INFANCY

THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES

Infancy, ***20***(6), 601–633, 2015

ISSN: 1525-0008 print / 1532-7078 online
DOI: 10.1111/infa.12093

WILEY Blackwell

Consequences of Eye Color, Positioning, and Head Movement for Eye-Tracking Data Quality in Infant Research

Roy S. Hessels
Department of Experimental Psychology, Helmholtz Institute, Utrecht University and Department of Developmental Psychology, Utrecht University

Richard Andersson
Eye Information Group, IT University of Copenhagen and Department of Philosophy & Cognitive Science Lund University

Ignace T. C. Hooge
Department of Experimental Psychology, Helmholtz Institute Utrecht University

Marcus Nyström
Humanities Laboratory Lund University

Chantal Kemner
Department of Experimental Psychology, Helmholtz Institute, Utrecht University and Department of Developmental Psychology, Utrecht University and Brain Center Rudolf Magnus University Medical Centre Utrecht

Correspondence should be sent to Roy S. Hessels, Heidelberglaan 1, 3584 CS Utrecht, The Netherlands. E-mail: royhessels@gmail.com

Eye tracking has become a valuable tool for investigating infant looking behavior over the last decades. However, where eye-tracking methodology and achieving high data quality have received a much attention for adult participants, it is unclear how these results generalize to infant research. This is particularly important as infants behave different from adults in front of the eye tracker. In this study, we investigated whether eye physiology, positioning, and infant behavior affect measures of eye-tracking data quality: accuracy, precision, and data loss. We report that accuracy and precision are lower, and more data loss occurs for infants with bluish eye color compared to infants with brownish eye color. Moreover, accuracy was lower for infants positioned in a high chair or in the parents' lap compared to infants positioned in a baby seat. Finally, precision decreased and data loss increased as a function of time. We highlight the importance of data quality when comparing multiple groups, as differences in data quality can affect eye-tracking measures. In addition, we investigate how two different measures to quantify infant movement influence eye-tracker data quality. These findings might help researchers with data collection and help manufacturers develop better eye-tracking systems for infants.

According to Aslin (2007), "It is no exaggeration to say that without looking time measures, we would know very little about infant development." Over the years, researchers have employed measures of looking behavior to investigate infants' detection, discrimination, and preferences for visual stimuli in a time when they cannot express their preferences or respond verbally (Aslin, 2007). In earlier studies, observers would code where infants were looking at to estimate global looking times at visual stimuli. The rise of eye tracking as a tool in infant research has induced a shift from investigating the macrostructure to the microstructure of infant looking behavior (Aslin, 2007, 2012). Eye trackers are now used to study, for example, infant oculomotor characteristics (Wass & Smith, 2014), object perception (Amso & Johnson, 2006), face processing in typically developing infants (Wheeler et al., 2011), and in infants at risk for autism spectrum disorders (Jones & Klin, 2013).

Remote video-based eye trackers are devices that commonly illuminate the eye with an infrared light that reflects off the cornea. The reflection of this infrared light (the corneal reflection) and the position of the pupil are then registered by the eye tracker. A calibration sequence is subsequently run to transform the position of the corneal reflection and the pupil into gaze position on the screen. Remote video-based eye trackers are particularly popular for infant research: They can be positioned in front of the infant and do not interfere with the infant unlike, for example, EEG measurements. In addition, they allow infants to move their head in front of the

eye tracker. Finally, remote eye trackers are often easy to operate. As a result, eye trackers are currently common in many infant research facilities (Aslin, 2012; Oakes, 2012). The switch from having observers code infants' gaze direction to having a machine compute, it has several advantages. First, gaze estimation by an eye tracker is objective. Second, the spatial and temporal resolutions with which gaze direction can be determined are markedly higher for eye tracking than for manual coding of videos.

While eye tracking seems an attractive method for investigating all aspects of infant looking behavior, there are several pitfalls to consider. First, while eye-tracking methodology for adults has received considerable attention (see Holmqvist et al., 2011 for an extensive overview), eye-tracking methodology for infant research has not (see, e.g., Wass, Forssman, & Leppänen, 2014). This is particularly important, as infants tend not to behave as adults would in front of an eye tracker. Adults can usually be instructed on how to behave during an eye-tracking experiment. In addition, their head movements can be restrained using a chinrest, headrest, or bite bar, thereby limiting the interference of movement with the output of the eye tracker. Infants, on the other hand, cannot readily be instructed to perform a certain task, to remain still, or to focus on a particular stimulus during calibration. Whether and how the quality of eye-tracking data is affected by the behaviors typically seen in infant research is currently unclear (although see Wass et al., 2014 for some first examples). Secondly, analyzing and interpreting eye-tracking data is a laborious endeavor (Gredebäck, Johnson, & von Hofsten, 2009). It requires not only knowledge of the physiology of the eye, which may differ between adults and infants, but also of signal processing and the technical limitations of the eye tracker (Oakes, 2010). In the present study, we investigate which factors affect data quality in infant eye tracking. In doing so, we hope to help researchers achieve higher data quality and include data from more participants and help manufacturers develop better systems for infant eye-tracking research.

Data quality and its consequences for eye-tracking data analyses

How does one judge the quality of eye-tracking data? A prerequisite is to know what exactly eye-tracking data quality is (Holmqvist, Nyström, & Mulvey, 2012; Nyström, Andersson, Holmqvist, & van de Weijer, 2013). Data quality refers to a property of the raw eye-tracker data and is often operationalized in several measures, three of which we describe here: spatial accuracy, spatial precision, and data loss (Holmqvist et al., 2011). We will define these data quality measures one at a time and discuss the possible consequences for event detection, that is, identifying eye movements

(saccades) and periods that the eye is relatively still (fixations) in the eye-tracker signal. Moreover, we will discuss the possible consequences for the calculation of eye-tracking measures when data quality is low. First, it should be noted that the accuracy, precision, and amount of data loss that is observed during a measurement stems from a combination of both participants, environmental and eye-tracker-related factors. The accuracy and precision that can be achieved using a particular eye tracker are often reported by the manufacturer. These values are usually achieved under optimal conditions (i.e., using artificial eyes or cooperative adult participants) and are difficult to achieve when conducting eye-tracking research with infants (see Hessels, Cornelissen, Kemner, & Hooge, 2014 for a discussion). Regardless of whether decreased accuracy or precision stems from the eye tracker or the participant, they impose restrictions on what conclusions can be drawn from the eye-tracking data.

Accuracy refers to the systematic offset between the gaze position as reported by the eye tracker and a known target position. Accuracy is calculated by having participants fixate a target and calculating the offset between this target and the gaze position reported by the eye tracker. Figure 1b depicts an example trial in which accuracy might be calculated. The distance between the center of the fixation target and the mean of the gaze coordinates from the left and right eye is an estimate for the accuracy. Stimuli should be of sufficient size and surrounded by a sufficiently large margin to account for the (in)accuracy during a measurement. If accuracy is low, this might result in a decreased total dwell times on area of interests (AOIs) if the margins around the AOIs are too small (Holmqvist et al., 2012).

Precision refers to the sample-to-sample error during a recording (Holmqvist et al., 2012). Precision can be computed in a number of ways (Holmqvist et al., 2011), but a common measure for precision is the root mean square (RMS) of the Euclidean distances between samples during a period where the eye is still. The cluster of gaze coordinates near the center of the fixation target in Figure 1b is an example of an imprecise fixation. Low precision has consequences for the separation of fixations and saccades. If a fixed velocity threshold is used to separate fixations from saccades, low precision (i.e., a higher RMS value) might result in a larger number of samples spuriously exceeding this threshold. The number of fixations may then apparently increase (Wass, Smith, & Johnson, 2013). If a velocity algorithm with a threshold that adapts to the noise level is used, this might result in an overall higher threshold. Small saccades might thereby remain below the velocity threshold and the number of fixations apparently decreases (Holmqvist et al., 2012). When the number of detected fixations changes, so does the average fixation duration associ-

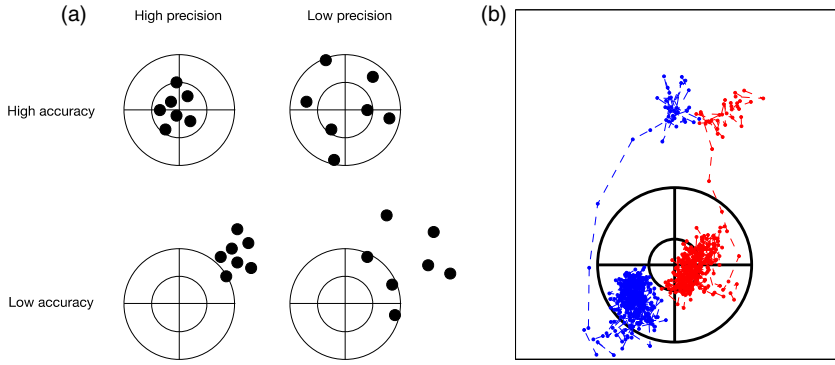

Figure 1 (a) Schematic overview of accuracy and precision, given that the participants gazes at the center of the bulls eye. Dots represent consecutive gaze positions reported by the eye tracker. (b) Example data from a validation trial. Red dots represent data from the right eye; blue dots represent data from the left eye. The bull's eye is a schematic representation of a fixation target.

ated with these fixations, as multiple fixations are either merged into one fixations or split into multiple fixations (see Shic, Chawarska, & Scassellati, 2008 for possible consequences thereof). Both accuracy and precision may vary independently in eye-tracking data. Figure 1a depicts a schematic overview of the possible combinations of accuracy and precision.

