The Effects of Calibration Target, Screen Location, and Movement Type on Infant Eye-Tracking Data Quality

Karola Schlegelmilch[1] and Annie E. Wertz[1]

Author Affiliations:

[1] Max Planck Institute for Human Development, Max Planck Research Group Naturalistic Social Cognition, Lentzeallee 94, 14195 Berlin, Germany.

Correspondence:

Karola Schlegelmilch, Max Planck Institute for Human Development, 94 Lentzeallee, 14195 Berlin, Germany. Phone: +49 (030) 82406-295; Email: schlegelmilch@mpib-berlin.mpg.de

Keywords:

Eye-tracking, data quality, EyeLink 1000 Plus, infant research methods, calibration

Acknowledgements:

We thank our participants and their parents, J. Eichelsdörfer, D. Erdt, Jann Wäscher and the members of the Max Planck Research Group *Naturalistic Social Cognition* for their assistance. We also thank Sam Hutton for supporting information related to the eye-tracking system. During this research, Karola Schlegelmilch was a pre-doctoral fellow of the International Max Planck Research School on the Life Course (LIFE, www.imprs-life.mpg.de). This research was funded by the Max Planck Society. The authors declare no conflicts of interest with regard to the funding source for this study.

Abstract

During infant eye-tracking, fussiness caused by the repetition of calibration stimuli and body movements during testing are frequent constraints on measurement quality. Here, we systematically investigated these constraints with infants and adults using EyeLink 1000 Plus. We compared looking time and dispersion of gaze points elicited by stimuli resembling commonly used calibration animations. The adult group additionally performed body movements during gaze recording that were equivalent to movements infants spontaneously produce during testing. In our results, infants' preference for a particular calibration target did not predict data quality elicited by that stimulus, but targets exhibiting the strongest contrasts in their center or targets with globally distributed complexity resulted in the highest accuracy. Our gaze measures from the adult movement tasks were differentially affected by the type of movement as well as the location where the target appeared on the screen. These heterogeneous effects of movement on measures should be taken into account when planning infant eye-tracking experiments. Additionally, to improve data quality, infants' tolerance for repeated calibrations can be facilitated by alternating between precise calibration targets.

Many insights into infant development are based on the study of gaze behavior. Eye-tracking technology allows an increasingly more detailed analysis of infant gaze behavior and is used to investigate a wide range of phenomena, such as categorization, object and face perception, and social cognition (for reviews see e.g., Aslin, 2007; Gredebäck, Johnson, & von Hofsten, 2009; Oakes, 2012). While the availability of high temporal and spatial measuring resolution expands the possible experimental designs and dependent measures, typical problems that might occur during infant eye-tracking can markedly effect data quality. Therefore, researchers must remain cautious to avoid overestimating its measurement accuracy (Aslin, 2012) and continue to address the inherent challenges of infant eye-tracking (Oakes, 2012).

The major challenges are body movements or inadequate looking behavior during calibration and during the later stages of the experiment. Haith (2004) estimated that an average of 50% of infants recruited for eye-tracking studies did not provide usable data as a result of such failures. In cases where individual infants are not fully excluded from the datasets, rejected trials of otherwise acceptable individual performance increase the proportions of unusable data (for procedures to reduce data loss in post hoc data optimization, see Leppänen, Forssman, Kaatiala, Yrttiaho, & Wass, 2015, for Tobii systems; Renswoude et al., 2018, for EyeLink technology).

A comparison of data quality in infant eye-tracking based on exclusion rates alone is difficult because exclusion criteria are adjusted according to the sensitivity of the phenomena under investigation. For example, psychophysical investigations that are sensitive to stability of gaze might be particularly prone to confounds related to differences in body movement, making more conservative exclusion boundary values necessary (e.g., an average calibration error of <1° or a data yield >80%; Alahyane et al., 2016). Infant studies that include data from adult participants often also employ more conservative exclusion boundaries to facilitate comparisons across differentially behaving participant groups (e.g., a data yield >80%;

Morgante, Zolfaghari & Johnson, 2012). Similarly, studies that assess infants' attention to the details of an image depend on high spatial accuracy to produce interpretable results (e.g., Constantino et al., 2017). In contrast, studies that assess attention to larger visual targets that are clearly separated in the visual field can achieve valid data in spite of higher calibration errors or lower proportions of recorded gaze (e.g., Kulke, Atkinson, & Braddick, 2015; LoBue, Buss, Taber-Thomas, & Pérez-Edgar, 2017). Despite the diverse demands of different experimental paradigms on data resolution, all approaches to infant eye-tracking would benefit from the following: (1) infant participants who are more attentive throughout the experimental session, and (2) enhanced measurement accuracy.

The present study therefore targets the most common pitfalls of infant eye-tracking: the calibration procedure and body movement during remote mode recording. We compared several animated calibration targets for their attractiveness to infants and their ability to direct infants' gaze to their centers. Enhancing the calibration stimuli and procedures used during this essential part of data collection will lead to more reliable recordings. In addition, we systematically investigated the ways in which body and head movements affect the accuracy of gaze recordings. More knowledge about the impact of these factors can help elucidate the best steps to take during and after data recording and adapt experimental procedures accordingly.

Infant Calibration Targets

The accuracy of infant eye-tracking data relies to a large extent on calibration quality (Gredebäck et al., 2009; Oakes, 2012). In standard adult calibration procedures, adults are explicitly instructed to fixate 5 to 13 point-like visual targets as precisely as possible. Infants of course cannot be instructed in this way. Instead, infants' spontaneous attention needs to be captured and held by animated calibration targets. Further, infants are commonly expected to perform calibrations with only 5 to 6 targets because of their limited attention span (Gredebäck et al., 2009). Inattentiveness of an infant during calibration makes repetitions of

this procedure necessary, which can lead to annoyance and further inattentiveness. Animations that facilitate infants' attention and result in bundled fixations during calibration should therefore produce more reliable data. Indeed, the design of calibration targets has an impact on fixation stability even for adults, who voluntarily try to keep their gaze still (Thaler, Schütz, Goodale, & Gegenfurtner, 2013).

Determining which features facilitate calibration in infancy is a difficult task. Visual acuity relating to spatial frequency and contrast are not yet as developed in infancy as in adulthood, making less detailed stimuli easier for infants to process. However, patterns that are easy for infants to perceive can become boring when presented too frequently. A family of commonly applied calibration targets therefore consists of looming concentric spheres or rings, which are expected to provoke central fixations. Because concentric forms are not processed in an adult-like way until adolescence (Doucet, Gosselin, Lassonde, Guillemot, & Lepore, 2005), it is not yet clear how this processing difficulty interacts with infants' attention, especially if the target is additionally flashed, moved, or its contour density is intensified to increase salience (Aslin & Smith, 1988; Zihl & Dutton, 2015). There is reason to suspect that the combination of these features may be problematic because visual patterns that are too stimulating can cause the infant to turn away (Bornstein & Benasich, 1986). Nevertheless, calibration targets must have features that make them sufficiently noticeable when appearing at unexpected locations on the screen because the area covered by the visual field is still increasing during infancy.

Inter-individual variability in the development of the fundamental issues we have raised makes it difficult to rely on theoretical assumptions alone when predicting the impact of calibration targets on infants' gaze behavior. Therefore, a systematic experimental investigation of the applicability and impact on data quality of calibration targets with different features is necessary.

Infant Eye-tracking Accuracy

Several factors that generally lead to a reduction of data quality during eye-tracking are present in infant eye-tracking experiments: movement, sitting position, geometry of the set-up, and the operators' experience with calibration procedures (for an extended discussion of these factors see Holmqvist, Nyström, & Mulvey, 2012). Movement during the recording sequence is particularly challenging because it causes changes in the geometry on which the calibration was based. In addition, the pupils might become partially covered, or move out of the area observable by the eye-tracker's camera, resulting in less robust data recording. Common dependent variables like the number of fixations or response time latencies are systematically influenced by interruptions of contact to the eye-tracking camera (Wass, Smith, & Johnson, 2013).

