Eye-tracking data quality as affected by ethnicity and experimental design

Article *in* Behavior Research Methods · April 2013

DOI: 10.3758/s13428-013-0343-0 · Source: PubMed

2 authors:

Pieter Blignaut
University of the Free State
79 PUBLICATIONS **630** CITATIONS

SEE PROFILE

Daniël Jacobus Wium
University of the Free State
7 PUBLICATIONS **109** CITATIONS

SEE PROFILE

Behav Res (2014) 46:67–80
DOI 10.3758/s13428-013-0343-0

BRIEF COMMUNICATION

Eye-tracking data quality as affected by ethnicity and experimental design

Pieter Blignaut · Daniël Wium

Published online: 23 April 2013

Abstract Lack of accuracy in eye-tracking data can be critical. If the point of gaze is not recorded accurately and reliably, the information obtained or action executed might be different from what the user intended. This study reports trackability, accuracy, and precision as indicators of eye-tracking data quality as measured at various head positions and light conditions for a sample of participants from three different ethnic groups. It was found that accuracy and precision for Asian participants was worse than that for African and Caucasian participants. No significant differences were found between the latter two ethnic groups. Operating distance had the largest effect on data quality, since it affected all indicators for all ethnic groups. Illumination had no significant effect on accuracy or precision, but the accuracy achieved by African and Caucasian participants was better when the stimulus was presented on a dark background. Large gaze angles proved to be detrimental for trackability for African participants, while accuracy and precision were also affected adversely by larger gaze angles for two of the ethnicities.

Keywords Eye tracking · Data quality · Accuracy · Precision · Ethnicity

Introduction

Eye tracking can be used to obtain information about how people acquire and process information while they read (Rayner, Pollatsek, Drieghe, Slattery & Reichle, 2007), browse a Web site (Goldberg, Stimson, Lewenstein, Scott & Wichansky, 2002), shop (Vikström, Wallin & Holmqvist, 2009), drive a motor car (Crundall, Chapman, Phelps & Underwood, 2003), interpret medical images (Donovan, Manning & Crawford, 2008), and perform other tasks where the ability to decipher the visual world around them is critical. Eye tracking can also be used as input modality during computer interaction, such as eye typing (Abe, Ohi & Ohyama, 2007) or other gaze-contingent systems. No matter what the application area, it is important that the quality of data is sufficient to support the investigation or action at hand.

Data quality may be affected by one or more of three primary sources of error—namely, *participant characteristics*, *equipment features*, or the *experimental setup*. First, the output from eye-tracking devices may vary with individual differences in the shape or size of the eyes, such as the corneal bulge and the relationship between the eye features (pupil and corneal reflections) and the foveal region on the retina. Ethnicity, viewing angle, head pose, eye color, cleft and texture, position of the iris within the eye socket, and the state of the eye (open or closed) all influence the appearance of the eye (Hansen & Ji, 2010) and, therefore, the quality of eye-tracking data (Holmqvist, Nyström & Mulvey, 2012). Equipment-related problems might occur because of unstable or unsuitable sampling frequency, low camera resolution, incorrect or unstable identification of the pupil and glint centers, the fixation identification algorithm and associated parameter settings, the calibration procedure, and other hardware- and software-related issues. Experimental conditions, such as light conditions, head position, and stimulus can also largely affect experimental outcomes, along with incorrect analysis of data. It is also not always obvious what experimental setup would be the best to obtain optimum eye-tracking results. Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka and Van de Weijer (2011) lists a wide variety of possible sources of error, with illustrative values.

Several separate constructs may be used to quantify data quality (or the lack thereof) (Holmqvist et al., 2011). *Spatial accuracy* refers to the distance between the actual and reported gaze positions, while *temporal accuracy* (a.k.a. *latency*) refers to the time difference between the actual and reported gaze events. *Spatial precision* (sometimes referred to as *noise*) and

P. Blignaut (✉) · D. Wium
Department of Computer Science and Informatics (IB65),
University of the Free State, PO Box 339,
9300, Bloemfontein, South Africa
e-mail: pieterb@ufs.ac.za

temporal precision refer to the variance in position and latency measures, respectively. *Trackability* (or *robustness*) refers to the proportion of raw data samples that are lost during a recording; the more data are lost, the worse the trackability is. *Spatial resolution* refers to the smallest eye movement that can be detected reliably in the data. Spatial resolution is highly dependent on spatial precision, since it is impossible to detect eye movements that are smaller than the sample-to-sample variability in the data.

This article focuses on trackability, spatial accuracy (or just accuracy), and spatial precision (or just precision) as indicators of the variance in data quality between participants from three different ethnic groups—namely, Africans, Asians, and Caucasians—under varying but controlled experimental conditions.

Measures of data quality

ISO standard 5725–1 defines precision as the "closeness of agreement between independent test results obtained under stipulated conditions" (ISO 5725–1, 1994). Precision should not be confused with accuracy, which is defined as the "closeness of agreement between a test result and the accepted reference value" (ISO 5725–1, 1994). Informally, precision refers to the spread (or dispersion) of the recorded raw gaze data samples, while accuracy refers to the offset between the observed and actual fixation positions (Hornof & Halverson, 2002). Figure 1 (left) shows imprecision around the stimulus (crosshair), while in Fig. 1 (right), the precision is better but accuracy is poor.

Both accuracy and precision can be computed separately for the two eyes and separately in the horizontal and vertical directions. Trackability, accuracy, and precision values are calculated from samples that are recorded when a participant (or pair of artificial eyes) is assumed to fixate a stationary target.

Trackability

Trackability (or robustness or versatility) refers to how well an eyetracker works for a variety of participants (Holmqvist et al., 2011). Trackability can be quantified by dividing the number of valid recorded raw data samples by the total number of samples that were supposed to be captured in the tracking period.

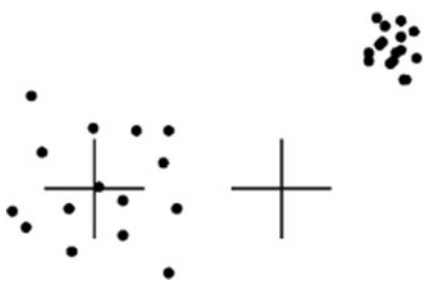

Fig. 1 Lack of precision (*left*) and lack of accuracy (*right*). (Examples from Hornof & Halverson, 2002)

Loss of data typically occurs when some of the critical features of the eye image—for example, the pupil and/or corneal reflection—cannot be reliably detected (Holmqvist et al., 2011). Sometimes, it might be possible to do a regression to fill the gaps, but long periods of data loss cannot be addressed in this way. Holmqvist shows how data loss can reduce the number of processed fixations, while increasing the fixation duration.

Typically, glasses and contact lenses may cause reflections that can either obscure the corneal reflection or be regarded by the eyetracker as being actual corneal reflections. We have found that tilting the glasses somewhat might have a huge impact on data quality (Fig. 2).