Data loss, the final aspect of data quality, refers to the proportion of valid samples that the eye tracker reports (Nyström et al., 2013). An invalid sample occurs when the eye tracker does not report a gaze position. This might be because a participant is looking outside the tracking area (e.g., away from the screen) or because a participant's eyelids are closed due to a blink. Data loss can, however, also occur when the participant is directed toward the screen and the participant's eyes are open. This might be for a number of reasons: It could be that the eye tracker is unable to detect the eyes, the pupil, or the corneal reflection. There are several possible consequences of data loss: Detecting fixations or saccades cannot be accomplished for periods of data where gaze positions cannot be computed, unless interpolation methods are applied to resolve data loss. Furthermore, if a short period of data loss occurs during a fixation, it might appear as if that fixation is actually two separate, shorter, fixations. In addition, data loss might lead to apparently longer latencies, and increased variability, of gaze shifts toward a target in the data (Wass et al., 2014).

Although the discussion on eye-tracking data quality has become more prominent over the last few years, reporting data quality measures is still uncommon in the field of infant eye tracking. Current guidelines for pub-

lishing infant eye-tracking studies do not specify that measures of accuracy, precision, and data loss achieved during measurement to be included (Oakes, 2010). Inclusion of participants or trials based on data quality is mostly based on the amount of data that was recorded. Chawarska and Shic (2009); Hunnius, de Wit, Vrins, and von Hofsten (2011); Shic, Macari, and Chawarska (2013), for instance, excluded trials based on inattention, and Amso, Haas, and Markant (2014) excluded participants with <30% valid data achieved during a recording. In addition, there are studies that specify a minimum accuracy at calibration before preceding with the study (e.g., Johnson, Amso, & Slemmer, 2003), or studies implementing post hoc procedures to correct for inaccurate data (Frank, Vul, & Saxe, 2011). We are unaware of any studies using estimates of precision as exclusion criteria in infant eye-tracking research. The issue of data quality is not only limited to post hoc exclusion or inclusion of data but achieving high data quality for the data that are ultimately included is equally important. The question that then arises is how exactly high data quality can be achieved? What are the factors that affect data quality, and what are their relative contributions?

How can higher data quality be achieved?

Using adult participants, several studies have examined which factors might influence data quality of video-based eye trackers. Kammerer (2009) notes that for participants with "bright-colored" (i.e., bluish color) eyes, accuracy was lower compared to participants with darker eyes using a Tobii 1750 remote eye tracker. In addition, accuracy was lower for participants with glasses compared to participants without glasses or with contact lenses (Kammerer, 2009). Two studies systematically investigated factors that might influence data quality: Nyström et al. (2013), using a SMI HiSpeed 500 Hz tower-mounted eye tracker, and Blignaut and Wium (2013), using a remote Tobii TX300 eye tracker. Nyström et al. (2013) investigated whether calibration method, visual aids, eyelash direction, eye color, the use of mascara, eye dominance, pupil diameter, recording number, and the position of a target on screen affected accuracy, precision, and data loss for 149 adult participants. They confirm the results by Kammerer (2009) that data quality is lower for participants with bright eye colors compared to participants with dark eye colors. In addition, Nyström et al. (2013) highlight the differences in data quality based on the calibration method used: operator-, participant-, or system-controlled. Blignaut and Wium (2013) report that accuracy and precision are lower for Asian participants compared to African and Caucasian participant. Among others, the operating distance from the eye tracker affected

accuracy, precision, and trackability (i.e., the complement to the proportion of data loss), and the vertical head position affected trackability. Finally, Holmqvist et al. (2011) highlight several more specific details of video-based eye trackers that might affect precision and accuracy based on their observations. All in all, there are a large number of factors that may increase or decrease data quality for adult participants.

Eye tracking with infant participants has, however, been less investigated. As Aslin and McMurray (2004) already pointed out, eye tracking with infants can be more difficult than can be expected with adult participants. The differences between eye tracking with infants and eye tracking with adults are most easily observed in raw data. Figure 2 depicts an example trial of both an adult and an infant participant. The raw data for the adult participant are precise and the eye tracker continuously reports gaze. The raw data for the infant participant, however, are less precise, and contain periods of data loss due to inattention by the infant and instable tracking of the eyes. The differences between adult and infant eye-tracking data may be due to a number of reasons. First, choosing an eye tracker that is suitable for infant research is a difficult matter, particularly when the technical specifications of an eye tracker do not necessarily predict its ability to cope with infant behavior during measurement (Hessels et al., 2014). Second, seating an infant in front of an eye tracker can be performed in numerous ways; we do not yet know how this affects eye-tracker data quality. Infants can, for instance, be positioned in the lap of the parents (e.g., Gredebäck, Fikke, & Melinder, 2010; Wass & Smith, 2014), in a high chair on the floor, or strapped in a baby seat (e.g., Jones, Carr, & Klin, 2008; Shic et al., 2013). Each type of seating may place different requirements on how the eye tracker can be maneuvered to achieve the right geometry of the setup. In addition, each type of seating may allow more or less movement during a measurement. When possible, infants tend to move around during measurements, which changes both the position and orientation of the eyes relative to the eye tracker (Wass et al., 2014). When movements that are often observed in infant research (i.e., looking away from the screen, and shifted head orientations) were modeled by adult participants, data loss and possibly systematic offsets were observed (Hessels et al., 2014). Notably, the amount of data loss and the severity of possible systematic offsets depended on the eye tracker tested. Third, infants are more difficult to calibrate due to inattention to small calibration targets (Aslin & McMurray, 2004). While larger, moving calibration targets have been used as replacements for infants, little is known about how this might affect the accuracy of measurements. Finally, there might be physiological differences between infants and adults that render eye tracking more

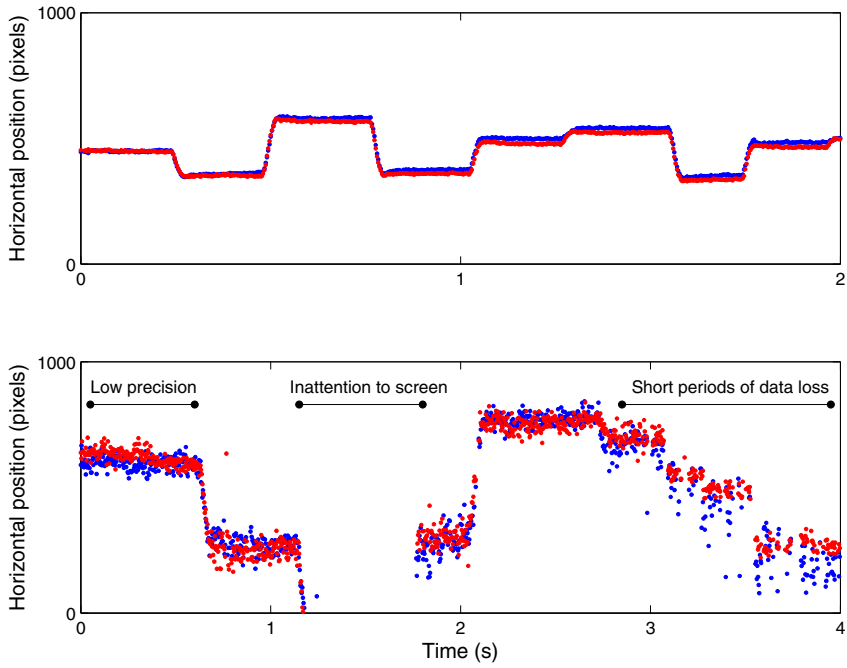

Figure 2 Raw data from an example trial for an adult participant (top panel) and an infant participant (bottom panel) recorded with the Tobii TX300 at 300 Hz. Red dots are coordinates from the right eye; blue dots are coordinates from the left eye. The infant data are less precise than the adult data and contains periods of data loss due to inattention and other acquisition problems. For both adult and infant data, fixations and saccades can still be identified.

difficult. Wass et al. (2014), for example, suggest that an infant's increased pupil size can make tracking problematic, as they are hard to detect for pupil detection algorithms built for adult pupil size. Furthermore, Tobii, a well-known manufacturer of eye trackers in infant research, developed an illumination mode for their TX300 model that is specifically designed for infants.[1] We speculate that this is for differences in the reflection of infrared light between infants' and adults' skin, making detection of the pupil and corneal reflection more troublesome as they become darker in the image when the white balance is adjusted. However, little is known about whether this infant illumination mode

[1]http://www.tobii.com/Global/Analysis/Downloads/User_Manuals_and_Guides/Tobii_TX300_Eyetracker_UserManual.pdf

increases the data quality or tracking performance of the eye tracker, or how skin reflectance of infrared light affects eye-tracking data quality in general.