The circumstances of the infant eye-tracking situation make a more tolerant procedure necessary. Infants sitting on the lap of their caregiver can be expected to move in all spatial dimensions, even if they are interested in the experiment. Although some laboratories successfully use infant seats in eye-tracking studies for certain age groups (e.g., Saez de Urabain, Nuthmann, Johnson, & Smith, 2017), constrictions of movement can be uncomfortable and distracting for infants. Therefore, researchers must account for deviations from a stable position during infant testing. Remote mode eye tracking comes with a moderate spatial tolerance to account for such instability. Some systems also provide the ability to do drift checks to assess whether the measured gaze points have shifted during trial sequences (e.g., EyeLink 1000 Plus). If the reported fixation error is too large, a recalibration procedure should be implemented. A single drift check measurement might not be sufficient if the moment to accept the fixation was poorly chosen or if the infant's saccade towards the validation target was not precise. If the indicated gaze positions on the eye-tracking monitor or on a visual data output give the impression that fixations are systematically displaced, some eye-tracking software offers the possibility to adjust them later during analysis by

carefully shifting them to their assumed correct locations (e.g., EyeLink Data Viewer User's Manual, 2002-2015), and researchers have developed procedures for post hoc corrections as well (e.g. Frank, Vul, & Saxe, 2012).

The success of all these factors—the tolerance of the eye-tracking device, drift checks, or subsequent corrections—depend on understanding the effects of movement on the data. The algorithms of the eye-tracker that correct head movements in remote mode might not function properly if participants move too much (Hessels, Cornelissen, Kemner, & Hooge, 2015b; Niehorster, Cornelissen, Holmqvist, Hooge, & Hessels, 2017). Additionally, movement might result in blurred camera images leading to noise and a different variance of gaze points (Holmquist et al., 2012; Wass et al., 2014) and changes in the angle of the participant's head in relation to light sources might affect accuracy (Wass, Smith, & Johnson, 2013). Previous investigations of infant eye-tracking described reduced precision as a function of trial number(Hessels, Andersson, Hooge, Nyström, & Kemner, 2015a), and high unpredictability of the magnitude or angular direction of inaccurate fixation measurement (Morgante et al., 2012). Therefore, more precise insights into the effects of unstable sitting positions on gaze data are needed.

The Current Study

We compared the impact of different factors on the eye-tracking data quality of infant (8- to 12-month-olds) and adult participants. Our goals in the current study were twofold. First, we compared several different calibration targets for their impact on infants' attention and their ability to guide infants' gaze to their centers. Some of the animated calibration targets we tested were already in regular use in laboratories conducting infant eye-tracking experiments, while two additional novel calibration targets were developed for this study based on the sensitivity of the early visual system and infant perceptual abilities. Second, we systematically assessed effects of certain types of head and body movements during the

recording session by asking adults participants to perform movements similar to those typically made by infant participants during fixation sequences.

To our knowledge, this is the first investigation to address attention to different calibration stimuli with infants. The study was conducted in remote mode with the eye-tracking system EyeLink 1000 Plus (SR Research Ltd. 2015). The EyeLink system has been predominantly used with adult participants. Its high sampling rate could enhance the detection of inadequate gaze shifts, but be less robust to unrestricted movement and cause measurement artifacts (Niehorster et al., 2017). Investigations of accuracy and precision with infants were thus far conducted with Tobii eye-tracking technology (Hessels et al., 2015a; Morgante et al., 2012; Wass et al., 2013; Wass et al., 2014). The Tobii system assesses fixations on dispersal based algorithms instead of the velocity based algorithm of the EyeLink system, and data quality or dependent variables may be affected in a different manner if another technical system is used (Hessels et al., 2015b). Moreover, the Tobii system uses different calibration procedures that allow missing calibration points and graphically indicate gaze distance to the calibrated target (Tobii Studio User's Manual, 2016; for a discussion of the procedure see Morgante et al., 2012). In spite of the differences between eye-tracking systems, our investigation of the effects of different calibration targets and movement types on accuracy using EyeLink technology will provide valuable insights for infant eye-tracking studies using other technical systems.

Method

Participants

The present study was conducted according to guidelines laid down in the Declaration of Helsinki, with written informed consent obtained from a parent or guardian for each child before any assessment or data collection. All procedures involving human subjects in this study were approved by the Ethics Committee of the Max Planck Institute for Human

Development. The final sample of infant participants recruited from urban and suburban regions of a large European city were 29 healthy, full term infants (age: *M* = 10 months, 8 days, range = 8 months, 0 days to 12 months, 13 days; 14 female). All infants had normal vision without correction. An additional four infants were recruited but excluded from the final sample because they could not be calibrated due to excessive movement (2 infants), or their eyes were not detected by the eye-tracker (2 infants). We did not assess eye color because it was outside of the scope of the present investigation (for a discussion of eye color affecting infant eye-tracking data quality, see Hessels et al., 2015a). The adult sample consisted of 25 participants (age: *M* = 24.9, *SD* = 3.96, *range* = 19 – 34 years; 11 female). All adult participants had normal vision without correction and all adult participants were included in the analysis. Our infant and adult sample sizes were chosen based on those used in similar investigations (e.g., Dalrymple et al., 2018; Morgante et al., 2012; Wass et al., 2013) and to be within the recruiting capabilities of a wide range of infant labs. All participants were recruited from participant databases and tested in the Max Planck Institute for Human Development, Berlin, Germany. Both participant groups received 10 Euros and infants additionally received a participation certificate.

Stimuli

The six calibration targets we tested were animated geometric forms (see Figure 1a.). The calibration targets we focused on included: a.) differing concentric forms (spiral, star-like, or circular), b.) blurred contours vs. equally distributed contrasts, and c.) different types of motion around a center (twisting, looming or blinking). We focused on abstract symmetrical forms because stimuli that resembled naturalistic figures (e.g., faces, ducks) were expected to guide infants' gaze to non-central areas of interest (e.g., eyes and mouth of a face, head or tail of an animal). Symmetrical forms equally surround the target's center so that attention is not drawn by irregularities of the silhouette. We therefore sought to compare the gaze elicited by different types of symmetrical forms, some with blurred contours at the outer

edges and some without. All of our targets also exhibited some movement to attract infants' attention. The zooming in and out motion gives the impression that the targets are looming towards the participant and receding again. In addition, spirals provide concentric movement effects when they twist. Due to the limitations of infants' attention, we did not parametrically vary all possible feature and movement combinations. Instead, we investigated whether combinations of graphical forms and movement would elicit more central attention.

Contrast and size values were chosen to fit the visual capability of the infant age group (Aslin & Smith, 1988). The calibration targets expanded to a maximum diameter of up to 5° visual angle, and shrank to minimal diameters of between 2.5° and 0.5°, depending on their specific design and behavior. All calibration targets were accompanied by sounds corresponding to their looming and twisting behavior. Video examples of the calibration targets are provided online (https://osf.io/3k8jp/?view_only=e8075dc7bf0e4ab780c5e620b8f4860f). The calibration targets used for the initial calibration procedure were presented on a grey background, while repetitions for validation or as part of the trial sequences were presented on different monochromatic backgrounds of the same luminance level as the grey (see Figure 1b and the section *infant experimental design* for further descriptions).

The part of the experiment that was exclusively performed by adults (see Movement Task section below) used the 13 point calibration procedures provided by the manufacturer (SR Research Ltd. 2015). The stimuli that were used during the trial sequences of the adult movement block consisted of small filled circles (Ø = 0.5°) with a crosshair centered on it and a thin blurred circle surrounding the center at Ø = 3° to facilitate peripheral detection (see Figure S1). They appeared at 9 different screen locations (see Figure 7a and S2) in randomized order.