Participant-related eye physiology—for example, droopy eyelids or narrow eyes—may also obscure the glint or pupil or part thereof with subsequent loss of data (Fig. 3). If the eyetracker allows, such problems can be addressed by adjusting the eye camera in order to manipulate the gaze angle (angle at the eye between eye camera and gaze target).

Other aspects that could make a difference to trackability are adjustment of resolution, number of eye cameras, quality of camera sensors, and number of and positioning of infrared illuminators, as well as parameters to the image-processing algorithms. Unfortunately, researchers very seldom have access to these variables, since manufacturers mostly hide these from the user in favor of ease of use.

Accuracy

Explanation and significance

Lack of accuracy, also known as systematic error, may not be a problem when the areas of interest are large (e.g., 8º of visual angle) and are separated by large distances (Zhang & Hornof, 2011), but in studies where the stimuli are closely spaced, as in reading, or irregularly distributed as on a Web page, uncertainty of as little as 0.5°–1° can be critical in the correct analysis of eye-tracking data. Accuracy is also of great importance for gaze-input systems (Abe et al., 2007). With regard to reading research, for example, Rayner et al. (2007) states 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" (p. 522). Lack of accuracy may also result from bad calibrations, head movements, astigmatism, eyelid closure, or other sources that are strongly dependent on the particular characteristics of the individual participant (Hornof & Halverson, 2002).

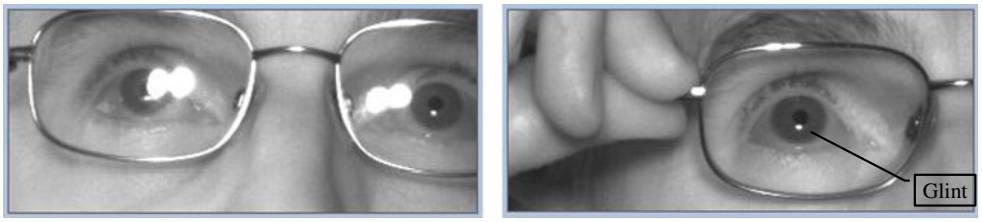

Fig. 2 Glasses can cause reflections which make eye tracking impossible. By using the eye video, the glasses can be repositioned so that reflections are minimized or even eliminated

In principle, the accuracy of eye tracking refers to the difference between the measured gaze direction and the actual gaze direction for a person positioned at the center of the head box (Borah, 1998; Tobii, 2010). It is measured as the distance, in degrees, between the position of a known target point and the average position of a set of raw data samples, collected from a participant looking at the point (Hornof & Halverson, 2002; Holmqvist et al., 2011). This error may be averaged over a set of target points that are distributed across the display.

In order to compare and replicate research results, it is essential that researchers report the accuracy of eye tracking. Stating the manufacturers' specifications could be misleading, since it is known that accuracy can vary considerably across participants and experimental conditions (Blignaut & Beelders, 2009; Hornof & Halverson, 2002). Some researchers—for instance, Tatler (2007) and Foulsham and Underwood (2008)—make a point of recalibrating until the measured accuracy is below 0.5° and only then start recording.

Reported accuracy of video-based eyetrackers

Video-based eye tracking is the most widely practiced eye-tracking technique (Hua, Krishnaswamy & Rolland, 2006). This technique is based on the principle that when near infrared light is shone onto the eyes, it is reflected off the different structures in the eye to create four Purkinje reflections (Crane & Steele, 1985). The vector difference between the pupil center and the first Purkinje image (also known as the glint or corneal reflection is tracked. Corneal reflection/pupil devices are largely unobtrusive and easy to operate.

While an accuracy of 0.3° has been reported for tower-mounted high-end systems operated by skilled operators (Holmqvist et al., 2011; Jarodzka et al., 2010), remote systems are usually 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 mean accuracy over participants was 1°. They regarded systematic errors of less than 1.7° as being acceptable. Johnson, Liu, Thomas and Spencer (2007), using an alternative calibration procedure for another commercially available eyetracker with stated accuracy of 0.5°, found an azimuth error of 0.93° and an elevation error of 1.65°. Zhang and Hornof (2011) also found an accuracy of 1.1°, which is well above the 0.5° as stated by the manufacturer. Using a simulator eyeball, Imai et al. (2005) found that a video-based eyetracker has an *x*-error of 0.52° and a *y*-error of 0.62°. Van der Geest and Frens (2002) reported an *x*-error of 0.63° and a *y*-error of 0.72°. Chen, Tong, Gray and Ji (2008) proposed a "robust" video-based eye-tracking system that has an accuracy of 0.77° in the horizontal direction and 0.95° in the vertical direction, which they deemed acceptable for many HCI applications that allow natural head movements. Brolly and Mulligan (2004) also proposed a video-based eyetracker with accuracy in the order of 0.8°. Hansen and Ji (2010) provided an overview of remote eyetrackers and reported the accuracy of most model-based gaze estimation systems to be in the order of 1°–2°.

Fig. 3 Participant with droopy eyelid that partially obscures the pupil

Hornof and Halverson (2002) thoroughly studied the nature of systematic errors in a set of eye-tracking data collected from a visual search experiment. They found that the systematic error tends to be constant within a region of the display for each participant. Specifically, the magnitude of the disparities between the target visual stimuli and the corresponding fixations were "somewhat evenly distributed around 40 pixels (about 1º of visual angle) and most were between 15 and 65 pixels" (Hornof & Halverson, 2002, p. 599). Horizontal and vertical disparities remained constant to a certain degree for each participant. Thus, accuracy was not randomly distributed across all directions or sizes but was, as the name implies, systematic.

The role of calibration

Calibration refers to a procedure to gather data so that the coordinates of the pupil and one or more corneal reflections in the coordinate system of the eye video can be converted to *x*- and *y*-coordinates that represent the participant's point of regard in the stimulus space. The procedure usually consists of asking the participant to look at a number of predefined points at known angular positions while storing

samples of the measured quantity (Abe et al., 2007; Kliegl & Olson, 1981; Tobii, 2010). Theoretically, this transformation should remove any systematic error, but eyetrackers often maintain some systematic error even directly after careful calibration (Hornof & Halverson, 2002).

Precision

The precision of an eyetracker is defined as the ability to reliably reproduce a measurement (Holmqvist et al., 2011). Precision should be calculated from data samples that are recorded when the eye is fixating on a stationary target (Holmqvist et al., 2011). The use of artificial eyes presents a good way to ensure that there is no movement of the eyes and that the precision measured is a property of variation in the eyetracker only.

High precision is needed when measuring small fixational eye movements such as tremor, drift, and microsaccades. Poor precision can also be detrimental to fixation and saccade detection algorithms. Holmqvist et al. (2012) found that the number of fixations that are identified by an adaptive velocity threshold algorithm increases with precision, while the duration of fixations decreases. They surmise that higher noise levels prevent the algorithm from identifying shorter saccades, with the result that nearby but separate fixations are merged into longer fixations.

Poor precision can be caused by a multitude of technical and participant-specific factors like hardware limitations and eye color (Holmqvist et al., 2011). Our own analysis of eye videos and comparisons between human and artificial eyes have led us to conclude that poor precision is usually caused by unstable pupil and glint detection by the imaging software and is not necessarily participant related. This is especially the case when the images taken by the eye camera is of lower quality.