We are only aware of one study that has specifically examined data quality in infant eye-tracking research. Wass et al. (2014) report a negative correlation between amount of head movement and precision, and precision appears to be lower late in a measurement than early in the measurement. In addition, Wass et al. (2014) report that both head movement and time since the start of the measurement are correlated with data loss. They employ the average duration of valid data periods as a measure for data loss: Shorter periods of continuous valid data would indicate more data loss. Wass et al. (2014) suggest that this measure better reflects data loss due to unstable tracking of the eyes than data loss due to inattention of the infant. The reasoning behind this is that data loss in infant studies differs from that in adult studies. In adult studies, data loss is often reported as the proportion of invalid samples during a trial or experiment. Usually, the adult participants are instructed to attend to the screen the entire time. If we assume participants follow the instruction, any data loss stems either from blinks or technical difficulties during the measurement. In infant eye-tracking research, however, participants are not expected to attend to the screen the entire time. Data loss can therefore stem not only from blinks or technical difficulties, but also inattention (i.e., looking away from the screen). In order to measure data loss in infant eye-tracking research due to unstable tracking of the eyes, periods of data loss stemming from blinks and inattention should therefore be excluded. We introduce such a measure for data loss in this study.

Concluding, it would seem that both time since the beginning of a measurement and movement during a measurement decrease data quality in infant eye-tracking research. There have, however, been no endeavors to systematically investigate participant and measurement characteristics that affect eye-tracking data quality with infants. We address this gap in the literature by investigating the effect of predictors we deem most influential on infant eye-tracking data quality in a group of 10-month-old infants who visited our laboratory twice. Although the purpose of the two visits was to examine test–retest reliability in a visual search task, this study only concerns data quality assessment. The following predictors were investigated. Eye physiology—the color of the eye, the direction of the eyelashes, and the size of the opening of the eye—is suggested to affect eye-tracking data quality for adult participants (Blignaut & Wium, 2013; Nyström et al., 2013). For infants, however, the pigment

that produces the color of the eye is still being formed; most European infants (i.e., with a light skin color) have bright eyes at birth, with pigment formation occurring throughout the first year after birth. In addition, eyelashes for infants tend to be sparser than for adults. Whether eye-tracking data quality is then also affected for infants is a question in this study. Moreover, where adult participants can be positioned to the researcher's liking and movement can be restricted, the same does not apply for infants. We investigate whether the manner of seating the infant, and the amount of movement it allows, affects data quality. Finally, we wonder whether the behavior and contentedness of the infant during the measurement affects data quality. In infant research (in general, not only eye-tracking research), excluding participants due to "fussiness" is common, sometimes up to a third of the participants (Cassia, Turati, & Simion, 2004; Leppänen, Moulson, Vogel-Farley, & Nelson, 2007). Although it is difficult to find specific criteria for excluding fussy infants, common criteria include the infant being upset, hungry, or sleepy. We investigate whether factors related to fussiness also affect eye-tracking data quality. Our findings might help researchers achieve higher data quality and increase the throughput of infants from data recording to data analysis. In addition, our findings might help manufacturers develop better eye-tracking systems for infant eye-tracking research.

METHODS

Participants

Seventy-seven infants were invited into the laboratory center for a larger study, recruited through the local municipality. Of the 77 infants invited, 75 (39 male, 36 female) participated in a first session of the present eye-tracking study. Sixty-one (29 male, 32 female) of the 75 that completed the eye-tracking experiment on their first visit (which we will refer to as a session) returned to the laboratory center for a second session. A total of 136 sessions were thereby recorded for this study. Mean age during the first session was 302.8 days (SD = 12.8 days); mean age during the second session was 307.5 days (SD = 11.2 days). Infants were only invited to participate if the parents indicated that the infants were not born preterm (i.e., before 37 weeks of pregnancy) and had no impaired hearing or vision or developmental disorders. Parents gave written informed consent on the day of the first session, and the study was approved by the ethics committee of the local University Medical Centre

(Protocol ID 14-221). Parents received a 10 € compensation for each testing day, with another 5 € travel compensation if required.

Operators

As Nyström et al. (2013) report that different operators may achieve varying levels of data quality, we incorporated the operator into our statistical analysis (see Statistical analysis for details). Four operators performed the data recordings. Two operators had previous experience with infant eye-tracking recordings using the eye tracker in this study. One operator had extensive experience with eye tracking in adult participants using multiple systems, but not with infants. The final operator was newly trained and had recorded data in approximately five eye-tracking sessions with infants and five with adults. The present experiment was part of a larger pilot study in which infants completed a number of tasks across the day. Consequently, the specific time of the day and operator for the eye-tracking sessions was not set in advance; whoever was available performed the data recording. All operators, regardless of experience level, were given the same training in eye tracking with infants by the first author prior to the start of the study.

Apparatus

Stimulus presentation was handled by MATLAB R2013a and the Psych-Toolbox (version 3.0.11; Brainard, 1997) running on a MacBook Pro with OS X 10.9. Stimuli were presented on an external 23-inch screen belonging to the Tobii eye tracker at a resolution of 1920 by 1080 pixels and a refresh rate of 60 Hz. The Tobii TX300 eye tracker running at 300 Hz was used for tracking infants' eye movements. The TX300 is capable of recording at 0.4° accuracy (binocular) and 0.14° precision under ideal conditions.[2] As the accuracy and precision that can be achieved while recording infants is one of the aims of this study, the actual precision and accuracy values will be presented in the results section. The Tobii SDK was used for communication between MATLAB and the eye tracker.

Stimuli

The experiment consisted of 24 visual search trials (based on Amso & Johnson, 2006), used for the calculation of flicker and RMS noise (see

[2]http://www.tobii.com/Global/Analysis/Training/Metrics/Tobii_TX300_Eye_Tracker_Accuracy_and_Precision_Test_Report.pdf

"Data analysis" section). Each visual search display consisted of 28 white lines (3.3° by 0.9°) as target candidates on a black background. The lines were arranged in a grid of 14 columns by two rows and subsequently jittered between −1.6° and 1.6° in horizontal and between −6.3° and 6.3° in vertical direction. All lines except the target line were aligned vertically. The target line was tilted 30°, 60°, or 90° clockwise and could appear in one of eight fixed locations. Each combination of target line angle and location was presented once, resulting in 24 trials. The visual search trials were interspersed with validation trials after the first and every additional fifth trial for a total of five validation trials. Each validation trial contained one validation target that was identical to one of the calibration targets. The validation targets were used to calculate offset (see "Data quality measures").

Preceding the visual search experiment was a 5-point calibration sequence. Each calibration and validation stimulus consisted of a colored spiral (red, green, yellow, purple, or blue) on a black background. The spiral changed in size between 4.0° and 5.4° at 0.8 Hz following a sinusoidal wave. In addition, the spiral rotated at 0.8 Hz. Following a key press of the operator, the spiral shrank in size to 0.5° over a period of 0.5 sec. The spiral then remained on screen for 0.2 sec. For the calibration sequence, a point was calibrated at the start of this 0.2-sec period. For the validation trials, data was recorded throughout (see Accuracy and precision for more details).

Procedure

Positioning

The infants and parents were welcomed into the eye-tracking room and familiarized with the experimental setup. Thereafter, the infants were strapped in a baby seat, and the parent was seated on a height-adjustable chair. The baby seat was subsequently placed on the parent lap, with the infant placed parallel to the screen of the eye tracker. Positioning the infant in a baby seat was performed as this would give the most stable positioning through the recording and limit the infants' movements. If, however, the parent indicated that the baby seat would probably result in a restless or upset infant, the infant was seated without a baby seat in the parents lap or in a high chair. The decision for either the parents lap or the high chair was up to the judgment of the operator, that is, which of the two would work best for the particular infant. While this manner of seating introduces a selection bias for the conditions other than baby seat, the seating predictor is included for the following reason: If data quality is

reduced after choosing a different type of positioning, this is important to consider when conducting eye-tracking research in infancy, regardless of whether the seating itself is the (only) cause of this reduced data quality. We outline this issue further in the discussion. After positioning the parent and infant, the position of the eye tracker was adjusted so that the eyes of the infant were at 65 cm from the eye tracker and at the same height as the center of the screen.