{ Place Figure 1 here }

Apparatus

An EyeLink 1000 Plus (SR Research Ltd. 2013 - 2015) eye-tracking system was installed on a host PC with 32bit operating system Intel(R) Core(TM)^2 Duo processor with 2.80GHz and 2Gb Ram. Gaze was recorded using an EyeLink 1000 Plus High-speed Camera with a 16 mm / 1:14 lens and an CL Illuminator TT890. Monocular gaze position was recorded without head stabilization in remote mode. The device has a recording accuracy of 0.25° - 0.5° and a precision (RMS) of < .05 visual angle, as specified by the manufacturer. Pupil and corneal reflection was assessed in a sampling rate of 500 Hz. A target sticker was placed on participants' faces (cheek or forehead) and the camera of the eye-tracker was placed approximately 60 cm in front of the target sticker as recommended by the manufacturer (the possible range is 40 cm - 70 cm for remote mode tracking; EyeLink, 2015). The presentation monitor (Samsung UE50H6470SS, 80 cm by 63 cm, 50" display, with 1280 by 1024 pixel resolution, and 400Hz CMR refresh rate) was set at a distance of 140 cm away from the participants' eyes to approximately fit the trackable area of 32° by 26° visual angle in accordance to the manufacturers suggestion.

Procedure for Infant Experiment

Infants were seated on their caregiver's lap with a small bullseye sticker placed on their forehead that was recognized by the eye-tracking camera. Parents were reminded to sit quietly and not direct their infant's attention during the experiment. Corneal reflection and contrast sensitivity of the eye-tracker were adjusted while an introductory animation clip was shown. The room was dimmed and the eye-tracking device was operated quietly from behind a curtain. The presentation could last up to 9 minutes maximum, but was terminated early if the infant showed fatigue, did not attend to the screen anymore, or if the caregiver requested to end the session.

Experimental design. The infant experiment consisted of six trial sequences. Each trial sequence started with a five point calibration using one of the six calibration targets (see Figure 1a). A different calibration target was used for this initial calibration before each of the six trial sequences; the order of the trial sequences was randomized across participants. Calibration success was determined by evaluating the symmetry of the pattern of gaze points shown on the eye-tracking monitor after the infant had attended to all five target locations. Following the instructions provided by the manufacturer, these gaze point locations were of equal distance to each other (EyeLink, 2015). If gaze points were registered at less than five locations, the calibration procedure was not accepted by the eye-tracker and needed to be repeated. We stopped the experiment if three calibration attempts were unsuccessful.

After successful calibration, if the infant still seemed interested in the screen, a five point validation was performed with the same calibration target on a differently colored background. This was done to tentatively assess calibration success during infant eye-tracking, similar to how it is commonly done during adult eye-tracking. If the infant lost interest and started to move during this validation procedure, the validation was stopped immediately and the trial sequence (see below) was initiated so that the accuracy of the calibration would not be impaired through intermediate movement. If the infant already began fidgeting during calibration, the experimenter skipped the validation entirely and went directly on to the trial sequence. After the initial calibration procedure, infants were shown three types of trials in the trial sequence (Example videos for the three trial types are provided online (https://osf.io/3k8jp/?view_only=e8075dc7bf0e4ab780c5e620b8f4860f):

a.) *Preference* trials*:* These trials examined infants' preference for looking at the six different calibration targets (see Figure 1a) when they were presented simultaneously on the screen. To do this, the different calibration targets were shown four at a time, evenly spaced in four quadrants of the screen (see S3 and Figure S2). Infants were shown three different combinations of four calibration targets during one trial, such that each of the six calibration

targets appeared twice. The stimuli were shown at the same four screen locations for each of the combinations. Each combination was shown for 8s, resulting in a 24s total duration for the trial. The Preference trials were accompanied by music and occurred only once in each trial sequence.

b.) *Verification* trials*:* In these trials, we assessed the accuracy and precision of infants' gaze elicited by each calibration target (see Figure 1a). A calibration target was presented in parallel at three of the five screen locations used in the initial calibration procedure; the configurations across the five possible locations were randomly selected out of several potential combinations and varied across trials to avoid confounds from particular screen locations (see Figure S3). The calibration target used in each Verification trial was always different from the target used for the initial calibration procedure. A Verification trial lasted for 12s and was accompanied by one of two rhythmic Marimba sounds. The parallel and synchronous movement of the three identical calibration targets was intended to maintain infants' interest during these trials while their gaze to each of the targets was recorded. Three verification trials occurred in each trial sequence with alternating calibration targets.

c.) *Spread* trials: Here, we compared the accuracy of gaze elicited by variants of the six calibration targets (see Figure 1b) during the time course of a trial. We created variants of the calibration targets for these trials in order to understand which visual attributes elicit more accurate gaze (see Figure 1 for a precise description of the modifications). A single target was presented at central location on the screen and loomed from a size of 1° to 17° peaking at 2s, and decreased back to 1° until the trial terminated 6s later. In these trials, the target variants were shown one at a time. Three Spread trials occurred in a trial sequence; each Spread trial showed a different target variant.

Taken together, there were seven trials in each trial sequence (1 *Preference* trial, 3 *Verification* trials, and 3 *Spread* trials) that were shown in randomized order within each of the six trial sequences. Moreover, we randomized the order of the six trial sequences across

participants. Finally, two versions of the experiment were alternated to balance the combinations of targets used for the initial calibration procedure and the targets shown in the trial sequence.

Procedure for Adult Experiment

For the adult participants, the same eye-tracker setup was used as with the infants. The adult version of the study lasted approximately 30 minutes. At several pre-defined time points during the experiment, participants were offered a short break.

Experimental design. The adult version of the experiment consisted of four blocks. The first block was a sequence of practice trials consisting of instructions and examples of the respective trials. During this first block, adults were instructed to view the target videos played during the Preference, Verification, and Spread trials freely while keeping their head and body in a central and stable position. Adults were informed that during the movement tasks (see below), they would be asked to perform certain movements at predetermined points in the trials, and that the type of movement would be indicated on the screen. Adults were instructed to look at the targets that appeared during these trials as precisely as possible during or after performing the respective body movements (described in detail below). If necessary, the instructions were explained orally. Adults were also asked to practice the body movements described on the screen with the guidance of the experimenter.

The second block of the experiment was almost identical to the infant version described above, including the five point calibration, except that adults performed two fewer *Preference* trials to reduce the total testing time. The third block investigated the effect of head and body movements on data quality and was unique to the adult version of the experiment. It began with a 13 point calibration followed by four movement sequences. Each sequence started with instruction slides. Participants were asked to perform movement tasks while a static target appeared at one of nine locations distributed grid-like over the screen (see Figure 7a and S2 for details). The target was a small filled circle (0.5°) in front of a cross-hair

pattern (3°); the same target was used throughout the movement sequence. The distance between the target locations was approximately 9° in the horizontal and vertical dimension. The target was presented for 1s at a location, with inter stimulus intervals of 1s. The movement tasks adults were asked to perform were:

a.) *Fix*: Keep their head still and focus on the targets as precisely as possible by only moving their eyes (control condition).

b.) *Head Movement*: Focus on the targets as precisely as possible with the direction of their head following the direction of their eyes. This task mimicked infants' tendency to follow visual stimuli with their head as well as their eyes.

c.) *Side Movement*: Turn their head and upper body out of the area tracked by the eye-tracking camera in the direction indicated by arrows, and then directly return to the central position to fixate precisely on the following targets until the next directional arrow was shown. The arrows appeared three times during the task, pointing to the left, to the right, and upwards. With this movement task, we assessed data quality after the eye-tracking camera had to deal with fast movement and loss of the eyes and the bullseye sticker, as frequently occurs when infants look away from the screen.

d.) *Bend Movement*: Bend about 10 cm (4 inches) forward towards the monitor and stay in this position while directing their gaze on the subsequent visual targets as precisely as possible. Changes in the distance towards the screen are another common occurrence during infant eye-tracking.