Participant-related causes of imprecision can be addressed by using a chinrest or bite-board, while hardware-related causes can be addressed by using different eye-tracking hardware—for example, by using an eyetracker with a high-quality eye camera and/or using a tower- or head-mounted unit instead of a remote unit. If none of these are possible, the variability can be hidden through averaging the gaze points by means of a fixation identification algorithm (Hornof & Halverson, 2002).

The standard deviation (σ) of a set of data samples is a measure of the spread around the mean or central value. It is defined as $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_1^2}$ where N is the number of samples and d_i is some distance measure between an individual sample, i, and the central value. In two dimensions, the standard deviation would be $\sqrt{\left(\sigma_x^2 + \sigma_y^2\right)/2}$ where $\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$ and $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$.

Methodology

The trackability, accuracy, and precision of eye-tracking data under various conditions were determined by asking participants of three ethnic groups to follow some dots on the screen.

Hardware and software

A Tobii TX300 eyetracker with a frame rate of 300 Hz, screen resolution of 1,920×1,080 and a pixel size of 0.26 mm was used. According to the product description (Tobii, 2010), the operating distance is specified as 50–80 cm, and the freedom of head movement at 65 cm is 37 cm horizontally and 17 cm vertically. The firmware version for the TX300 at the time of testing was 3.0.1. The Metrics software (version 2.1.7.65534) developed by Tobii was used to capture the raw data, but it was not used to analyze the data.

It is important to note that this study is about the differences in data quality between three different ethnic groups, and not about the specific instrument that was used to capture the data. Another eyetracker, of the same model, might produce different results, and the absolute values for data quality should not be used to evaluate or compare the quality of this model of eyetracker with that of another.

Participants

Participants were recruited so as to provide a balance over ethnic group as far as possible. Eventually, 71 participants were tested, of which 26 were African, 22 Asian, and 23 Caucasian.

In order to focus on the effects of ethnicity and avoid possible influence of sight correction on data quality, participants who could not see clearly on the screen without glasses were not recruited. Although no formal screening was done with regard to other demographic or participant-related factors, some figures are provided in Table 1. It might seem that eye color is quite unbalanced, but it must be kept in mind that all Africans and most Asians have brown eyes.

Laboratory setup

The eyetracker was placed on a special table that is adjustable in three dimensions (Fig. 4) to control the head position relative to the eyetracker. A chinrest was used to ensure that the head position stays constant for a specific configuration. The center position for the eyetracker was defined as the position where the cameras are 650 mm from the participant's eyes and with eyes in the horizontal and vertical center of the head-box. The screen was perpendicular to the table at all times.

Table 1 Number of participants per demographic factor

Factor	Value	Participants
Gender	Male	47
	Female	24
Age group	2–10	1
	11–20	12
	21–30	48
	31–60	10
Dominant eye	Left	24
	Right	45
	Unknown	2
Eye color	Blue	12
	Brown	53
	Green	6
Makeup	Yes	11
	No	60

Five adjustable lights were used to adjust the lighting conditions inside the laboratory. One light, attached to the ceiling, was directly above the eyetracker, while the other four lights were placed in the corners of the room, just below the ceiling. The room was darkened so that illumination of approximately 0 lux could be achieved when all the lights were switched off. Two light meters, one facing the participant and one facing the eye camera, were used to measure lux levels. The light conditions were adjusted until similar readings were recorded on both light meters.

Experimental procedure

A participant session consisted of a calibration routine followed by a series of tests where a set of dots were displayed, and participants had to fixate each one of the dots. A participant session lasted about 30–45 min, on average.

Calibration was done once per participant at the start of a session by expecting participants to fixate nine white dots on a black background at random positions. Calibration was done in the center of the head box at 300 lux. A calibration was regarded as valid if data could be captured for each dot. If necessary, the calibration routine was repeated for one or more dots. A session was terminated if the eyetracker was unable to track a participant's eyes.

Fig. 4 Test participant and facilitator in front of a Tobii TX300. Note the chinrest, adjustable table, and light meters

The trackability, accuracy, and precision of the eye-tracking data were determined for each of the following conditions:

- Head position with respect to horizontal and vertical position relative to the eyetracker, as well as the operating distance (distance from the eye camera to the eyes)
- Light conditions
- Gaze angle
- Stimulus background

A participant session consisted of 21 tests, each with the eyetracker in a different position or with different light conditions, dot positions, or background color. Table 2 summarizes the various combinations. From the center position, the eyetracker was moved along the *x*-, *y*- and *z*-axes in 5-cm intervals, in both the positive and negative directions, and a test was done at each position (see Fig. 5 for the definition of the axes). Note that the *z*-axis is not perpendicular to the *x*–*y* plane but, rather, gives an indication of the operating distance to the eye camera.

Note also that −10 cm means that the table was lowered (or moved to the left) 10 cm and that the head is thus 10 cm above (or to the right of) the center position. Similarly, + 10 cm means that the table was raised (or moved 10 cm to the right), with the result that the head was 10 cm below (or to the left of) the center position.

Stimuli

Unless stated otherwise (Table 2), the stimuli for a single test consisted of nine dots, each within a larger disk (diameter 36 pixels) (Fig. 6). The dots were arranged in a 3×3 grid (Fig. 7) but appeared one at a time in random order for 2 s each, with 1 s in between. For all tests but that for set F, the dots were displayed on a black background, while for set F, the dots were displayed in white within a black disk on a white background. For all sets but set G, the dots were displayed at positions such that the two upper corner dots appeared at 20° (see angle α in Fig. 8). For set G, two dots at 25° and two dots at 30° were displayed in each corner (Fig. 9). The test in darkness was always performed last, and 3 min was allowed for the participant's pupil size to adjust before the test commenced.

Analysis

In order to maximize the probability that a participant's eyes were fixating on a specific dot, only data collected between

Table 2 Experimental configurations

Set	*X* (Horizontal)	*Y* (Vertical)	*Z* (Operating distance)	Light conditions	Background	Dots[1]
A	Center	Center	65 cm	300 lux	Black	9 at≤20°
B	Center −10 cm Center −5 cm	Center	65 cm	300 lux	Black	9 at ≤20°
	Center +5 cm					
	Center +10 cm					
C	Center	Center −10 cm Center −5 cm	65 cm	300 lux	Black	9 at ≤20°
		Center +5 cm				
		Center +10 cm				
D	Center	Center	50 cm 55 cm	300 lux	Black	9 at ≤20°
			60 cm			
			70 cm			
			75 cm			
			80 cm			
E	Center	Center	65 cm	0 lux 600 lux	Black	9 at ≤20°
				1,000 lux		
F	Center	Center	65 cm	300	White	9 at ≤20°
G	Center	Center	65 cm	300 lux	Black	2×2 at 25°
						2×2 at 30°

[1] See the Stimuli section for details)

1,000 and 1,500 ms since the dot's appearance were used during analysis. The raw data for each participant were exported to a database, and no filtering, event detection, or gap filling of any kind was done.