Calibration and experiment

After positioning, a 5-point calibration sequence was started. Calibration stimuli were serially presented in the four corners and center of the screen. The order of points was random each time the calibration was run. The infant was monitored with a webcam. The operator judged from this video whether the infant looked in the direction of the calibration stimulus and pressed the spacebar to calibrate the current point. After the calibration sequence, the calibration output[3] was examined. As the Tobii SDK does not provide an objective measure for the accuracy of a calibration, the calibration output was examined for two features. Calibration points that were either without data or with data that were inconsistent and characterized by dispersed gaze points around the calibration point were re-calibrated by the operator. Each re-calibration was noted down as an additional calibration run. After calibration was deemed successful, or when the infant started losing attention, the experiment was initiated. A central static attention getter (i.e., a colorful picture) preceded each visual search trial. The operator initiated the trial by pressing the spacebar when the infant was judged to look at the screen. After the first, and each additional fifth trial, a validation target was presented at one of the five calibration locations. When the operator judged from the video that the infant looked in the direction of the validation target, the spacebar was pressed and the validation point shrank in size (see the section "Stimuli"). If the infant did not attend the validation target for whatever reason, the operator pressed the spacebar as well. If, during the experiment, the infant was not attending to the screen, the operator could present attention getting sounds or videos with sound in the center of the screen. The entire experiment, including calibration and positioning, lasted approximately 10–15 min.

[3]The calibration output is outlined in the manual of the Tobii SDK, see http://www.tobii.com/en/eye-tracking-research/global/products/software/tobii-analytics-software-development-kit/

Collected predictors

The predictors that were hypothesized to influence eye-tracking data quality were recorded for all infants where possible. The predictors were divided into three groups: eye physiology, measurement characteristics, and infant contentedness and behavior. The predictors that were included in the statistical analysis are given in Table 1 with the number of levels and data points that were available for each predictor.

For eye physiology, the eye color, direction of eyelashes, and size of the eye opening were determined. The operator made the initial assessment of eye color, and pictures were taken to allow a second assessment. Eye color was scored as either "bluish" or "nonblue," to match previous research (Nyström et al., 2013). The eyelash direction was scored as either "upward" or "downward," and the size of eye opening as "open" or "narrow." After assessment of eyelash direction and size of the eye opening, there were five or less instances out of all participants in which "downward" or "narrow" was scored. Both eyelash direction and size of the eye opening were therefore not included in the analysis.

The measurement was characterized by three factors: number of calibrations, seating, and time since the start of the measurement. Number of calibrations was scored between one and four, where a value of four indicated that more than three calibration sequences were run. Seating was

TABLE 1
Predictors Included in the Statistical Analysis

Predictor	*Values*
Intercept	Not a predictor. Is given in the results of the statistical model. The intercept of the model represents the value for a theoretical group of infants having the reference value for all categorical predictors, and all numerical predictors set to zero
Eye color	**Bluish** (45), nonblue (26), could not be determined for four participants
Seating	**Baby seat in parents lap** (111), directly in parents lap (15), in high chair (10)
Number of calibrations	Numerical value between 1 and 4, see *collected variables* for details
Movement	Numerical value between 0 and 4, see *collected variables* for details
Time since fed	Numerical value in hours
Time since awoken	Numerical value in hours
Trial	Numerical value indicating trial number. 1–5 for accuracy. 1–24 for precision and flicker

Bold-faced text indicates reference level of categorical predictors. Values in parentheses are number of instances in dataset for value of the predictor.

noted as either "baby seat in parents lap," "directly in parents lap," or "in high chair." Finally, a predictor for trial was included to investigate the development of data quality over time since the start of the experiment.

To characterize infant contentedness and behavior during the experiment, three separate measures were collected. The time since the infant was last fed and last awoken was collected. This was performed by having the parent report the time the infant last ate and woke up. If infants slept or ate during the time that they were in the laboratory center, this was noted down. Finally, time since last fed and last awoken was retrieved in minutes by calculating the time difference to the start of the experiment. In addition, the operator scored the overall movement of the infant during the experiment. Operators were instructed to monitor the infant movement throughout the experiment and estimate the percentage of the total time in the experiment that the infant moved. Movement was scored between zero and four. Zero indicated that the infant did not move noticeably during the experiment. A score of one indicated the infants moved between 0 and 25% of the time, a score of two between 25 and 50%, three between 50 and 75%, and four between 75 and 100%. As the present coding scheme was used in prior studies, operators were familiar with this coding scheme as an estimate for movement prior to the start of this study.

Data analysis

To investigate the influence of the suggested predictors on eye-tracking data quality (i.e., measures for accuracy, precision and data loss), several data reduction steps were taken. First, periods of data loss were extracted from the raw eye-tracking data for the data quality measure of data loss. Second, raw eye-tracking data were reduced into fixations. Third, data quality measures for accuracy and precision were extracted. Finally, statistical models were applied to three data quality measures. We will outline the separate steps in more detail below.

Data loss

Based on experience with infant eye-tracking data and previous research (Wass et al., 2014), we assume that data loss due to unstable tracking of the infants' eyes by the eye tracker (sometimes referred to as technical difficulties) is mainly represented in short periods of data loss. We therefore only investigate data loss where its duration is below 100 ms. We choose this cutoff to exclude periods of data loss due to blinks (roughly 100–400 ms) or periods of inattention to the screen (typically longer than 400 ms). As visible from Figure 3, roughly 90% of all

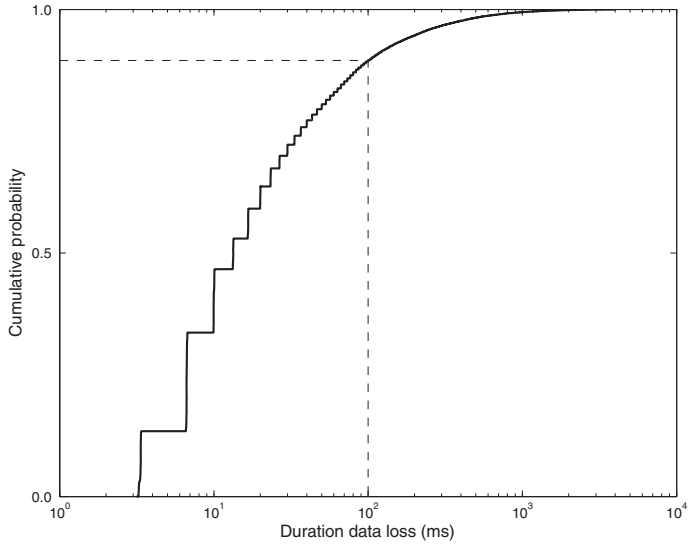

Figure 3 Cumulative probability of the duration of data loss periods in the present study. The dashed line indicates the probability of data loss period occurrences of 100 ms or lower. The steps in the graph correspond to the intersample interval (3.33 ms for 300 Hz data): Data loss duration is a multiple of the intersample interval.

periods of data loss observed in this study had durations of 100 ms or less. We will henceforth refer to these periods of data loss shorter than 100 ms as flicker. The proportion of flicker was calculated by dividing the total duration of flicker in a trial by the total duration of flicker and total duration of all valid samples (i.e., when a gaze position is reported by the eye tracker) combined. We thereby excluded all periods of data loss above 100 ms in the calculation of this measure. For the example eye-tracking data in Figure 2, the measure would be as follows: the proportion of the sum of all data loss noted under "short periods of data loss," to the amount of data points available (including the short periods of data loss). The longer period of data loss noted under "inattention to screen" is not included in this measure. For each infant, up to 24 values for proportion flicker as a measure for data loss were obtained in this manner, one for each of possible 24 visual search trials.

Eye-tracking data reduction

Raw position signals from the left and right eye were first combined into an average position signal. If gaze position was only available from one eye, that signal was used. Hereafter, a fixation detection algorithm specifically

designed for use in infant data was applied. The algorithm operates as an adaptive dispersion algorithm, with which fixation detection can be achieved across larger variations in noise levels, both local and between participants or trials. The algorithm, Identification by 2-Means Clustering (I-2MC), is based on a procedure called k-means clustering (where $k = 2$), which is used to determine whether one or two fixation clusters are present in a small moving window. As the I-2MC algorithm employs a moving window in which clustering is carried out, it is robust to variations in local noise. In this study, we used a moving window of 200 ms width.