The movement sequences consisted of 27 trials.

The final block was the *Calibration-Repetition* block which was also unique to the adult version of the experiment. This block began with another 13 point calibration, then all six calibration targets (see Figure 1a) were repeated one at a time in random order at five screen locations identical to those during the five point calibration procedure used with

infants. *Calibration-Repetition* was intended to compare our two accuracy measures *Displacement* and *Instability* (described below) for each of the targets. As in the first block, participants were asked to direct their gaze towards the stimuli in a way that reflected their natural interest (free viewing), but not to move their head or body during this part of the experiment. The stimuli were shown on a grey background with their original sound for 6s each.

Data Preparation

Trials were excluded from analysis if the recorded gaze proportion was below 50% of the full trial duration (infants $N = 88$; adults $N = 9$). This exclusion criterion, which may seem liberal for studies comparing infants with adults (see e.g., Morgante et al., 2012), was set because variance in data quality was necessary for the analysis. In addition, if single calibrations during the experiment could not be performed satisfactorily because of temporary movement of the participant (infants $N = 3$) or because of technical problems (infants $N = 1$; adults $N = 6$), that particular trial sequence was excluded.

For saccade detection, a velocity based algorithm was used, with thresholds of velocity 30°/sec, acceleration 8000°/sec^2, and motion 0.1°, and a heuristic filter was applied to reduce velocity noise in favor of saccade detection, as implemented by the manufacturer. Gaze was defined as fixation if it was not recognized as saccade or blink. We used these preinstalled settings because they are the most commonly used criteria and because every change in the thresholds will affect the outcomes (Holmqvist et al., 2011) and would reduce the generalizability of our results. Fixations that were shorter than 50ms, which is one of the post-recording thresholds of the EyeLink software, remained in the analysis because they were considered an indicator of reduced data quality.

We assessed the participants' head distance change after calibration. This was done by subtracting the head camera distance at the moment the calibration was accepted from all other data points of the trial sequence. This measure allowed us to estimate the amount of

movement for each participant. The EyeLink 1000 Plus data output provides the distance between the eye tracking camera and the bullseye sticker on the participant's head in millimeters. Note that this measure does not indicate the exact direction of movement[1].

To assess the proportion of recorded gaze, all samples with gaze data were divided by the total number of possible samples during a trial. For inferences about data quality, only points of gaze (POG) within a fixation were used. To further exclude POGs that most likely were not related to a distinct task, areas of interest (AOI) and periods of interest (POI) were defined. The AOIs covered the calibration target and a radial space around it large enough to include misplaced POGs due to inaccurate measurement, but small enough to exclude gaze that was directed at the screen for other reasons, such as gaze at the empty screen center, or intermittent fixations. The POIs began from the first moment when participants' visual attention was directed at one of the targets during our trial sequences. We defined this moment as the first time point when the average of all participants' fixation positions was inside the AOI of the specific trial. The POIs ended when less than the average of all participants' fixation positions were inside the AOI. The POIs excluded orienting and anticipatory fixations at the beginning of a trial. Because POIs were contingent on the AOI of the specific trial, the starting and ending points of POIs differed between the trial types (see Table S1 for a precise description of the AOIs and POIs).

[1] EyeLink 1000 Plus also provides coordinates for sideways or vertical movements, but their units are not clearly defined. EyeLink notes that all values indicating head movement in the data output "are intended for a qualitative indication of subject head position in the camera coordinate. If you need quantitative data output for the head movements and rotation angle, you will need an independent head tracker" (EyeLink Data Viewer User's Manual, 2015, p. 131).

Dependent Variables: Precision and Accuracy Measures

For our study we defined precision in line with Holmqvist et al. (2011) as the ability of the eye-tracker to reproduce a measurement, and spatial accuracy as the offset between the expected and the recorded gaze position. We assessed precision in two ways: first as a root mean square inter-sample distance of POGs (termed *RMS*, Holmqvist et al., 2012) and second as the distance between POG coordinates and their centroid during a fixation, divided by the amount of included POGs (termed *Dispersion*; Komogortsev, Jayarathna, Koh, & Gowda, 2010). Higher values of both precision measures indicate lower precision. During infancy, gaze points during a fixation cover a larger area than during adulthood (Luna, Velanova, & Geier, 2008; Zihl & Dutton, 2015), which must be kept in mind when precision is based on distances between POGs. Nevertheless, impaired precision can affect the proportional looking time to AOIs (Wass et al. 2014).

Accuracy was calculated in two ways as well. For trials following 13 point calibrations during the adult experiment, spatial accuracy of a fixation was assessed as the mean Euclidian distance between all fixational POGs and the stimuli center (termed *Displacement*). In the part of the experiment that was performed by infants and adults and that used animated calibration targets, accuracy was scored differently in order to separate calibration related displacements from gaze spread elicited by the stimuli. We calculated the Euclidean distance between all fixational gaze points occurring during the POI of a trial and their centroid. This score provides an estimate of the spatial spread or density of fixations (termed *Instability*)[2]. Displacement and Instability address distinct characteristics of accuracy. In contrast to Instability, Displacement does not distinguish between fixations that are close together and others that are wide spread if they have a similar distance to the target's center; therefore the

[2]Note that Gredebäck et al. (2009) used the same measure but termed as RMS.

two measures might lead to diverging values. To validate the use of Instability as a measure of accuracy, we compared both accuracy measures in the adult Calibration-Repetition task.

The units of all gaze related measures are degrees of visual angle.

Results

Statistical Analysis

In the part of the study that was performed by both infants and adults, infants successfully completed 761 trials (M_{infant} = 26.2 trials per participant, SD = 10.7, *min* = 8, *max* = 42) and adults completed 976 trials (M_{adult} = 39.4 trials per participant, SD = 3.3, *min* = 25, *max* = 40). In the infant sample, there were no differences between male and female infants in the proportion of the recorded gaze (M_{female} = .88, M_{male} = .88, t = .05, df = 125, p = .95), or in the precision measure Dispersion (M_{female} = .38, M_{male} = .41, t = 1.55, df = 117, p = .12). There was also no correlation between infants' age and Dispersion (*cor* = .24, t = 1.3, df = 27, p = 0.2) or proportion of recorded gaze (*cor* = -.12, t = .61, df = 27, p = 0.5). The covariates age and sex were therefore not included in the main analysis. Further descriptives of the data for the joint infant-adult part of the experiment are provided in the supplementary material S5.

We assessed the effects of our independent variables via linear mixed-effects models using the lme4 package (Version 1.1-12; Bates, Mächler, Bolker, & Walker, 2015) in R (Version 3.3.3). Linear mixed-effects models (LME) are suitable for our study because they tolerate the unequal number of trials provided by our participants (for an application see: Laubrock, Engbert, Rolfs, & Kliegl, 2007). In the models, random slopes were specified for variations of the variable of interest between participants (Pinheiro & Bates, 2000).

Analysis Calibration Targets

Preference trials. To examine which targets attracted participants' attention, we analyzed how long they spent looking at the different calibration targets using dwell time. Total dwell time towards the calibration targets was calculated for the POI of an individual trial by the summing up all gaze points during fixations in the AOI of a target. Dwell time was then transformed by taking its square root to fit the data to a normal distribution. An LME model was conducted to infer how dwell time to a stimulus was explained by the kind of target video presented. The effect of calibration target was taken as random at the participant level, and participant group was included as fixed effect covariate. Calibration target ($F(5) = 47.9$, $p < .001$), participant group ($F(1) = 15.5$, $p < .001$), and their interaction ($F(5) = 8.7$, $p < .001$) substantially contributed to the model, which is confirmed by likelihood ratio tests, indicating that removing video ($\chi^2(5) = 77.8$), group ($\chi^2(1) = 9.5$) or their interaction ($\chi^2(5) = 32.9$) significantly decreased the goodness of fit (all $p < .005$). The estimated random effects accounted for a large part of the variance. Figure 2 illustrates dwell times estimated by the model as a function of the calibration videos and the participant groups.