SQL queries were used to aggregate the results over the various combinations, and a statistical package was used to conduct an analysis of variance (ANOVA) test for each one of the factors, while controlling for the others. In other words, the Metrics software was used only to present the stimuli and capture the raw data, while the analysis was performed independently. That is, the data quality results as provided by Metrics were not used.

The accuracy and precision were averaged over all dots for a specific participant–test combination. All results were compared with the results for set A, which was regarded as the benchmark configuration.

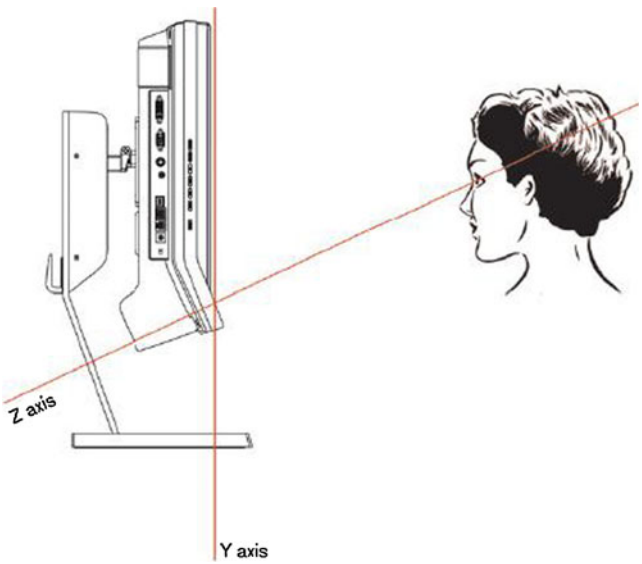

Fig. 5 Definition of axes for different head positions

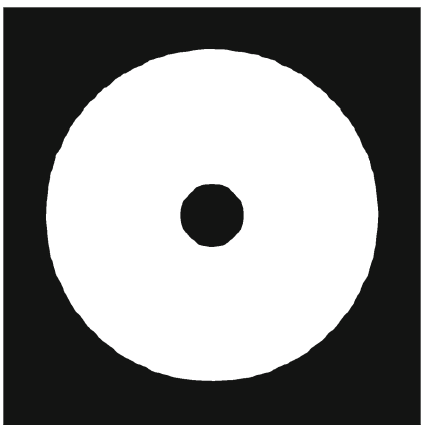

Fig. 6 Stimuli dots within a larger disk

Fig. 7 Positioning of the dots with gaze angles ≤20°.

Fig. 9 Positioning of dots with gaze angles at 25° and 30°

Results

Note the following conventions that were used in the subsequent analysis:

- Unless otherwise stated, significant levels were taken as α=0.01.
- All independent variables were categorical, but when it was possible to interpret the values on a continuous scale—for example, operating distance—line graphs were deemed to be more informative than bar charts. Bar charts were used where no sensible continuous scale was possible—for example, ethnicity or background color.
- Since all independent variables were categorical, even though they were displayed as line graphs, no meaning should be attached to slight offsets on the *x*-axis. Values were separated somewhat in order to avoid clutter with the display of confidence intervals.

Trackability

The trackability for each participant−test combination was calculated by expressing the number of raw data samples that were captured as a percentage of the maximum number of samples that could be captured in the interval between 1,000 and 1,500 ms after a point was displayed. It is acknowledged that blinks (about 100 ms) will cause loss of data within the 500-ms period of analysis, but it is assumed that the frequency and probability of blinks are the same for all test conditions and would, therefore, not affect the comparative value of the findings.

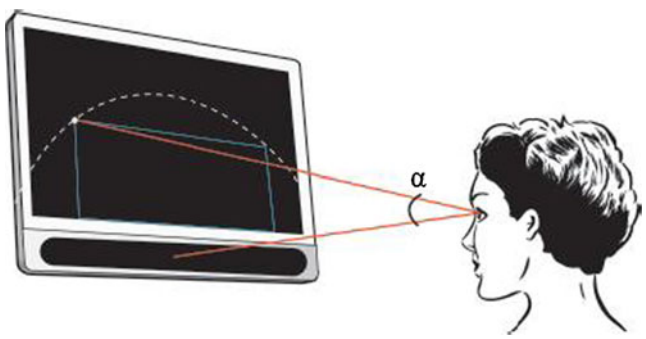

Fig. 8 Positioning of nine dots within 20°

A one-way ANOVA for the effect of ethnicity on trackability under ideal conditions (set A in Table 2) showed that the trackability was worse for Asian participants (87.2 %) than for Africans (95.5 %) and Caucasians (99.1 %). While the overall effect was not significant at the α=0.05 level, $F(2, 68)=1.99$, p=0.145, Fisher's post hoc test for the least significant difference (Fisher's LSD) indicated that the individual differences between Asians and Caucasians were significant at the α=0.1 level.

The data for all target points of a specific test were combined, and a one-way ANOVA over participants was done for the effect of head position (three dimensions), illumination, background color, and gaze angle on trackability, while controlling for ethnicity. The results are summarized in Table 3 and presented graphically in Fig. 10.

Referring to Table 3, horizontal head position (experimental conditions A and B) does not have a significant effect on trackability for any of the ethnic groups, although there seems to be an indication that participants performed somewhat worse at the far-left position (table moved 10 cm to the right). Vertical head position (conditions A and C) has a significant influence on trackability. Using Fisher's post hoc test again, it was found specifically that the trackability at the −10 and +10 positions (head position 10 cm above and 10 cm below the center, respectively) was significantly worse than the trackability in the center of the head box. This is clearly illustrated in Fig. 10.

In hindsight, it was realized that it would have been better to rather test trackability at −8, −4, 0, 4, and 8 cm, since the −10 and +10 positions were outside the 17-cm permissible vertical tolerance as specified for the specific model of eyetracker (Tobii, 2010). If the ANOVA in Table 3 was restricted to −5, 0, and +5 positions, vertical head position had no significant difference on trackability for any of the ethnic groups (Table 3).

The effect of operating distance (experimental conditions A and D) was significant for all ethnic groups (Table 3). Fisher's post hoc test showed specifically that the trackability at 50 and

Table 3 Analysis of variance for the effect of various factors on trackability (significant *p* values [α=0.01] are boldfaced)

Factor	Range	African			Asian			Caucasian		
		df	*F*	*p*	*df*	*F*	*p*	*df*	*F*	*p*
Horizontal head position	−10 to +10 cm	4,125	2.24	0.068	4,105	0.284	0.888	4,110	1.13	0.348
Vertical head position	−10 to +10 cm	4,125	49.25	**0.000**	4,105	27.26	**0.000**	4,110	81.60	**0.000**
	−5 to +5 cm	2,75	0.035	0.966	2,63	0.017	0.983	2,66	0.153	0.859
Operating distance	50 – 80 cm	6,175	39.44	**0.000**	6,147	14.54	**0.000**	6,154	29.40	**0.000**
	60 – 80 cm	4,125	0.122	0.974	4,105	0.941	0.443	4,110	4.64	**0.002**
Illumination	0, 300, 600, 1,000 lux	3,100	1.92	0.132	3,84	0.039	0.990	3,88	4.69	**0.004**
Background color	Black, White	1,50	0.014	0.906	1,42	0.097	0.757	1,44	0.001	0.972
Gaze angle	≤20°, 25°, 30°	2,75	3.99	0.023	2,63	0.214	0.808	2,66	0.530	0.591

55 cm was significantly worse than the trackability at the center of the head box. While the specified operating range for the eyetracker in this study was 50–80 cm (Tobii, 2010), it is clear that it works much better at 60–80 cm. When the ANOVA was repeated for this range, the effect of operating distance was significant for Caucasians only.