Accuracy and precision

Accuracy was estimated by determining the offset between the center of the validation targets and the fixation closest in space to this center. To increase the reliability of the fixation position chosen for determining offset, cubic spline interpolation was performed prior to the event detection. Interpolation was performed for periods of data loss with a duration of <100 ms (Frank, Vul, & Johnson, 2009). This interpolation increased the robustness of fixation detection when short bursts of data loss could occur, thus making the calculation of offset possible more reliable in the presence of data loss. As there were five validation trials, a maximum of five estimates of accuracy per infant were obtained. Note that offset is a measure for accuracy, and a higher offset means a lower accuracy.

Precision was estimated by calculating RMS noise for all fixations during the 24 visual search trials. Note that no interpolation was performed for the estimates of precision, in order to ensure that the RMS noise was not affected or determined by the interpolation method chosen. The number of estimates of precision acquired for each infant depends on the number of fixations they made over the entire visual search experiment. Note that RMS noise is a measure for precision and a higher RMS noise means a lower precision.

Histograms for offset, RMS noise, and proportion flicker are presented in Figure 4. As the distributions for offset and RMS noise were highly skewed and contained a number of large values, a log-transformation was applied to better portray the distribution. As visible from Figure 4, there are quite a few large values that have been obtained for offset. While these values may seem absurdly large, we cannot be sure that these values are incorrect. As described above, the validation stimulus was already 5.4° at its largest, and the infant may have looked anywhere on or around this target. We will therefore not exclude high offset values and discuss this in more detail in the "Discussion".

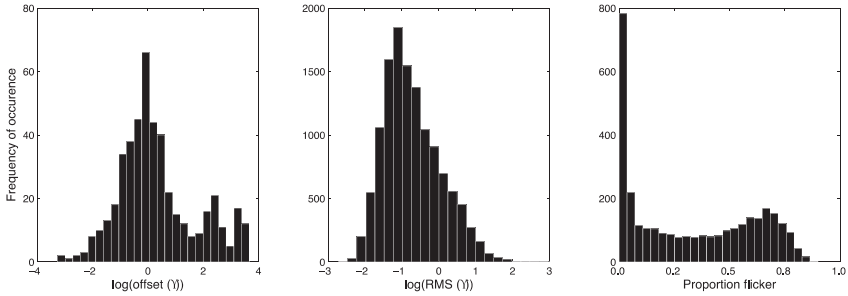

Figure 4 Histograms for the log-transformed offset (estimate for accuracy), log-transformed RMS noise (estimate for precision), and proportion flicker (measure of data loss).

Statistical analysis

The influence of the predictors given in Table 1 on data quality measures for accuracy, precision, and data loss was statistically tested by fitting linear mixed-effects models (LMEMs) using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014). Linear mixed-effects models are particularly suited for analyzing data with repeated measurements on both continuous and categorical variables. In addition, LMEMs are superior compared to common statistical analyses (e.g., ANOVAs) when dealing with unbalanced and missing data (Baayen, Davidson, & Bates, 2008).

For all three data quality measures, LMEMs were constructed with operator and participant session as random effects with random intercepts, and the predictors in Table 1 as fixed effects. Random factors are recommended to have at least six levels, and we have only four levels. To test whether this was a problem, we tested all predictors using operators as a fixed effect, both as a main effect and an interaction with the other predictors, and we found that the operators had little effect on the predictors of interest. So in order to simplify the models, and because we were not interested in the operators *per se*, we modeled the operators as random intercepts. As mentioned, the distributions of offset, RMS noise, and proportion flicker were skewed. Log-transformations were applied to offset and RMS noise to acquire Gaussian-like distributions. In addition, a logit transformation was applied to the proportion of flicker. Similar transformations were applied in a previous data quality study with adult participants (Nyström et al., 2013). Estimates for the effect of predictors on the data quality measures were acquired by constructing models with the largest maximal random effects structure that

would converge (Barr, Levy, Scheepers, & Tily, 2013): The effect of each predictor was modeled as a fixed effect and also as a random slope for each participant to account for the variance in effect for each participant. Additionally, the variance of the participants was modeled as random intercepts, and intercept–slope correlations were used to capture any interaction between the participants' intercepts and the effect of the predictor. These models also had operator as random intercepts. If the model failed to converge, the model was simplified to a model with random intercept and uncorrelated slope for that predictor. If that model too failed to converge, the estimate was acquired from the full model with only random intercepts for operator and participant session. *p*-Values for statistical significance of each predictor were acquired by comparing the full model with all predictors and random intercepts for operator and participant session, to the model without the predictor of interest using parametric bootstrapping (Halekoh & Højsgaard, 2014). The alpha level was set at 0.05.

As the intercept and estimates reflect log- or logit-transformed values, the values should be transformed back in order to be interpreted in their respective units (degrees or proportions). The predictors in the model were treatment-coded: An estimate for the accuracy of a new participant given a set of predictors can be acquired using Formula 1, where x is the assigned value in the model (see Table 1) for a predictor and b the estimate for that predictor. The intercept is noted as a.

$$\log(\text{accuracy}) = a + b_1 * x_1 + b_2 * x_2 + \ldots + b_n * x_n. \quad (1)$$

The value can subsequently be transformed back to degrees by taking the exponent.

RESULTS

Collected data

Before we turn to the statistical analyses, we first address the amount of data that was observed over the course of the experiment. In infant research, the amount of data collected in the beginning of an experiment is commonly larger than the amount of data collected at the end of the experiment, when more infants have dropped out due to, for example, inattention or sleepiness. We first examine the extent to which that is the case here. For 124 of 136 sessions, at least one observation for accuracy was acquired. Mean number of observations for accuracy (i.e., fixated val-

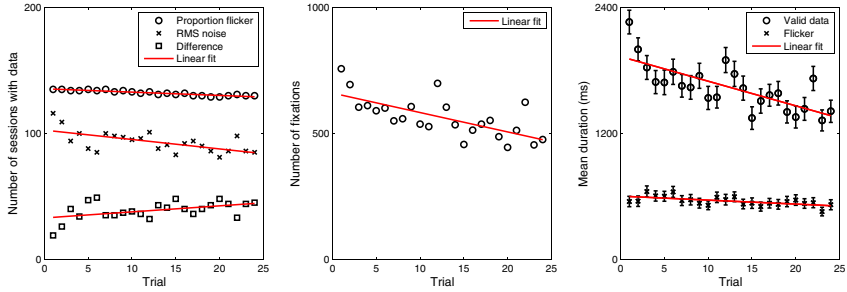

Figure 5 Number of data points and duration of data available as a function of trial. Left panel: number of sessions in which a data point was available for the proportion of flicker, RMS noise, and the difference between these two. Middle panel: number of fixations available per trial; each fixation is one estimate of precision in the linear mixed-effects model. Right panel: mean duration of valid data and mean duration of flicker (i.e., sum of periods of data loss < 100 ms) per trial. Error bars depict standard error of the mean.

idation targets) per session was 3.8 (SD = 1.5) of a maximum five. A total of 472 observations for accuracy were available.

For 134 of 136 sessions, at least one observation for precision was acquired. As visible from the left panel of Figure 5, the number of sessions for which at least one observation was acquired decreased as a function of trial number. The linear fit of number of sessions with at least one data point over trials decreased from 102 sessions at trial 1 to 85 sessions at trial 24. Mean total number of observations (i.e., fixations) per session was 101.0 (SD = 73.3). A total of 13,530 observations for precision were available. As visible in the middle panel of Figure 5, the number of observations also decreased as a function of trial number. The linear fit of number of observations over trials decreased from roughly 650 fixations for trial 1 to roughly 475 fixations for trial 24. This decrease in number of fixations as a function of trial was not due to fixation duration increasing as a function of trial.

For 135 of 136 sessions, at least one observation for flicker was acquired, and as visible from the left panel of Figure 5, the number of sessions for which flicker could be calculated remained fairly stable as a function of trial number. Mean number of observations per session for flicker was 23.3 (SD = 2.6) of a maximum of 24 trials. A total of 3,172 observations for flicker were available.

We will later discuss the implications of the changes in amount of data available over time.