For the infant group, Popflake I received the most attention. Popflake I dwell time was higher than for Bullseye ($\beta = 1885$ms, $SE = 138.8$, $p < .001$), Nautilus ($\beta = 1742$ms, $SE = 174$, $p < .001$), and Purple ($\beta = 1421$ms, $SE = 219.7$, $p < .05$). Bullseye was attended to for a shorter time than the other targets (all $t \geq 2.7$, $p < .01$) except Nautilus and Purple (all $t < .9$, *n. s.*).

{ Place Figure 2 here }

Verification trials. To assess the accuracy of infants' gaze elicited by the different calibration targets, we asked if Instability was affected by the calibration target. We also asked if the precision measure Dispersion was affected by the calibration target that was used for the initial calibration procedure of the respective sequence (for means and standard deviations see supplementary Table S5).

The dependent variables (DVs) were log transformed to fit normal distributions. Instability was best explained by the covariate participant group ($F(1) = 109.2$, $p < .001$), the attended calibration target ($F(5) = 12.1$, $p < .001$), head distance change ($F(1) = 10.8$, $p < .001$) and the interaction between group and target ($F(5) = 3.9$, $p < .01$). Removing any of the model terms led to a significant reduction of fit (all *p's* < .01). Intraclass correlation associated with the participants was controlled for by specifying participants as random intercept and target at the participant level as random slope. Instability in the infant group was higher than in the adult group ($\beta = .27°$, $SE = .04$, $t = 5.9$, $p < .001$), and a larger change of head distance after calibration led to higher instability ($\beta = .002°$, $SE = .0006$, $t = 3.4$, $p < .001$). Nautilus elicited the lowest Instability in the infant group, differing from Bullseye with $\beta = -.13°$, $SE = .047$, $t = -2.8$, $p < .01$ (see Figure 3). No other comparisons were significant.

The usage of a particular target for the initial calibration procedure only marginally predicted the precision measure Dispersion ($F(5) = 1.97$, $p < .10$). Instead, Dispersion was best estimated in an LME model that included participant group ($F(1) = 161.9$, $p < .001$), head distance change ($F(1) = 31$, $p < .001$), and as random slope head distance change at the participant level. Adding initial calibration target changed the model fit by ($\chi^2(5) = 9.8$, $p = .08$, *n.s.*), and removing any of the other variables significantly reduced its fit (all $p < .05$). Infants' fixations had a higher Dispersion than adults' fixations ($\beta = .16°$, $SE = .012$, $t = 12.7$, $p < .001$), and if head - camera distance increased after calibration for 10 mm, Dispersion increased for .013° ($SE = .0005$, $t = 2.6$, $p < .05$; see Figure S6).

{ Place Figure 3 here }

Spread trials. The 6s time course of the Spread trials was segmented into bins using the following data driven procedure. First, we identified turning points in the slope over which the infant Instability measure developed over time with the R package strucchange (Zeileis, Kleiber, Krämer, & Hornik, 2003). We then defined six bins of approximately similar length around each turning point. Participants' Instability values within a bin, and within the entire POI for a particular target, were then aggregated in order to analyze differences in gaze accuracy towards the target variants over time. Because the targets in the Spread trials increased in size and decreased again over the course of the trial, the six bins also captured gaze toward the target at different sizes.

When looking at the whole POI, Instability was best explained in a model including target variant ($F(5) = 16.3$, $p < .001$), participant group ($F(1) = 12.8$, $p < .001$) and bin ($F(5) = 2.9$, $p < .05$), and their interactions (target variant - bin ($F(25) = 4.6$, $p < .001$; group - bin ($F(5) = 12.5$, $p < .001$). Target variant at the participant level was specified as a random slope. Including the target x group interaction does not improve the fit ($\chi^2(30) = 1.5$, *n.s.).*

Gaze towards the stimuli during the subsequent bins was then analyzed. Within each bin, the effects of target variant and the interaction between participant group and target variant on Instability of gaze were estimated, with participants as a random intercept (Figure 4a). To account for multiple comparisons, we will only report differences related to infants' instability of gaze towards the stimuli at a significance level $p < .01$ (Figure 4b).

Infants' gaze became less stable over time and varied by target (Figure 4a, right panel). In the earliest segment between 0.8 and 1.7s, Bin 1, only CentBlink triggered lower Instability than ContrRings and FacetTwist. Bin 2 between 1.7 and 2.55s, which included the

fully expanded stimuli, revealed three target variants with better accuracy than the other target variants. Precisely, CentrBlink and Popflake II led to more stable and central fixations than ContrRings, FacetTwist and BlurRings, and SpiralTwist elicited more accuracy than ContrRings. This same pattern of results occurred within Bin 3, this time showing the largest discrepancies of the entire trial. In Bin 4, between 3.3 and 4.15s, only Popflake II differed from the three lower accuracy target variants. However, in Bin 5 gaze towards Popflake II increased in Instability, and only CentrBlink and SpiralTwist differed from FacetTwist, the latter as well from ContrRings. In Bin 6 all targets were viewed with similar, increasingly high Instability (for coefficients, standard errors and significance values see Table S5.).

{ Place Figure 4 here }

Calibration-Repetition trials. Next, we assessed adult participants' accuracy scores with our DVs Displacement and Instability. This allowed us to compare the performance of these two accuracy measures.

The LME model that explained Displacement best included the factor calibration target ($F(5) = 4.4$, $p < .01$), the factor target location ($F(1) = 4.6$, $p < .05$), the continuous variable head distance change ($F(1) = 22.1$, $p < .001$), and participant as a random intercept. Removing calibration target ($\chi^2(5) = 18.4$, $p < .01$), target location ($\chi^2(1) = 4.2$, $p < .05$) or head distance change ($\chi^2(1) = 4.3$, $p < .05$) would significantly decrease in the model's goodness of fit. Displacement increased with calibration targets presented at a peripheral location ($\beta = .04°$, $SE = .019$, $t = 2.1$), and with a larger head distance from the screen ($\beta = .007°$, $SE = .002$, $t = 4.7$). The calibration target Nautilus was attended to with the lowest Displacement and differed from all other videos except Harp (all t's < 2.5), while Purple was attended to with the highest Displacement differing from Nautilus and Harp, with all t's > 2.5.

Instability was best explained by calibration target ($F(5) = 9.5$, $p < .001$), target location ($F(1) = 17$, $p < .001$) and calibration target at the participant level as random slope. Instability increased with targets shown at a peripheral location ($\beta = .04°$, $SE = .009$, $t = 4.1$). Here as well, Nautilus was attended to with the lowest Instability and differed from all other calibration target except Harp (all $t's > 2.5$). Of those targets with low accuracy it was Purple which led to highest Instability scores, differing from all calibration targets except Bullseye, with all $t's > 2.7$ (see Figure 5). The measures Dispersion and Instability were correlated ($r_{df\ 714} = .35$, $t = 10$, $p < .001$), indicating an association of medium effect size between the two measures.

{ Place Figure 5 here }

Adult Movement Tasks

Finally, using our adult participants, we asked how head and body movements (see section *Procedure for Adult Experiment*) affect accuracy (Displacement) and precision (Dispersion, RMS) compared to recordings without movement (the control condition *Fix*), and if there is an effect of target location on the gaze measurement. The targets appeared at nine screen locations, and were grouped as *Center* (central), *Central-Peripheral* (central on one axis but peripheral at the other axis) or *Peripheral* (all four corners). In all LME models, movement type at the participant level was included as a random slope.