All ethnic groups performed worse in the dark than in any of the other light conditions (conditions A and E), although only the difference for Caucasians proved to be significant (Table 3). Background color (conditions A and F) had no significant effect on trackability for any one of the ethnicities. Trackability for African participants was significantly worse (α=0.05) at 30° than at smaller angles. No significant differences with respect to gaze angle (conditions A and G) were found for the other two ethnicities.

Accuracy

The binocular accuracy (average of left and right eyes) for every participant–test combination was expressed as the

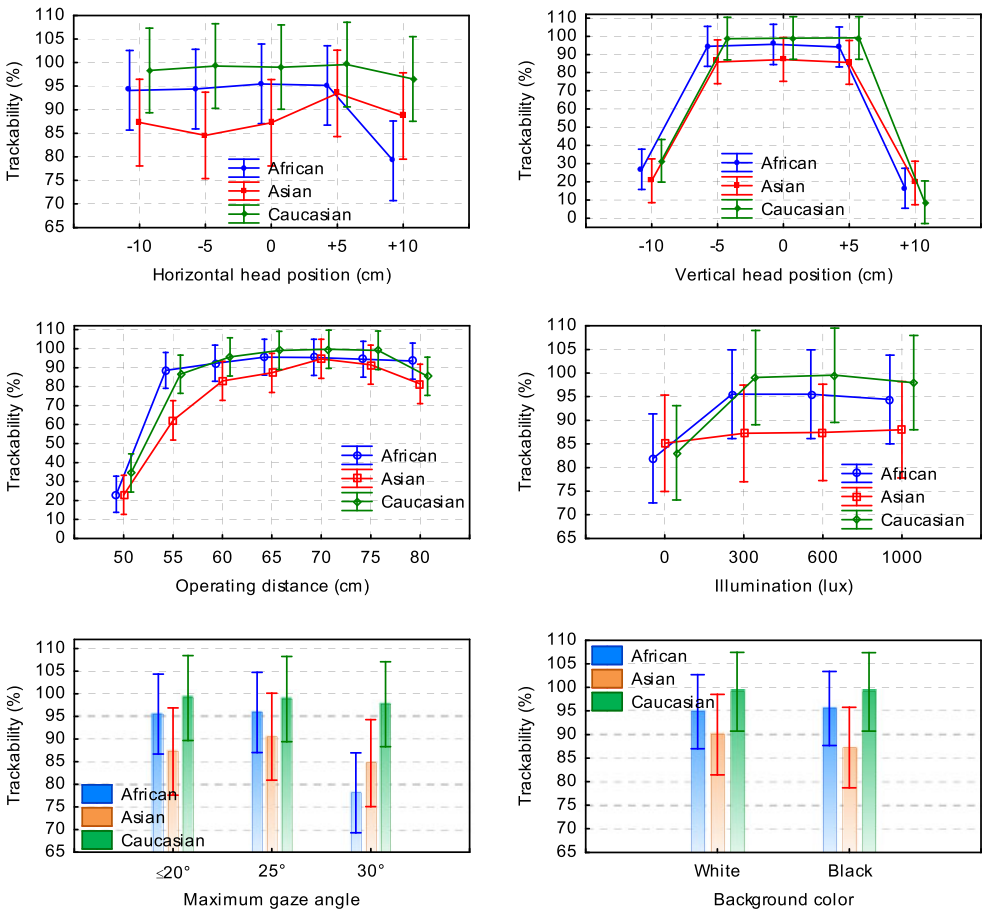

Fig. 10 Trackability against various factors per ethnicity

average offset (Euclidean distance between the sample and the target) of samples in the interval between 1,000 and 1,500 ms after a target point was displayed.

A one-way ANOVA for the effect of ethnicity on accuracy under ideal conditions (set A in Table 2) showed that the accuracy was worse for Asian participants (0.91°) than for Africans (0.57°) and Caucasians (0.61°). While the overall effect was not significant at the α=0.05 level, $F(2, 65)=3.08$, $p=0.052$, Fisher's LSD post hoc test confirmed that the individual differences between Asians and Africans and Asians and Caucasians were indeed significant at this level (0.05).

The data for all target points of a specific test were combined, and a one-way ANOVA over participants was done for the effect of head position (three dimensions), illumination, background color, and gaze angle on accuracy while controlling for ethnicity (Table 4). Since not all participants could be tracked in all experimental conditions or target positions, the number of observations (and thus the degrees of freedom) is not always the same. The results are presented graphically in Fig. 11.

Although accuracy seemed to be a bit better in the center of the head box than at the edges, it could not be proven that either horizontal or vertical head position is a significant indicator of accuracy for any of the ethnicities (Table 4). The relative large confidence intervals at −10 and +10 vertical positions are probably largely due to the bad trackability at these positions (see the Trackability section above), since only a small number of observations was available.

With respect to operating distance, the accuracy was significantly better for all ethnic groups in the center of the head box (55 and 60 cm) than at the near and far ends (50 and 80 cm). The bad accuracy at 50 cm should be interpreted in the light of the ideal head box for the specific eyetracker (see the Trackability section above).

Illumination did not influence accuracy significantly, although it seems as though Asian participants performed somewhat worse in dark conditions, while, Caucasian participants performed somewhat worse in bright conditions. African and Caucasian participants performed significantly better (α=0.05) on a black background than on a white background. While it seemed to be the other way around for Asian participants, the effect was not significant.

African and Caucasian participants showed significantly worse accuracy at 30° than at smaller ones (α=0.01). Although a similar trend was observed for Asian participants, it was not significant.

Precision

Precision was calculated for each target point of each participant–test combination as the standard deviation based on the pooled variance in the x- and y-dimensions. A one-way ANOVA for the effect of ethnicity on precision under ideal conditions showed that precision was not significantly influenced by ethnicity, $F(2, 65)=1.75$, $p=0.183$. Although the precision was somewhat worse for Asians (0.15°) than for the other two ethnic groups (both 0.11°), Fisher's LSD test showed that the effect was significant only at the α=0.1 level.

The data for all target points of a specific test were combined, and a one-way ANOVA over participants was done for the effect of head position (three dimensions), illumination, background color, and gaze angle on precision, while controlling for ethnicity. The results are summarized in Table 5 and presented graphically in Fig. 12.