TABLE 2
Statistical Results from the Linear Mixed-Effects Models for Offset, RMS Noise, and Proportion of Flicker

	Offset			*RMS noise*			*Proportion of flicker*		
Predictor	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*	*Estimate*	*SE*	*p*
Intercept	0.234	0.384	–	−0.543	0.166	–	−1.028	0.359	–
Nonblue eye color	−0.568	0.184	.003**	−0.497	0.083	.001**	−1.429	0.238	.001**
On parents lap	0.765	0.342	.022*	0.136	0.153	.727	0.504	0.438	.481
In high chair	0.780	0.329	.022*	−0.026	0.158	.727	0.354	0.427	.481
Number of calibrations	0.302	0.089	.006**	0.034	0.045	.245	0.231	0.109	.102
Movement	−0.156	0.091	.240	0.035	0.042	.421	−0.074	0.096	1.000
Time since fed	−0.071	0.065	.514	−0.034	0.042	.185	−0.107	0.106	.338
Time since awoken	0.015	0.065	1.000	−0.039	0.030	.211	−0.044	0.077	.376
Trial[†]	−0.025	0.041	.723	0.002	0.002	.005**	0.008	0.004	<.001***

[†]Note that there was a maximum of five trials for offset and 24 trials for RMS noise and proportion of flicker.
*$p < .05$, **$p < .01$, ***$p < .001$.

Linear mixed-effects models for offset, RMS noise, and proportion of flicker

Offset

As summarized in Table 2, infants with nonblue eyes produced significantly lower offsets than infants with bluish eyes. Moreover, infants seated either directly in the parents lap or in the high chair produced significantly higher offsets than did infants seated in the baby seat on the parents lap. Finally, as the number of calibrations increased, so did the offset for that particular recording session. The total movement during the recording as scored by the operator did not produce higher offsets. Neither did the time since the infant awoke, the time since the infant was fed, or the validation trial number—a measure for time since the start of the experiment.

RMS noise

As summarized in Table 2, infants with nonblue eyes produced significantly lower RMS noise than infants with bluish eyes did. Moreover, the RMS noise increased significantly as a function of trial number. The seating of the infant, the number of calibrations, the total movement during the experiment, and the time since the infant awoke of was fed did not predict RMS noise significantly better than chance.

Proportion of flicker

As summarized in Table 2, infants with nonblue eyes produced a significantly lower proportion of flicker than did infants with bluish eyes. In addition, the proportion of flicker increased over trials—a measure for time since the start of the experiment. No other predictors affected the proportion flicker.

For all models, correlations between fixed effects were insubstantial to small, except for the correlation between time since fed and time since awoken, which was moderate: between -0.35 and -0.36 depending on the specific model.

DISCUSSION

The purpose of the present study was to examine factors that might influence data quality in infant eye-tracking research. We investigated eye physiology, measurement characteristics, and infant behavior. Identifying the factors that influence data quality in infant eye-tracking research might

help researchers achieve higher data quality and increase the throughput of infants from data recording to data analysis. In addition, our findings might help manufacturers develop better eye-tracking systems for infant eye-tracking research. We showed that eye color, seating, number of calibrations, and trial number, significantly affect data quality measures.

We collected data quality measures for accuracy, precision, and data loss—according to our new definition of data loss, that is, flicker. First, there were a few observations that could be made from the amount of data available over time. While the number of sessions with at least one observation for precision decreased as a function of trial number, the number of sessions with an observation for flicker did not (see Figure 5). This indicates that the number of trials with at least some valid data, which allows calculation of the proportion of flicker, remains stable over the course of the experiment. The number of trials in which event detection can be achieved—which requires more valid data to detect periods of fixation—however, decreases over the course of the experiment. A possible explanation is that inattention to the screen increases as the experiment progresses, leaving little-to-no data available for event detection in some cases. Another possible explanation is that infants do look at the screen, but for some reason there is a decrease in stable tracking of the eyes over time. This may cause event detection to become impaired, as there are less consecutive periods of valid data left. Flicker, on the other hand, might still be calculated, as it does not require consecutive periods of valid data, only the presence of valid data. The two explanations are not mutually exclusive.

A first indication that the proportion of flicker per trial increases as a function of trial number comes from Figure 5. The right panel of Figure 5 shows a decrease of the mean summed duration of valid data (i.e., the sum of valid data periods per trial, averaged over participants) over trials. The mean summed duration of flicker (i.e., the sum of periods of data loss < 100 ms, averaged over participants) remains fairly stable over trial. As the proportion of flicker is defined as the duration of flicker divided by the duration of valid data, we can expect the proportion of flicker to increase over time.

We subsequently modeled whether participant eye physiology, seating, movement, contentedness, and measurement characteristics affected offset (measure for accuracy), RMS noise (measure for precision), and proportion of flicker (measure for data loss due to unstable tracking). We discuss the findings for each eye-tracking data quality measure separately and then summarize the results for data quality in general.

Accuracy: offset

The offset between a known target location and the fixated location by the infants served as our measure for accuracy. When the offset is higher, accuracy is lower, and we henceforth discuss the results in terms of accuracy. Accuracy was significantly higher for infants with nonblue eye color compared to infants with bluish eye color. The finding that bluish color results in lower accuracy has been previously reported by Kammerer (2009), and we extend this finding to infants. In addition, accuracy was significantly lower for participants seated directly in the lap of the parent or high chair compared to an infant positioned in a baby seat on the parents' lap. We should note, however, that the decision to place an infant directly in the parents' lap or a high chair was not made in advance, and a selection bias was thereby introduced. Not placing an infant in the baby seat on the parents' lap was only performed when either the parent indicated that a baby seat would not work, or when the operator determined that the infant would not relax in the baby seat. This means that a lower accuracy for infants in a high chair or directly in the parents' lap might have been a result of external factors. One explanation might be that the willingness of the infant to be restricted in their movement is lower compared to infants who would be positioned in the baby seat. Regardless, the infants who had to be positioned in a different type of seating compared to the baby seat produced lower accuracy. A tentative conclusion might be that a combination of positioning and infant willingness to be restricted in their movement may affect accuracy during recording. Another option may be that the seating outside the baby seat may have allowed infants to shift their position more after calibration. Earlier research has suggested that changing position after calibration affects accuracy in adult eye-tracking research (Cerrolaza, Villanueva, Villanueva, & Cabeza, 2012). Future research with randomized types of positioning will be necessary to substantiate such claims, however.

Finally, accuracy decreased as the number of calibrations before the start of the experiment increased. The intuitive explanation is that the operator gradually reduces the threshold for allowing a calibration. The quality of the final calibration would then be lower than what would be acceptable on a first run. Another explanation might be given if the reasons for re-calibration are considered. When re-calibration was necessary, this was often for one of three reasons. (1) Calibration data for one or more of the points were noisy.3 (2) Points were not looked at. (3) Data for the bottom or top points were not registered, while the points were actually being looked at. The latter reason usually indicated that the initial positioning of the infant was not adequate. Particularly the angle between

eye tracker and the infant had to be adjusted in order to be able to register the missing points. Future research will have to determine whether a gradually decreasing threshold for acceptance of calibration or an incorrect positioning before the calibration sequence produces lower accuracy. One possibility to consider is taking infants who would have to be repositioned out of the baby seat, and do a second positioning after a short break. This might minimize losing the infants' attention when substantial adjustments have to be made to the infants' position.

It is important to note that there are several limitations to the estimate of accuracy here. In adult eye-tracking research, the common method to estimate accuracy is to ask the participant to fixate a target location. The offset between the known target location and the fixated location by the participant then serves as the measure for accuracy. In infant eye-tracking research, however, we cannot instruct the participant to fixate a validation target. To estimate accuracy, it is therefore common to present an attractive stimulus on screen (e.g., a moving picture accompanied by sound), which the infant is assumed to fixate. Subsequently, the stimulus shrinks in size and the offset between the fixated location and the center of the validation target are taken as the estimate for accuracy. Whether the infant follows the shrinking of this validation target has thus far received little attention. In addition, it is difficult to determine which offsets generally occur in an eye-tracking study with infants. In the present study, most of the offsets we observed were below 2.5°, although a number of higher offsets occurred, some even larger than 10°. Note that at its maximum size, the validation target spanned 5.4° (which corresponds to 6.1 cm on screen). If an infant fixated the edge of the target and remained fixation as the target shrunk an offset of 2.7° would be recorded, that is when we assume the eye tracker is recording the infant's gaze with perfect accuracy. As the calibration stimuli were identical to the validation stimuli, the same problem holds for calibration. For the present analysis, we used log-transformed offset, which reduced the distance between median values and the extreme values observed in the present study. Most notably, these high offsets were recorded in infants seated directly in the parents' lap or high chair.

Precision: RMS noise

The RMS noise in all detected fixations served as estimates for precision. When RMS noise is high, precision is low, and we henceforth discuss the results in term of precision. We included all estimates based on all fixations (i.e., as opposed to the lowest or median value) as infant data may increase or decrease in noise over short periods of time. Precision was

significantly higher for infants with nonblue eye color compared to infants with bluish eye color. This extends the finding by Nyström et al. (2013) that blue-eyed participants produce noisier data, from adult participants to infants.