Displacement was best predicted with movement type ($F(3) = 64$, $p < .001$), target location ($F(2) = 88.9$, $p < .001$), and their interaction ($F(6) = 5.4$, $p < .001$; Figure 6a). All movement types led to increased Displacement (Side Movement: $\beta = .14°$, $SE = .04$, $t = 3.7$; Head Movement: $\beta = .17°$, $SE = .04$, $t = 3.9$; Bend Movement: $\beta = .55°$, SE = .07 $t = 8.4$), and non-central target locations led to larger Displacement than centrally presented targets

(Central-Peripheral: β = .08°, *SE* = .02, *t* = 3.5; Peripheral: β = .14°, *SE* = .03, *t* = 5.5). Bending towards the screen significantly increased Displacement at Peripheral locations (β = .13°, *SE* = .04, *t* = 3.6; all $p < .001$).

Dispersion was predicted by movement type only ($F(3) = 34.8$, $p < .001$). Adding target location to the model did not improve the fit ($\chi^2(2) = 1.3$, $p = .53$), and although the interaction of target location and movement type improved the model fit ($\chi^2(6) = 14$, $p = .03$), we decided against including it for parsimonious reasons and because the interaction without a main effect of target location would not be meaningful here. Head Movement increased Dispersion (β = .046°, *SE* = .008, *t* = 5.6, while Bend Movement reduced Dispersion (β = -.029°, SE = .004, *t* = 8; both $p < .001$). Dispersion elicited by Side Movement did not differ from the stable position.

RMS was best predicted by movement type ($F(3) = 63.2$, $p < .001$) and target location ($F(2) = 11.5$, $p < .001$; Figure 6b). In a similar pattern as Dispersion, RMS was reduced in Bend Movement (β = -.0023°, *SE* = .0003, *t* = 7.1), but increased in Head Movement (β = .0025°, *SE* = .0003, *t* = 8.6) and Side Movement (β = .0003°, *SE* = .0001, *t* = 2.3). Non-central target locations led to decreased RMS than centrally presented targets (Central-Peripheral: β = -.0004°, *SE* = .0001, *t* = 3.1; Peripheral: β = -.0006°, *SE* = .0001, *t* = 4.7; all $p < .05$).

{ Place Figure 6 here }

{ Place Figure 7 here }

Discussion

In the present study, we investigated the impact of different infant calibration targets and movements during gaze recording on eye-tracking data quality with infant and adult

participants using EyeLink 1000 Plus Remote Mode technology. We found that certain visual attributes of the calibration targets, as well as the duration of their presentation, influenced infants' gaze instability. Targets with interesting centers and low contrast at their periphery resulted in better gaze recording outcomes. Body movement substantially contributed to gaze instability and fixation dispersion. All movement types we tested with adults negatively affected accuracy, as did the eccentricity of a target's location. Movement towards the screen particularly increased peripheral gaze displacement and following a target with head turns resulted in less precise gaze.

Calibration Targets Influence Stability of Gaze

Infants fixated our calibration targets with different gaze stability, demonstrating that some characteristics of an animated graphical form elicited more accurate gaze than others. Interestingly, our results showed that infants' preference to look at a particular calibration target was not predictive of the data quality elicited by that same target in our study. Infants fixated on the target Nautilus for the least amount of time in the Preference trials, but Nautilus nevertheless led to the highest stability of gaze points in the Verification trials. The calibration target which elicited the greatest preference, Popflake I, led to similar gaze stability as Nautilus.

By reducing the attributes of our calibration targets in the Spread trials, we were able to infer which visual characteristics contributed to stable gaze. Our results showed that animations with an interesting center but low contrasts in their periphery (CentrBlink, SpiralTwist), as well as very complex concentric animations (Popflake II), elicit the most stable gaze over time and are therefore better suited for infant calibration. CentrBlink and SpiralTwist share two important attributes with the (not reduced) Nautilus target that performed well in the Verification trials: a blurred periphery and an interesting (blinking and high contrast) center. The target variants leading to less stable gaze consisted of blurred

concentric forms without a clear center (as in BlurRings), or point symmetrical patterns with distributed contrast which were not blurred in their eccentric parts (ContrRings and FacetTwist; for a detailed description of all target variants see Figure 1).

The decision of when to accept infant's gaze to a target during the calibration procedure is another important criteria for calibration success. Our results from the Spread trials, in which the target variants appeared to loom over the course of the trial, indicated that accuracy dropped similarly for all target variants over time in our infant sample. About four seconds after stimulus onset, infants' gaze started to be less stable even for targets that were fixated more accurately, and after 5 seconds, differences between targets could no longer be found (see Figure 4). This is in contrast to adults, who attended to the shrinking targets with increasing gaze stability over time.

To better understand why infants' gaze decreased in stability over time, we compared our Instability measure to the more common accuracy measure of Displacement (also termed "offset" by Hessels et al., 2015a) in the adult Calibration-Repetition task. The correlation between Displacement and Instability was of medium effect size in the adult Calibration-Repetition trials, indicating that the two measures were similar but not entirely overlapping. The most obvious difference between the two measures occurred for the visually demanding video Popflake I (see Figure 5). Given that the Calibration-Repetition trials were the last block of the experiment, the adult participants were already well acquainted with the targets, and more likely to direct their gaze to details of Popflake I's silhouette as is reflected in the higher Instability score for this target. We therefore interpret the increase of Instability in the later portions of the infant Spread trials as less central gaze, because by this point of the trial, infants became inattentive and increased exploratory gaze around more distributed screen areas. Alternatively, the increase of Instability can also be understood as a loss of interest in the target decreasing in size. These explanations are not mutually exclusive.

Our calibration targets clearly differed in how they elicited central gaze, therefore it was surprising that they only marginally modulated later gaze precision when they were used during the initial calibration procedure. This may have occurred for several reasons. First, fixation control develops until early adolescence (Buquet & Charlier, 1996; Ygge, Aring, Han, Bolzani, & Hellström, 2005). Infants' fixations generally cover a larger area than adults' and are less stable (Luna et al., 2008; Zihl & Dutton, 2015), which may have obscured potential differences during fixation and is in line with the significant effect of the covariate participant group (infant vs. adult). Additionally, movement during recording significantly contributed to the variance of our infants' Dispersion scores, leading to a loss of statistical power such that the effects of our calibration videos were only marginal (see Supplementary S8 for further discussion of this point).

It was difficult to implement a validation of calibration success for our infant participants as it is commonly implemented in adult eye-tracking. The repeated presentation of the identical target at all five calibrated screen locations directly following calibration typically led to infant impatience and inattentiveness. Therefore, in many of the cases we omitted the validation procedure from the trial sequences. Additionally, the generally poor accuracy score reported by the eye-tracker for the attempted validations may not have been attributable to calibration success per se, but instead to the effects of movement due to infant inattentiveness during the validation procedure (see S6 for a description of the validation attempts). As a result of these kinds of difficulties, experimenters often skip validation procedures with infants and instead rely on the pictorial pattern of the calibration map to infer calibration success. In future research it would be worth investigating whether the symmetry of the calibration coordinates provided by the EyeLink output can be quantified and included in the statistical analysis (for similar suggestions based on Tobii technology see Dalrymple, Manner, Harmelink, Teska, & Elison, 2018).

Movement Affects Accuracy and Precision in Opposite Directions

The accuracy of gaze measurement in our adult sample was affected by all of the movement types we examined. When the adult's eyes and the bullseye head sticker briefly moved outside of the area registered by the eye-tracking camera—a common occurrence during infant eye-tracking—and returned to central and stable position before the recording started, the target - POG distance systematically increased by .15°. This adds to the findings of Niehorster and colleagues (2017) who performed a similar task and found a right sided insensitivity of the EyeLink system towards the returning gaze. A similarly strong impact of movement on accuracy occurred during head turns towards the target leading to an increased offset of .17°. Turning the head in the direction of a stimulus is also a common movement during infant eye-tracking, since perceptuomotor coordination accompanies attentional strategies and learning in infancy (Gibson, 1969; Yoshida & Smith, 2008).

