time spent reading and the amount of degradation in the quality values. This means that the reduction in data quality is associated with certain events, such as sneezing, that can occur at any time during the experiment, and that additional events do not further decrease the data quality.

Other studies have found a difference in accuracy originating from eye dominance—for instance, Marmitt and Duchowski (2002) reported that in most cases, calibration was better for the participants' dominant eye (self-reported), but they did not define what "better" means. In contrast, Cui and Hondzinski (2006) found no evidence that eye dominance influenced accuracy. However, both of these studies used rather few participants (nine and six, respectively) and low-speed eyetrackers (60 Hz), making the analyses potentially insensitive to small differences between the dominant and the nondominant eyes.

We found smaller offsets for the dominant eye than for the nondominant eye. Since both eyes were calibrated simultaneously and independently, this finding does not necessarily mean that the dominant eye has better accuracy per se. Rather, it means that the dominant eye more reliably can realign its gaze direction to a previously shown target—that is, the target shown during calibration—than can the nondominant eye. Notice that, in general, we should not expect to find a perfect alignment between the eyes, since there is natural disparity between them. This is known, for instance, from studies investigating binocular coordination and disparity during reading. As a methodological note, Liversedge, White, Findlay, and Rayner (2006) recommended that monocular calibration should be conducted on each eye independently when studying binocular coordination, whereas Nuthmann and Kliegl (2009) argued that while binocular calibration may affect the magnitude of fixation disparity, it is unlikely to influence its direction.

We used a single tower-mounted eyetracking system based on the principle of dark-pupil eyetracking, and do not exclude the possibility that other recording techniques may provide results different from the ones found in this study (see, e.g., Smeets & Hooge, 2003, who found higher variability in the peak velocities and amplitudes of saccades recorded with scleral search coils as compared to saccades recorded with a video-based system). We also cannot exclude the possibility that other manufacturers, producing eyetrackers similar to the one that we tested, have implemented better calibration procedures, which display calibration targets so that participants look more steadily and exactly at them and can make more qualified decisions on when to sample feature values from the eye image. A calibration in which the participant not only clicks the keyboard or mouse button, but points at the calibration target while clicking, may help to keep up the motivation throughout a large number of calibration targets. This type of clicking to calibrate is not used as the default option in any commercial systems, however. Instead, the trend among manufacturers has been to develop calibration methods with as few targets as possible. How many calibration targets to use, how they should be presented to optimally guide participants' gazes, and whether or not to click on the targets are largely open questions that will require additional research.

## Conclusions

We have provided the first comprehensive set of data showing how the calibration method, the operator, participants' eye physiologies, and visual aids affect the quality of data recorded with a video-based eyetracker using the principle of dark-pupil and corneal-reflection tracking. Many of the factors that we investigated had a significant effect on the data quality, and we argue that the results presented in this article are of great importance to everyone who wants to record high-quality data to answer fine-grained questions about the psychological or neurological nature of the oculomotor system. We welcome further research replicating these findings on a wider range of eyetrackers.

**Author note** The authors thank the operators who helped collect the data: Kerstin Gidlöf, Nils Holmberg, and Roger Johansson. The School of Economics and Management, Henrik Gyllstad, and the Information Technology unit at the Centre for Languages and Literature in Lund are all acknowledged for valuable help with administration and logistics. Finally, our thanks to SensoMotoric Instruments for lending us two additional high-speed eyetrackers. Parts of the results in this article have been presented at conferences and workshops: the Scandinavian Workshop on Applied Eye Tracking (2010), the Vision Sciences Society (2011) meeting, and the European Conference on Eye Movements (2011). On page 130 of Holmqvist et al. (2011), we mentioned that participant-controlled calibration may give better accuracies and cited a poster presented at the 2011 annual meeting of the Vision Sciences Society.

## References

Baayen, R. H. (2010). languageR: Data sets and functions with analyzing linguistic data. A practical introduction to statistics (R package version 1.2) [Computer software]. Available from http://CRAN.R-project.org/package=languageR

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. doi:10.1016/j.jml.2007.12.005

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using s4 classes (R package version 0.999375-42) [Computer software]. Available from http://CRAN.R-project.org/package=lme4

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433–436. doi:10.1163/156856897X00357

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). London: Routledge Academic.