Moreover, precision decreased as a function of trial number: Precision was lower at the end of the experiment compared to the start. This supports an earlier report by Wass et al. (2014) that precision is higher at the start of the experiment than it is at the end of the experiment. The reason for decreased precision over time might, for instance, be due to changes in position in the head box since the beginning of the experiment or changes in head orientation.

Data loss: proportion of flicker

The proportion of flicker, calculated by dividing the sum of all periods of data loss shorter than 100 ms by the sum of all valid data in a trial, served as an estimate of data loss due to unstable tracking. A higher proportion of flicker means more data loss, and we henceforth discuss the results in terms of data loss. Data loss was significantly lower for infants with nonblue eye color compared to infants with bluish eye color. While the effect of lighter eye color on estimates of accuracy and precision has been described for adults (Kammerer, 2009; Nyström et al., 2013), and for infants here, no previous research has reported that lighter/bluish eye color produced more data loss. If we consider how decreases in precision and increases in data loss may occur, a possible explanation can be given. As has previously been suggested by Nyström et al. (2013), a possible explanation for blue eyes producing noisier data is that the blue iris is darker compared to a brown iris under infrared light, making it more difficult to distinguish from the pupil than a brownish iris. If detection of the pupil is unreliable, this may produce imprecise data, whereas a complete lack of detection produces data loss. If the detection of the eye and pupil is further impaired due to, for example, movement, one might observe more data loss as a result, instead of mere imprecision. Considering that infants are typically more difficult to restrain in terms of movement than adults, we might observe more data loss (in this case specifically flicker) for infants with bluish eye color compared to infants with nonblue eye color. In adults, this might not be the case as detection of the pupil is only slightly impaired due to the eye color, but does not yet become impossible at times altogether. If more data loss and higher noise is indeed the result of a lower contrast between pupil and iris, then using an eye tracker that illuminates the eye at a different angle might help. In so-called bright-pupil eye-tracking systems (Holmqvist et al., 2011, p. 25), the infrared light is

positioned in the same axis as the camera. The infrared light is reflected off the back of the retina, making the pupil appear bright in the camera image. When the pupil is bright and the iris dark under infrared light, it might be easier to determine the pupil-iris border compared to dark-pupil systems. In bright-pupil eye tracker, we therefore expect smaller differences in precision and data loss between eye colors.

Data loss also significantly increased as a function of trial number, which replicates previous reports with infants (Wass et al., 2014). As described before, the mean duration of valid data decreased over time, whereas the mean duration of flicker did not (see Figure 5). Data loss (as measure by the proportion of flicker), as a result, increases over time.

Summary of Eye-Tracking Data Quality Models

We report that for all three measures of data quality, infants with nonblue eye color produced data with higher quality compared to infants with bluish eyes. We hereby extend the finding that eye color influences data quality in adults (Kammerer, 2009; Nyström et al., 2013) to infants. As in the present study, both previous studies also used a dark-pupil eye tracker, where the contrast between a bluish iris and the pupil under infrared light is supposedly low (Nyström et al., 2013), thereby impairing data quality. This is an important finding when, for instance, testing-resources are limited. In this case, one might consider including mainly infants with darker eye colors. However, this is not a guarantee for adequate data quality, and moreover, the researcher should verify that eye color itself is not correlated with the outcome measure of interest. Moreover, both precision and stable tracking of the eyes are impaired as the experiment progresses, consistent with earlier reports (Wass et al., 2014). In addition, we provide possible explanations of how positioning may affect accuracy, although these remain speculative. What about the other predictors we investigated? While eye physiology and measurement characteristics affected data quality, it appeared as though infant contentedness and total movement during the experiment made little difference for the data quality. What might be the reason thereof?

Infant contentedness was operationalized using two measures: the time since the infant was last fed and last awoken. The reasoning was that infants who had not slept or eaten for a longer time were more likely to become fussy during the experiment. We found no effect of infant contentedness on all three data quality measurements. This finding is interesting for two reasons. Given that there was enough variability in the time since infants were last fed or last awoke, one option is that infant contentedness can be considered self-regulatory, where infants who are tired or hungry,

cry and are fed or fall asleep. In this manner, infants who start the experiment are neither sufficiently hungry nor sleepy, and unlikely to produce data with impaired data quality. A second possibility is that infant contentedness has, in fact, little-to-no effect on data quality, which is at least the case for the range that we have collected data in. Regardless of which might be the case, one should not necessarily have to expect impaired data quality due to hunger or sleepiness provided that parents and infants are left to feed and sleep at will.

An estimate for overall movement of the infant was obtained by having the operator score the percentage of time the infant moved during the experiment. This estimate of movement did not seem to affect any of the three data quality measures we investigated. Previous research, however, suggested that head movement—calculated from the velocity of eye position in the eye-tracker head box—was negatively correlated with precision and positively correlated with data loss (Wass et al., 2014). We consider several possibilities that might explain these discrepancies. First, it is quite possible that head velocity per trial predicts lower precision and more data loss (Wass et al., 2014), whereas a crude overall estimate of movement does not (c.f. the present study). An estimate for head velocity from the eye-tracker signal is objective in the sense that it is measured by a machine, whereas an estimate of total movement by the operator is subjective. It might very well be that the estimate by the eye tracker is a better predictor. In the present study, however, operator-coded movement was initially opted for as an estimate for overall movement that would be independent from the signal from the eye tracker. Independency from the eye tracker was particularly important as we were using it to predict another characteristic of the eye-tracker signal. Head velocity calculated from the eye position signal is not, however, independent from the eye tracker. In essence, using the head velocity signal from the eye tracker to predict changes calculated from the position signal in the eye tracker is potentially circular: The eye-tracker signal is used to predict the eye-tracker signal. If this is the case, then the head velocity signal should be affected by the same predictors as the measures for accuracy, precision, and data loss are.

To examine the effect of the predictors that affected data quality in the present study on head velocity, we constructed scatterplots for head velocity calculated as in Wass et al. (2014) against eye color, type of seat, and number of calibrations (see panels a–c Figure 6). As visible in panel a in Figure 6, infants with bluish eye color seemed to produce larger head velocity compared to infants with nonblue eye color. In addition, as visible from panel b in Figure 6, infants seated directly in the parents' lap or in the high chair appeared to have larger head velocities than infants placed in the baby seat. While the latter might not seem surprising—infants in a

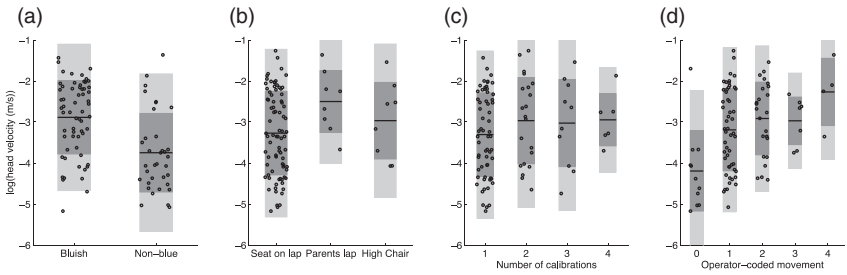

Figure 6 Raw data points for log-transformed head velocity (m/s). Each data point represents the participant mean of head velocities over all 24 visual search trials. Values are jittered with respect to the *x*-axis to avoid overlapping data points. Thick black lines indicate means, the dark gray area covers one standard deviation from the mean, and the light gray area covers two standard deviations from the mean. (a) Head velocity for infant with bluish and nonblue eye color. (b) Head velocity for different seats used during sessions. (c) Head velocity as a function of number of calibrations. (d) Head velocity as a function of operator-coded movement.

baby seat are strapped in and restricted in their movement—the finding that eye color predicts differences in head velocity did surprise us. While there is a possibility that all the infants with bluish eye color generally moved more in our sample than did infants with nonblue eye color, this appeared not to be the case if we compared operator-coded movement for the two groups.

We considered an alternative we deem more plausible. Bluish eye color has been suggested to be more difficult to track robustly due to that fact that the pupil is more difficult to detect in the infrared eye image where the bluish iris appears darker than a brownish iris. The head velocity is calculated from the position signal of the eyes in the head box, and if the same pupil detection method used to determine gaze position is used to determine eye position, it should suffer from the same difficulties. Bluish eye colors should therefore result in noisier head position signals calculated from the position of the eyes in the eye-tracker head box. Moreover, the differences between level means we observe in panels a–c in Figure 6 correspond to the direction of estimates we observe in the our statistical models: Infants with bluish eyes produced lower accuracy (higher offset), lower precision (higher RMS noise), and more data loss. In addition, infants seated in the parents' lap or high chair produced lower accuracy (higher offsets) than infants seated in the baby seat.