Movements toward the screen had the strongest effect on gaze accuracy. Displacement not only generally increased by .55° for this movement type, but was further augmented by .13° for targets presented in the four corners of the screen. In fact, our data revealed that during all trials the presentation of non-central targets systematically added between .08° and .14° to the measured gaze - target distance. This finding underscores the importance of using variable target locations during intermittent drift checks to verify calibration accuracy during infant eye-tracking. The full range of drift would not be detected if drift check targets are located only at the screen center. A warped POG map resulting from intermittent movement could also lead to imprecise post hoc adjustment of gaze data if a one directional displacement is assumed.

There was a different pattern of results for precision during the adult movement tasks. Fixation dispersion was unaffected by target location, while non-central target locations reduced RMS values. Head turns decreased precision as assessed by both scores. However, bending toward the screen seemingly increased precision as assessed by Dispersion and RMS.

This apparent increase in precision was surprising given that the bend movement led to the lowest gaze accuracy. The reason for this discrepancy, as Figures 7a,b show, is that movement towards the screen after calibration made the POGs drift towards the center of the monitor. This resulted in a reduction of the size of the POG map and in a shrinkage of the inter sample distances. At the same time, the offset of the measured POGs increased, resulting in higher displacement values especially at non-central target locations. This finding also illustrates the necessity of exploring data in multiple ways to avoid misinterpretation—here, better precision scores clearly do not reflect higher data quality.

The high inter sample distance during the Head Movement task may reflect data quality loss originating from the combination of head turns and movements as the adults turned their heads to follow the movement of the target during this task. A change in the angle of the eyes influences the assessed pupil size (Hayes & Petrov, 2016) which again affects the estimation of POGs (Choe, Blake, & Lee, 2016; *EyeLink 1000 Plus User Manual*, 2015; Nyström et al., 2016). Moreover, the bullseye sticker that indicates a participant's head position moves slightly sideways and in its angle during these kinds of movements. Infants usually spontaneously perform a combination of different movements, including more excessive angular positions than adults. Accordingly, these combinations of movement may have caused the considerably higher RMS values for the infant sample than those of the adult sample ($Md_{Infants}$, = .021, *min = .008, max = .063* compared to Md_{Adults} = .011, *min = .006, max = .02*). This finding emphasizes the care that needs to be taken when comparing participant groups of different age, even if no strong distance changes to the eye-tracking camera are obvious (see Supplementary S9 for further discussion).

Taken together, our findings for the movement tasks demonstrate that the consequences of unconstrained recording situations on gaze DVs are difficult to calculate. Specifications given by manufacturers are usually achieved under optimal conditions and differ from the specifications assessed with naturally behaving participants. Our data quality

scores were preprocessed (e.g., means of POIs or fixations limited by AOIs) to estimate the variability that may occur during analysis of gaze from participant groups that cannot be restrained. In the user manual of the EyeLink eye-tracker (2013-2015), head movement of 35 cm in vertical and horizontal direction are said to be tolerated without accuracy reduction for a camera distance of 60 cm (EyeLink, 2015). For movements towards the camera, the system reports a warning if the distance exceeds a 20 cm range, outside of which accuracy can not be guaranteed. However, in our study movement within these ranges clearly affected DVs (for further examples including angular movements and recovery of the eye-tracker after loss of the eye, see Hessels et al., 2015b; Niehorster et al., 2017). Future studies could investigate the usefulness of including the change in head distance registered by the eye-tracker as control variable during Remote Mode infant eye tracking.

Practical Implications

Our results point to several practical steps that infant researchers can take to improve eye-tracking data quality. Of course, the requirements for gaze accuracy depend on the specific context in which eye-tracking data are collected. Therefore, researchers should take into account the demands of their phenomena of interest and of their experimental design when implementing any of our suggestions.

First, we suggest using calibration targets with an interesting center and low contrast in their periphery or globally distributed complexity. Calibration targets with these characteristics—including some kind of movement to attract infants' attention as all of our stimuli did (e.g., looming, twisting, etc.)—elicit more accurate gaze. Even if the differences in accuracy between the types of target used might only seem marginal in some cases, it is nevertheless important to optimize as many aspects of the calibration procedure as possible. Calibration targets that are not controlled in their distribution of contrast or luminance—even if they are provided by some eye-tracking systems—should be avoided. The calibration

targets that worked well in our study are available online (see link in the conclusions section below).

Importantly, gaze toward calibration targets during the calibration procedure should be accepted within the first four seconds because attention towards the targets is higher during this phase. To further facilitate infants' attention when repeated calibrations or drift checks are necessary, calibration animations that elicit precise gaze can be alternated. Additionally, the background color of the screen on which the calibration target is shown can be changed to facilitate infants' interest in the display. Because alterations of the display's luminance level would result in changes in pupil size and affect gaze measurement, changes in brightness entering the participant's eye should generally be avoided in eye-tracking experiments. If the background color change is controlled for luminance, it will not interfere with accuracy (EyeLink, 2015). Moreover, depending on the constraints of the experimental conditions, trials can be accompanied by changing sounds or music. In our study, infants were repeatedly confronted with the same six calibration targets during the trial sequences, and we successfully used background color changes and music as described.

Calibration success is crucial for all infant eye-tracking studies, independent of the technology that is used. Poor calibration procedures have a particularly negative effect on infant eye-tracking procedures because the number of trials in these studies is limited by infants' shorter attention spans. Therefore, the risk of a high amount of missing data and incorrect data points can be mitigated by adopting higher quality calibration procedures.

In addition to optimizing calibration targets and procedures, the diverse effects of movement on our gaze measures in the present study should be kept in mind when planning infant eye-tracking studies. Movement towards the screen has an especially high impact on spatial accuracy, and if fixation positions on AOIs are assessed, researchers should expect misplaced POGs with large offsets especially at peripheral screen locations. In such cases, adapting the AOIs accordingly may avoid alterations of the variables of interest (Holmqvist et

al., 2012; Orquin, Ashby, & Clarke, 2016). For example, in paradigms that compare attention to multiple areas of the screen, AOIs could be reduced in size, so that misplaced POGs fall into neutral screen areas rather than being falsely attributed to the wrong AOI. A warped POG map, with larger peripheral offsets, could also lead to systematic errors between central and peripheral AOIs.

Experimenters should be attentive to movement throughout the recording sessions and have recalibration procedures prepared if infants exhibit excessive movement of any kind. The measurement of head target - camera distance provided on the EyeLink camera set-up screen as well as a blurred camera image of the eye can both be used as indicators for distance changes even if the eye-tracker does not provide a warning message. Additionally, implementing intermittent drift checks with central and non-central target locations can help to detect shifts of the POGs and possible skewness of the POG map. These checks can occur at regular intervals during the trials. POG shifts can also be assessed via additional software implemented in the experiment (e.g., Dalrymple et al., 2018; Frank et al., 2012).

Studies targeting psychophysical research questions that are more sensitive to fine grained changes in inter-sample distance should be especially aware of the diverse movement effects. If, for example, participant groups differ systematically in their motoric responses, as is the case for comparisons of infants and adults, the resultant systematic distortions in the assessed data could lead to false inferences about group differences. Studies that are particularly sensitive to dispersion of gaze points should consider the inclusion of drift checks and recalibrations at several predetermined intervals during the recording sequence.

Following these practical steps can help to mitigate the problems of infant eye-tracking and increase the quality of measured gaze.