Although it seems as though there is a trend of precision being somewhat worse at the left and right edges of the head box, this was significant for African and Caucasian participants only. Precision was significantly better for the Asian and Caucasian participants in the center of the screen than at the top (−10 cm) and bottom (+10 cm) of the screen.

Operating distance had a significant effect on precision (α=0.01), since precision was significantly better for all ethnic groups in the center of the head box (55 and 60 cm) than at the near and far ends (50 and 75 cm+).

Neither illumination nor background color had a significant effect on precision for any of the ethnic groups. African and Caucasian participants performed significantly worse at large gaze angles than at smaller gaze angles.

Discussion

Effect of various factors on data quality

Table 6 shows a summary of the significance of ethnicity, head movement, illumination, and gaze angle on trackability, accuracy, and precision of eye tracking as it was determined in this study. Each one of the factors is discussed below in more details.

Ethnicity

The narrow eyes of Asian participants cause the eyetracker to lose the glint, and therefore, trackability is worse for Asians than for the other two ethnicities. Asian participants also performed worse than the other two ethnicities with regard to accuracy and precision. It is acknowledged that not all Asians have narrow eyes and some African and Caucasian people may have eyes with a narrow cleft, but it is assumed that the participants in this study were representative of the majority of people from the respective ethnicities.

Table 4 Analysis of variance for the effect of various factors on accuracy

Factor	Range	African			Asian			Caucasian		
		df	*F*	*p*	*df*	*F*	*p*	*df*	*F*	*p*
Horizontal head position	−10–10 cm	4,117	1.20	0.313	4,96	0.04	0.997	4,110	1.93	0.110
Vertical head position	−10–10 cm	4,83	1.00	0.413	4,69	0.33	0.860	4,76	2.24	0.072
Operating distance	50–80 cm	6,159	18.40	<**0.000**	6,128	6.72	<**0.000**	6,146	15.33	<**0.000**
	60–80 cm	4,120	17.78	<**0.000**	4,100	4.04	**0.004**	4,110	14.95	<**0.000**
Illumination	0, 300, 600, 1,000 lux	3,95	1.98	0.122	3,76	0.48	0.707	3,87	1.92	0.133
Background color	Black, White	1,48	11.83	**0.001**	1,38	0.23	0.631	1,44	4.45	0.041
Gaze angle	≤20°, 25°, 30°	2,71	6.39	**0.003**	2,57	1.96	0.150	2,66	5.89	**0.004**

Head position

Accuracy was not affected by horizontal movement of the head, while trackability was affected somewhat for African participants only. Precision was affected significantly for Africans and Caucasians, but one should take into account that because of the small absolute precision values and the even smaller variances, a difference as small as 0.1° would prove significant.

Since head positions were used that are outside the head box as specified by the manufacturer, vertical head movement proved to be a significant indicator of trackability in this study. Researchers will have to take cognizance of the small vertical tolerance and take care to position participants such that their eyes are on the same level as the vertical center of the eyetracker. For those participants who could be tracked, vertical head movement did not affect accuracy, while precision was affected. Asians showed a particularly interesting trend (see Fig. 12b) that could be related to narrow eyes with the glint being clearly defined in the straight-forward gaze position only.

Operating distance clearly had the largest effect on data quality, since all indicators (trackability, accuracy, and

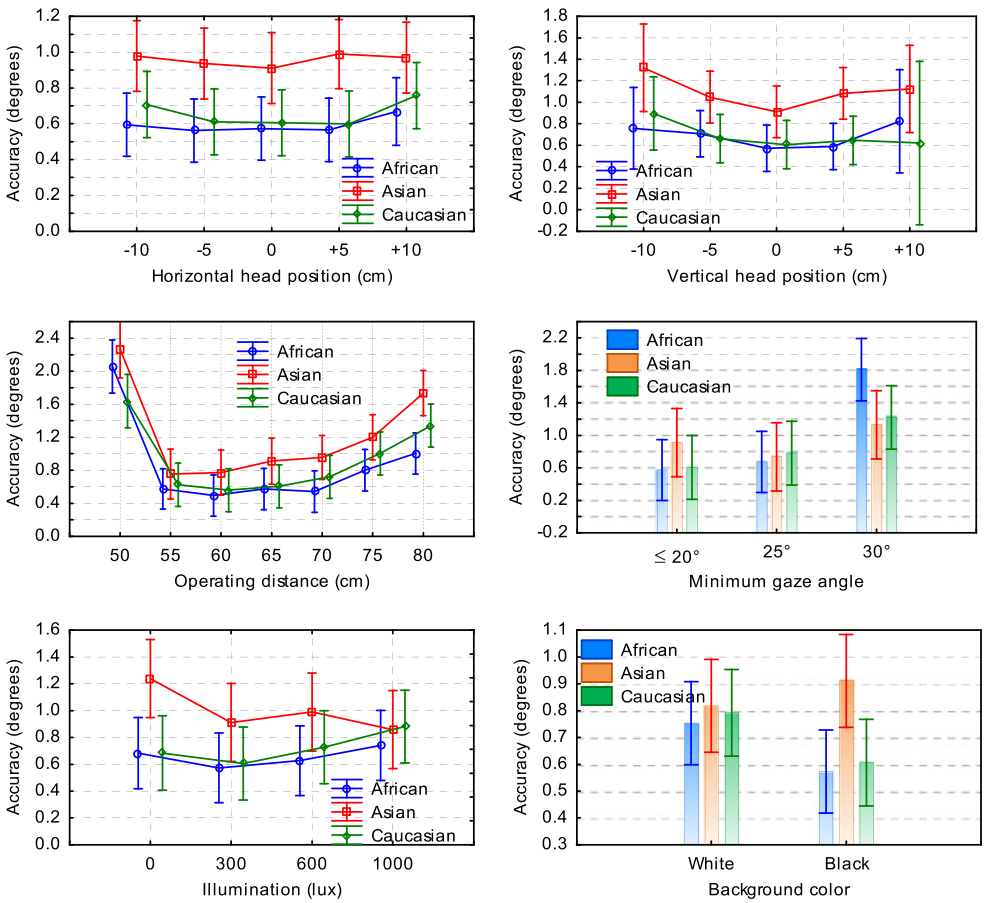

Fig. 11 Accuracy against various factors per ethnicity

Table 5 Analysis of variance for the effect of various factors on precision

Factor	Range	African			Asian			Caucasian		
		dF	F	p	dF	F	p	dF	F	p
Horizontal head position	−10 – 10 cm	4,117	3.59	**0.008**	4,96	0.64	0.633	4,110	3.79	**0.006**
Vertical head position	−10 – 10 cm	4,83	1.78	0.140	4,69	4.98	**0.001**	4,76	6.45	**0.000**
Operating distance	50 – 80 cm	6,159	42.0	**0.000**	6,128	23.6	**0.000**	6,146	21.79	**0.000**
Illumination	0, 300, 600, 1000 lux	3,95	0.41	0.746	3,76	0.47	0.703	3,87	0.59	0.622
Background color	Black, White	1,48	0.00	0.970	1,38	0.27	0.606	1,44	0.41	0.527
Gaze angle	≤20°, 25°, 30°	2,71	6.76	**0.002**	2,57	1.76	0.181	2,66	10.4	**0.000**

precision) were affected for all ethnicities. Although Tobii (2010) claims that the trackable range should be between 50 and 80 cm, the trackability at 50 cm was far below 50 %. Despite the fact that the eyetracker is said to allow free head movement within the head box, accuracy and precision were acceptable only in a narrow range from 55 to 70 cm.