Operator-coded movement did show a positive relation with head velocity as calculated from the eye tracker (panel d in Figure 6), although

it is difficult to determine whether this only reflects general impaired tracking or also actual movement. While overall operator-coded movement did not seem to affect data quality, this estimate is not objective and is crude. It is therefore premature to conclude that overall movement has little effect on eye-tracking data quality. Future research using objective measures of head movement, independent from the eye-tracker signal, will be necessary to fully understand the effects of movement on data quality. It might very well be that overall movement in a 5-min period does not affect data quality, whereas short bursts of movement, which may not be reflected in our overall measure of movement, temporarily decrease data quality (see also Hessels et al., 2014). Head movement in future research might best be estimated using 3D accelerometers, EMG, or video coding. If the latter is opted for, a separate video system separate from the eye tracker needs to be implemented (particularly given that the Tobii TX300 does not provide a video signal), and particular care should be taken to ensure adequate video quality to code head movement. At present, caution is advised when head movement estimated from the eye-tracker signal is used to predict other aspects of the eye-tracker signal.

CONCLUSIONS, LIMITATIONS, AND FUTURE RESEARCH

In the present study, we have investigated the influence of eye physiology, positioning, measurement characteristics, and infant contentedness and behavior on eye-tracking data quality in infant research. We report that eye color influenced all measure of data quality, precision decreased, and data loss increased as a function of trial number. Furthermore, we presented tentative arguments on the effect of positioning on accuracy, although future research is needed to substantiate the present findings. Moreover, we report that infant contentedness during measurement does not seem to affect data quality. Finally, we reported that operator-coded total movement during the experiment did not affect data quality and highlighted the problem when using head movement estimates from the eye-tracker signal. These findings are valuable for infant eye-tracking researchers, both for the interpretation of data as well as for data collection. Researcher might use the present findings to increase throughput in their eye-tracking studies and optimize data collection. In addition, researchers should be aware of the differences in data quality that can arise due to factors such as eye color. If, for instance, two groups are being compared on measures that are affected by the level of data quality (see, e.g., Shic et al., 2008), it is important to balance factors that are known to affect data quality over these groups, for example, eye color.

Finally, manufacturers might use the present results to improve eye-tracking systems for use with infant participants.

There are also several limitations of the present study to consider. First, the eye-movement data presented here were recorded using a Tobii TX300. While this eye tracker is a common eye tracker for use in infant research, the question remains how the present findings generalize to other systems. In addition, we wonder how the present study extrapolates over different age groups. While some findings we report here are similar as in adults and are likely to generalize over children, other findings are, at this point, still specific to the age group reported here. Finally, we should note that the Tobii TX300 and the eye trackers used in previous data quality studies (Kammerer, 2009; Nyström et al., 2013; Wass et al., 2014) are dark-pupil eye trackers. When a different eye illumination angle is used, such as that in bright-pupil eye trackers, we expect data quality to be differently affected by the predictors we investigated. Particularly, we expect the effect of eye color on data quality measures to be reduced, as the plausible explanation of low contrast between iris and pupil does not hold for bright-pupil systems (Nyström et al., 2013). We encourage and welcome researchers to investigate eye-tracking data quality for other systems, as well as different age groups. Other limitations in the present study that have already been discussed above were the absence of random assignment of the positioning during the experiment and the lack of estimates of movement at a higher temporal resolution (i.e., using motion capture or 3D accelerometer techniques). We welcome future research into eye-tracking data quality using these techniques.

ACKNOWLEDGMENTS

The authors would like to thank all employees at the KinderKennisCentrum of Utrecht University for help with data collection. In addition, author RH would like to thank Kenneth Holmqvist and the Lund Humanities laboratory for their hospitality while working on this project. The study was financed through the Consortium on Individual Development (CID). CID is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO Grant Number 024.001.003 awarded to author CK). Author RA furthermore acknowledges support from the Swedish Research Council, Grant Number 437-2014-6735.

REFERENCES

Amso, D., Haas, S., & Markant, J. (2014). An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes. *PLoS One*, *9* (1), e85701.

Amso, D., & Johnson, S. P. (2006). Learning by selection: Visual search and object perception in young infants. *Developmental Psychology*, *42*, 1236–1245.

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53.

Aslin, R. N. (2012). Infant eyes: A window on cognitive development. *Infancy*, *17*(1), 126–140.

Aslin, R. N., & McMurray, B. (2004). Automated corneal-reflection eye tracking in infancy: Methodological developments and applications to cognition. *Infancy*, *6*(2), 155–163.

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*(4), 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1. 1–7. Retrieved from http://CRAN.R-project.org/package = lme4.

Blignaut, P., & Wium, D. (2013). Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior Research Methods*, *46*(1), 67–80.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436.

Cassia, V. M., Turati, C., & Simion, F. (2004). Can a nonspecific bias toward top-heavy patterns explain newborns' face preference? *Psychological Science*, *15*(6), 379–383.

Cerrolaza, J. J., Villanueva, A., Villanueva, M., & Cabeza, R. (2012). Error characterization and compensation in eye tracking systems. Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12.

Chawarska, K., & Shic, F. (2009). Looking but not seeing: Atypical visual scanning and recognition of faces in 2 and 4-year-old children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *39*, 1663–1672.

Frank, M. C., Vul, E., & Johnson, S. P. (2009). Development of infants' attention to faces during the first year. *Cognition*, *110*(2), 160–170.

Frank, M. C., Vul, E., & Saxe, R. (2011). Measuring the development of social attention using free-viewing. *Infancy*, *17*(4), 355–375.

Gredebäck, G., Fikke, L., & Melinder, A. (2010). The development of joint visual attention: a longitudinal study of gaze following during interactions with mothers and strangers. *Developmental Science*, *13*, 839–848.

Gredebäck, G., Johnson, S., & von Hofsten, C. (2009). Eye tracking in infancy research. *Developmental Neuropsychology*, *35*(1), 1–19.

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models – The R package pbkrtest. *Journal of Statistical Software*, *59*, 1–32.

Hessels, R. S., Cornelissen, T. H. W., Kemner, C., & Hooge, I. T. C. (2014). Qualitative tests of remote eyetracker recovery and performance during head rotation. *Behavior Research Methods*. doi:10.3758/s13428-014-0507-6.

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

Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye tracker data quality: What it is and how to measure it. Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12, 45. doi:10.1145/2168556.2168563

Hunnius, S., de Wit, T. C. J., Vrins, S., & von Hofsten, C. (2011). Facing threat: Infants' and adults' visual scanning of faces with neutral, happy, sad, angry, and fearful emotional expressions. *Cognition & Emotion*, *25*(2), 193–205.

Johnson, S. P., Amso, D., & Slemmer, J. A. (2003). Development of object concepts in infancy: Evidence for early learning in an eye-tracking paradigm. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(18), 10568–10573.

Jones, W., Carr, K., & Klin, A. (2008). Absence of preferential looking to the eyes of approaching adults predicts level of social disability in 2-year-old toddlers with autism spectrum disorder. *Archives of General Psychiatry*, *65*(8), 946–954.

Jones, W., & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, *504*, 427–431.

Kammerer, Y. (2009). How to overcome the inaccuracy of fixation data? The development and evaluation of an offset correction algorithm (pp. 1–28). Presented at the Scandinavian Workshop on Applied Eye Tracking 2009.

Leppänen, J. M., Moulson, M. C., Vogel-Farley, V. K., & Nelson, C. A. (2007). An ERP study of emotional face processing in the adult and infant brain. *Child Development*, *78*(1), 232–245.

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.

Oakes, L. M. (2010). Infancy guidelines for publishing eye-tracking data. *Infancy*, *15*(1), 1–5.

Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy*, *17*(1), 1–8.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Shic, F., Chawarska, K., & Scassellati, B. (2008). The amorphous fixation measure revisited: with applications to autism. 30th Annual Meeting of the Cognitive Science Society, Washington, DC.

Shic, F., Macari, S., & Chawarska, K. (2013). Speech disturbs face scanning in 6-month-old infants who develop autism spectrum disorder. *Biological Psychiatry*, *75*(3), 231–237.

Wass, S. V., Forssman, L., & Leppänen, J. (2014). Robustness and precision: How data quality may influence key dependent variables in infant eye-tracker analyses. *Infancy*, *19*(5), 427–460.

Wass, S. V., & Smith, T. J. (2014). Individual differences in infant oculomotor behavior during the viewing of complex naturalistic scenes. *Infancy*, *19*(4), 352–384.

Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*(1), 229–250.

Wheeler, A., Anzures, G., Quinn, P. C., Pascalis, O., Omrin, D. S., & Lee, K. (2011). Caucasian infants scan own- and other-race faces differently. *PLoS One*, *6*(4), e18621.