Crossland, M. D., Culham, L. E., & Rubin, G. S. (2004). Fixation stability and reading speed in patients with newly developed macular disease. *Ophthalmic & Physiological Optics, 24,* 327–333. doi:10.1111/j.1475-1313.2004.00213.x

Cui, Y., & Hondzinski, J. (2006). Gaze tracking accuracy in humans: Two eyes are better than one. *Neuroscience Letters, 396,* 257–262.

Di Russo, F., Pitzalis, S., & Spinelli, D. (2003). Fixation stability and saccadic latency in élite shooters. *Vision Research, 43,* 1837–1845. doi:10.1016/S0042-6989(03)00299-2

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

Gagl, B., Hawelka, S., & Hutzler, F. (2011). Systematic influence of gaze position on pupil size measurement: Analysis and correction. *Behavior Research Methods, 43,* 1171–1181. doi:10.3758/s13428-011-0109-5

Goldberg, H. J., & Wichansky, A. M. (2003). Eye tracking in usability evaluation: A practitioner's guide. In J. Hyöna, R. Radach, & H. Deubel (Eds.), *The mind's eye: On cognitive and applied aspects of eye movement research* (pp. 493–516). Amsterdam: Elsevier.

Hammoud, R. I. (2008). *Passive eye monitoring: Algorithms, applications and experiments*. New York: Springer.

Hansen, D. W., & Ji, Q. (2009). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32,* 478–500.

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

Hornof, A. J., & Halverson, T. (2002). Cleaning up systematic error in eye-tracking data by using required fixation locations. *Behavior Research Methods, Instruments, & Computers, 34,* 592–604. doi:10.3758/BF03195487

Jarodzka, H., Balslev, T., Holmqvist, K., Nyström, M., Scheiter, K., Gerjets, P., & Eika, B. (2010). Learning perceptual aspects of diagnosis in medicine via eye movement modeling examples on patient video cases. In S. Ohlsson & R. Catrambone (Eds.), *Cognition in flux: Proceedings of the 32nd Annual Meeting of the Cognitive Science Society* (pp. 1703–1708). Austin: Cognitive Science Society.

Kammerer, Y. (2009, May). *How to overcome the inaccuracy of fixation data—The development and evaluation of an offset correction algorithm*. Paper presented at the Scandinavian Workshop on Applied Eye-Tracking (SWAET), Stavanger, Norway.

Komogortsev, O., & Khan, J. (2008). Eye movement prediction by Kalman filter with integrated linear horizontal oculomotor plant mechanical model. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (pp. 229–236). New York, NY: ACM Press.

Kosnik, W., Fikre, J., & Sekuler, R. (1986). Visual fixation stability in older adults. *Investigative Ophthalmology & Visual Science, 27,* 1720.

Kumar, M., Klingner, J., Puranik, R., Winograd, T., & Paepcke, A. (2008). Improving the accuracy of gaze input for interaction. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (pp. 65–68). New York, NY: ACM Press.

Liversedge, S. P., White, S. J., Findlay, J. M., & Rayner, K. (2006). Binocular coordination of eye movements during reading. *Vision Research, 46,* 2363–2374. doi:10.1016/j.visres.2006.01.013

Marmitt, G., & Duchowski, A. (2002, September). *Modeling visual attention in VR: Measuring the accuracy of predicted scanpaths*. Paper presented at Eurographics 2002, Saarbrücken, Germany.

Miles, W. R. (1929). Ocular dominance demonstrated by unconscious sighting. *Journal of Experimental Psychology, 12,* 113–126.

Minshew, N., Luna, B., & Sweeney, J. (1999). Oculomotor evidence for neocortical systems but not cerebellar dysfunction in autism. *Neurology, 52,* 917.

Mullin, J., Anderson, A. H., Smallwood, L., Jackson, M., & Katsavras, E. (2001). Eye-tracking explorations in multimedia communications. In A. Blandford, J. Vanderdonckt, & P. Gray (Eds.), *Proceedings of IHM/HCI 2001: People and computers XV—Interaction without frontiers* (pp. 367–382). Cambridge: Cambridge University Press.