Illumination and background color of the stimulus

Although one might argue that darker conditions, which cause the pupil to dilate, might be better for eye tracking, it was found that illumination had no significant effect on accuracy or precision, while trackability for Caucasians was

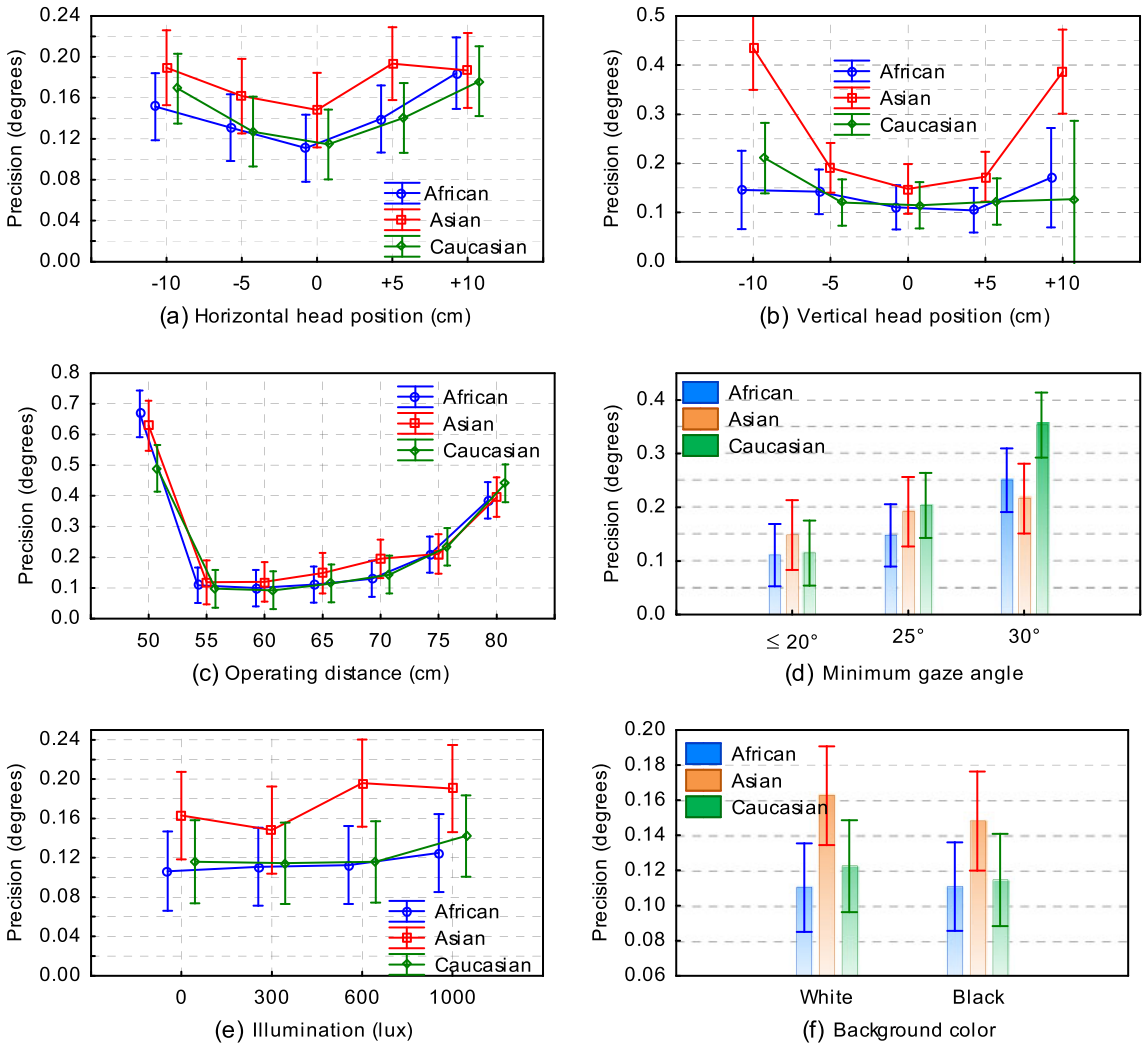

Fig. 12 Precision against various factors per ethnicity

Table 6 Summary of significant effects on data quality (✓, significant at α=0.1; ✓✓, significant at α=0.05; ✓✓✓, significant at α=0.01, ×, not significant)

		Trackability	Accuracy	Precision
Ethnicity (under ideal conditions)		✓	✓✓	✓
Horizontal head movement	Africans	✓	×	✓✓✓
	Asians	×	×	×
	Caucasians	×	×	✓✓✓
Vertical head movement[1]	Africans	×	×	×
	Asians	×	×	✓✓✓
	Caucasians	×	✓	✓✓✓
Operating distance	Africans	✓✓✓	✓✓✓	✓✓✓
	Asians	✓✓✓	✓✓✓	✓✓✓
	Caucasians	✓✓✓	✓✓✓	✓✓✓
Illumination	Africans	×	×	×
	Asians	×	×	×
	Caucasians	✓✓✓	×	×
Background color	Africans	×	✓✓✓	×
	Asians	×	×	×
	Caucasians	×	✓✓	×
Gaze angle	Africans	✓✓	✓✓✓	✓✓✓
	Asians	×	×	×
	Caucasians	×	✓✓✓	✓✓✓

[1]Since the −10/10 vertical positions were outside the head box for the specific eyetracker, the results for the smaller head box are used

worse in dark conditions than in the illuminated conditions. These results were as expected on the basis of the fact that remote eyetrackers rely on their own infrared illuminators to illuminate the eyes and provide a corneal reflection.

Similar to bright illumination in the room, a bright background color is also expected to cause the pupil to contract, while a dark background is expected to cause the opposite. It was found that background color does not affect trackability. Precision was also not affected by background color, but the accuracy achieved by African and Caucasian participants was better when the stimulus was presented on a dark background. These effects could be due to the software algorithm that is used by the manufacturer to identify the pupil in the eye video, and it might be different for another make of eyetracker or version of firmware.

Gaze angle

Gaze angle was a significant predictor of trackability for African participants only. The accuracy and precision achieved by African and Caucasian participants were also affected adversely by larger gaze angles. The accuracy and precision values of Asian participants were already significantly higher than those of Africans and Caucasians at small gaze angles and were not further increased by larger gaze angles. This result agrees with that of Blignaut, Holmqvist, Nyström and Dewhurst (2012) that accuracy of eye tracking is worse in the corners of the display than in the center thereof. Nyström, Andersson, Holmqvist and Van de Weijer (2013), who used another make and model of eyetracker, found, however, that the offset for off-center targets was not different from the offset for targets in the center of the screen.

Quality of eye tracking versus quality of the eyetracker

Although this research was done to study the effects of head position, illumination, and gaze angle with respect to three different ethnic groups, it is inevitable that a reader would use the results to infer some conclusions about the quality of the specific model of eyetracker that was used. It should be kept in mind, though, that any value for any aspect of data quality pertains to the specific unit and version of firmware that was used at the time. Another unit, another methodology, another policy with regard to participant screening and so forth could provide different results. For this reason, it is imperative that researchers use the same eyetracker for the duration of a study and do not compare values from one study with those from another.

Conclusions

This study endeavored to find trends with regard to the effect of various experimental conditions on the quality of eye-tracking data for different ethnic groups. It was found that trackability, accuracy, and precision for Asian participants was somewhat worse than that for African and

Caucasian participants. No significant differences were found between the latter two ethnic groups.

Trackability and accuracy were not affected by horizontal movement of the head, while precision was affected significantly for African and Caucasian participants. Vertical head movement proved to be a significant indicator of precision for all ethnic groups but did not affect trackability and accuracy as much.

Operating distance had the largest effect on data quality, since all indicators (trackability, accuracy, and precision) were affected for all ethnic groups

Illumination had no significant effect on accuracy or precision, while trackability was somewhat worse in dark conditions than in the illuminated conditions. Background color, on the other hand, does not affect trackability, but the accuracy achieved by African and Caucasian participants was better when the stimulus was presented on a dark background.

Large gaze angles proved to be detrimental for trackability of African participants, while the accuracy and precision achieved by African and Caucasian participants were also affected adversely by larger gaze angles.

The results from this study should not be used to compare different makes and models of eyetrackers with one another. Researchers should also not compare absolute values of data quality from one study with those from another, even if the experiments were conducted on the same model of eyetracker. The results of this study should only be used to ascertain that data quality may vary to some extent over ethnicity and experimental conditions.

The results have implications for participant recruiting, as well as execution of experiments. Participants should preferably be recruited from a homogenous ethnicity, and it is important to ensure that all participants for a study sit more or less in the same position in front of the eyetracker, especially with regard to operating distance. Room illumination and color of the stimulus might also have an effect on the outcome of results and should be kept constant for all participant sessions.

References

Abe, K., Ohi, S., & Ohyama, M. (2007). An eye-gaze input system using information on eye movement history. In C. Stephanides (Ed.), *Universal access in HCI, Part II. HCII2007, LNCS 4555* (pp. 721–729). Berlin Heidelberg: Springer-Verlag.

Blignaut, P., & Beelders, T. (2009). The effect of fixational eye movements on fixation identification with a dispersion-based fixation detection algorithm. *Journal of Eye Movement Research, 2*(5), 4. 1–14.

Blignaut, P. J., Holmqvist, K., Nyström, M., & Dewhurst, R. (2012). *Improving the accuracy of video-based eye-tracking in real-time through post-calibration regression*. Noosa: Proceedings of the 2012 Eye Track Australia Conference.

Borah, J. 1998. Technology and application of gaze based control. *RTO Lecture series on Alternative Control Technologies*, 7–8 October 1998, Brétigny, France and 14–15 October 1998, Ohio, USA

Brolly, X.L.C. & Mulligan, J.B. (2004). Implicit calibration of a remote gaze tracker. *Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop* (CVPRW'04), 8,134-146.

Chen, J., Tong, Y., Gray, W., & Ji, Q. (2008). *A robust 3D gaze tracking system using noise reduction*. Savannah: Proceedings of the Eye-tracking Research Applications Conference (ETRA).

Crane, H. D., & Steele, C. M. (1985). Generation-V Dual-Purkinje-Image Eyetracker. *Applied Optics, 24,* 527–537.

Crundall, D., Chapman, P., Phelps, N., & Underwood, G. (2003). Eye movements and hazard perception in police pursuit and emergency response driving. *Journal of Experimental Psychology Applied, 9*(3), 163–174.

Donovan, T., Manning, D. & Crawford, T. 2008. Performance changes in lung nodule detection following perceptual feedback of eye movements. Medical Imaging 2008 (SPIE).

Foulsham, T., & Underwood, G. M. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*(2), 1–17.

Goldberg, J., Stimson, M., Lewenstein, M., Scott, N. & Wichansky, A. 2002. Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 symposium on eye tracking research & applications*, 51–58.

Hansen, D. W., & Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*(3), 478–500.

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

Holmqvist, K., Nyström, M. and Mulvey, F. 2012. Eye tracker data quality: What it is and how to measure it. In *Proceedings of the 2012 Symposium on Eye Tracking Research and Applications* (*ETRA 2012*). Santa Barbara, CA, 45–52.

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*(4), 592–604.

Hua, H., Krishnaswamy, P., & Rolland, J. P. (2006). Video-based eyetracking methods and algorithms in head-mounted displays. *Optics Express, 14*(10), 4328–4350.

Imai, T., Sekine, K., Hattori, K., Takeda, N., Koizuka, I., Nakamae, K., ... Kubo, T. (2005). Comparing the accuracy of video-oculography and the sclera search coil system in human eye movement analysis. *Auris Nasus Larynx, 32,* 3–9.

ISO 5725–1. (1994). *Accuracy (trueness and precision) of measurement methods and results – Part1: general principles and definitions*. Geneva, Switzerland: International Standards Organisation.

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.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, Portland, 1703–1708.

Johnson, J. S., Liu, L., Thomas, G., & Spencer, J. P. (2007). Calibration algorithm for eyetracking with unrestricted head movement. *Behavior Research Methods, 39*(1), 123–132.

Kliegl, R., & Olson, R. K. (1981). Reduction and calibration of eye monitor data. *Behavior Research Methods and Instrumentation, 13,* 107–111.

Komogortsev, O. V., & Khan, J. I. (2008). *Eye movement prediction by Kalman filter with integrated linear horizontal oculomotor plant mechanical model* (pp. 229–236). Savannah: Proceedings of the 2008 Symposium on Eye Tracking Research and Applications (ETRA).

Nyström, M., Andersson, R., Holmqvist, K., & Van de Weijer, J. (2013). The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods, 45*(1), 272–288. doi:10.3758/s13428-012-0247-4. Springer.

Rayner, K., Pollatsek, A., Drieghe, D., Slattery, T., & Reichle, E. (2007). Tracking the mind during reading via eye movements: Comments on Kliegl, Nuthmann, and Engbert (2006). *Journal of Experimental Psychology: General, 136*(3), 520–529.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision, 7*(14), 1–17.

Tobii 2010. Product description for the Tobii TX300 eye tracker. Tobii Technology AB, November 2010. Product flyer downloaded from www.tobii.com on 6 July 2012.

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.

Vikström, K., Wallin, A. & Holmqvist, K. 2009. Yarbus goes shopping. Scandinavian workshop on on applied eye-tracking (SWAET), Stavanger, 5–7 May 2009.

Zhang, Y., & Hornof, A. J. (2011). Mode of disparities error correction of eye tracking data. *Behavior Research Methods, 43*(3), 834–842.