Nebes, R. D. (1978). Vocal versus manual response as a determinant of age difference in simple reaction time. *Journal of Gerontology, 33,* 884–889.

Nuthmann, A., & Kliegl, R. (2009). An examination of binocular reading fixations based on sentence corpus data. *Journal of Vision*, *9*(5), 31:1–28. doi:10.1167/9.5.31

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods, 42,* 188–204. doi:10.3758/BRM.42.1.188

Pernice, K., & Nielsen, J. (2009). *Eyetracking methodology—How to conduct and evaluate usability studies using eyetracking*. Berkeley: New Riders Press.

Putnam, N. M., Hofer, H. J., Doble, N., Chen, L., Carroll, J., & Williams, D. R. (2005). The locus of fixation and the foveal cone mosaic. *Journal of Vision*, *5*(7), 3:632–639. doi:10.1167/5.7.3

R Development Core Team. (2009). *R: A language and environment for statistical computing [Computer software manual]*. Vienna: R Foundation for Statistical Computing. Available from www.R-project.org.

Rascol, O., Sabatini, U., Simonetta-Moreau, M., Montastruc, J. L., Rascol, A., & Clanet, M. (1991). Square wave jerks in Parkinsonian syndromes. *British Medical Journal, 54,* 599–602.

Rayner, K., Pollatsek, A., Drieghe, D., Slattery, T. J., & Reichle, E. D. (2007). Tracking the mind during reading via eye movements: Comments on Kliegl, Nuthmann, and Engbert (2006). *Journal of Experimental Psychology. General, 136,* 520–529.

SR Research, Inc. (2007). *EyeLink user manual 1.3.0* [Computer software manual]. Mississauga, Ontario, Canada: Author.

Salvucci, D., & Goldberg, J. H. (2000). Identifying fixations and saccades in eyetracking protocols. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications* (pp. 71–78). New York, NY: ACM.

Schnipke, S., & Todd, M. (2000). Trials and tribulations of using an eye-tracking system. In *CHI'00: Extended abstracts on Human Factors in Computing Systems* (pp. 273–274). New York, NY: ACM Press.

SensoMotoric Instruments. (2009). *iView X system manual* (Version 2.4) [Computer software manual]. Berlin, Germany: Author.

SensoMotoric Instruments. (2010). *Experiment Center manual 3.0* [Computer software manual]. Teltow, Germany: Author.

Shic, F., Scassellati, B., & Chawarska, K. (2008). The incomplete fixation measure. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (pp. 111–114). New York, NY: ACM Press.

Smeets, J. B. J., & Hooge, I. T. (2003). Nature of variability in saccades. *Journal of Neurophysiology, 90,* 12–20. doi:10.1152/jn.01075.2002

Steinman, R., Cunitz, R., Timberlake, G., & Herman, M. (1967). Voluntary control of microsaccades during maintained monocular fixation. *Science, 155,* 1577.

Tarita-Nistor, L., González, E. G., Mandelcorn, M. S., Lillakas, L., & Steinbach, M. J. (2009). Fixation stability, fixation location, and visual acuity after successful macular hole surgery. *Investigative*

*Ophthalmology & Visual Science, 50,* 84–89. doi:10.1167/iovs.08-2342

Tobii Technology. (2010). *Tobii Studio user manual 2.X* [Computer software manual]. Stockholm, Sweden: Author.

Tobii Technology. (2011). *Tobii tx300 eye tracker user manual, release 1.0* [Computer software manual]. Danderyd, Sweden: Author.

van der Geest, J. N., & Frens, M. A. (2002). Recording eye movements with video-oculography and scleral search coils: A direct comparison of two methods. *Journal of Neuroscience Methods, 114,* 185–195.

Zhang, Y., & Hornof, A. J. (2011). Mode-of-disparities error correction of eye-tracking data. *Behavior Research Methods, 43,* 834–842. doi:10.3758/s13428-011-0073-0