Conclusion

During infant eye-tracking, uncertainty about calibration success, fussiness caused by the repetition of calibration stimuli, and body movements during testing are frequent constraints on measurement quality. Our systematic investigation of these constraints with infants and adults revealed some characteristics of calibration targets that elicit more reliable data. These calibration targets can be flexibly implemented in different calibration procedure designs and are provided online, together with the necessary information on the adjustment of the background color (https://osf.io/3k8jp/?view_only=e8075dc7bf0e4ab780c5e620b8f4860f). Using EyeLink 1000 Plus technology, we also discovered heterogeneous effects on accuracy and precision as result of movement types which are common during infant eye-tracking. These findings provide some insight into measures that can be taken to improve data quality when conducting infant eye-tracking studies.

References

Alahyane, N., Lemoine-Lardennois, C., Tailhefer, C., Collins, T., Fagard, J., & Doré-Mazars, K. (2016). Development and learning of saccadic eye movements in 7- to 42-month-old children. *Journal of Vision*, *16*(1), 6. https://doi.org/10.1167/16.1.6

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x

Aslin, R. N. (2012). Infant Eyes: A Window on Cognitive Development. *Infancy*, *17*(1), 126–140. https://doi.org/10.1111/j.1532-7078.2011.00097.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(i01).

Bornstein, M. H., & Benasich, A. A. (1986). Infant Habituation: Assessments of Individual Differences and Short-Term Reliability at Five Months. *Child Development*, *57*(1), 87–99. https://doi.org/10.2307/1130640

Buquet, C., & Charlier, J. R. (1996). Evaluation of sensory visual development based on measures of oculomotor responses. In F. Vital-Durand, J. Atkinson, & O. J. Braddick (Eds.), *Infant Vision* (pp. 291–306). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198523161.003.0019

Choe, K. W., Blake, R., & Lee, S.-H. (2016). Pupil size dynamics during fixation impact the accuracy and precision of video-based gaze estimation. *Vision Research*, *118*, 48–59. https://doi.org/10.1016/j.visres.2014.12.018

Constantino, J. N., Kennon-McGill, S., Weichselbaum, C., Marrus, N., Haider, A., Glowinski, A. L., … Jones, W. (2017). Infant viewing of social scenes is under genetic control and atypical in autism. *Nature*, *547*(7663), 340–344. https://doi.org/10.1038/nature22999

Dalrymple, K. A., Manner, M. D., Harmelink, K. A., Teska, E. P., & Elison, J. T. (2018). An Examination of Recording Accuracy and Precision From Eye Tracking Data From Toddlerhood to Adulthood. *Frontiers in Psychology, 9.* https://doi.org/10.3389/fpsyg.2018.00803

Doucet, M.-E., Gosselin, F., Lassonde, M., Guillemot, J.-P., & Lepore, F. (2005). Development of visual-evoked potentials to radially modulated concentric patterns. *Neuroreport*, *16*(16), 1753–1756.

EyeLink Data Viewer User's Manual. (2015) (Version 2.3.1). Mississauga, Ontario, Canada: SR Research Ltd.

EyeLink 1000 Plus User Manual. (2015) (Version 1.0.6). Mississauga, Ontario, Canada: SR Research Ltd.

Frank, M. C., Vul, E., & Saxe, R. (2012). Measuring the Development of Social Attention Using Free-Viewing. *Infancy*, *17*(4), 355–375. https://doi.org/10.1111/j.1532-7078.2011.00086.x

Gibson, E. J. (1969). *Principles of perceptual learning and development*. Englewood Cliffs, NJ: Prentice-Hall.

Gredebäck, G., Johnson, S., & von Hofsten, C. (2009). Eye Tracking in Infancy Research. *Developmental Neuropsychology*, *35*(1), 1–19. https://doi.org/10.1080/87565640903325758

Hayes, T. R., & Petrov, A. A. (2016). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods*, *48*(2), 510–527. https://doi.org/10.3758/s13428-015-0588-x

Hessels, R. S., Andersson, R., Hooge, I. T., Nyström, M., & Kemner, C. (2015a). Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, *20*(6), 601–633.

Hessels, R. S., Cornelissen, T. H. W., Kemner, C., & Hooge, I. T. C. (2015b). Qualitative tests of remote eyetracker recovery and performance during head rotation. *Behavior Research Methods*, *47*(3), 848–859. https://doi.org/10.3758/s13428-014-0507-6

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

Holmqvist, K., Nyström, M., & Mulvey, F. (2012). Eye Tracker Data Quality: What It is and How to Measure It. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 45–52). New York, NY, USA: ACM. https://doi.org/10.1145/2168556.2168563

Komogortsev, O. V., Jayarathna, S., Koh, D. H., & Gowda, S. M. (2010). Qualitative and quantitative scoring and evaluation of the eye movement classification algorithms. In *Proceedings of the 2010 Symposium on eye-tracking research & applications* (pp. 65–68). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1743682

Kulke, L., Atkinson, J., & Braddick, O. (2015). Automatic Detection of Attention Shifts in Infancy: Eye Tracking in the Fixation Shift Paradigm. *PLOS ONE*, *10*(12), e0142505. https://doi.org/10.1371/journal.pone.0142505

Laubrock, J., Engbert, R., Rolfs, M., & Kliegl, R. (2007). Microsaccades Are an Index of Covert Attention: Commentary on Horowitz, Fine, Fencsik, Yurgenson, and Wolfe (2007). *Psychological Science*, *18*(4), 364–366. https://doi.org/10.1111/j.1467-9280.2007.01904.x

Leppänen, J. M., Forssman, L., Kaatiala, J., Yrttiaho, S., & Wass, S. (2015). Widely applicable MATLAB routines for automated analysis of saccadic reaction times. *Behavior Research Methods*, *47*(2), 538–548. https://doi.org/10.3758/s13428-014-0473-z

Luna, B., Velanova, K., & Geier, C. F. (2008). Development of eye-movement control. *Brain*

and Cognition, *68*(3), 293–308. https://doi.org/10.1016/j.bandc.2008.08.019

Morgante, J. D., Zolfaghari, R., & Johnson, S. P. (2012). A Critical Test of Temporal and Spatial Accuracy of the Tobii T60XL Eye Tracker: *Infancy*, *17*(1), 9–32. https://doi.org/10.1111/j.1532-7078.2011.00089.x

Niehorster, D. C., Cornelissen, T. H. W., Holmqvist, K., Hooge, I. T. C., & Hessels, R. S. (2017). What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods*, 1–15. https://doi.org/10.3758/s13428-017-0863-0

Nyström, M., Hooge, I., & Andersson, R. (2016). Pupil size influences the eye-tracker signal during saccades. *Vision Research*, *121*, 95–103. https://doi.org/10.1016/j.visres.2016.01.009

Oakes, L. M. (2012). Advances in Eye Tracking in Infancy Research. *Infancy*, *17*(1), 1–8. https://doi.org/10.1111/j.1532-7078.2011.00101.x

Orquin, J. L., Ashby, N. J. S., & Clarke, A. D. F. (2016). Areas of Interest as a Signal Detection Problem in Behavioral Eye-Tracking Research. *Journal of Behavioral Decision Making*, *29*(2–3), 103–115. https://doi.org/10.1002/bdm.1867

Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-Effects Models in S and S-Plus*, 3–56.

Renswoude, D. R. van, Raijmakers, M. E. J., Koornneef, A., Johnson, S. P., Hunnius, S., & Visser, I. (2018). Gazepath: An eye-tracking analysis tool that accounts for individual differences and data quality. *Behavior Research Methods*, *50*(2), 834–852. https://doi.org/10.3758/s13428-017-0909-3

Saez de Urabain, I. R., Nuthmann, A., Johnson, M. H., & Smith, T. J. (2017). Disentangling the mechanisms underlying infant fixation durations in scene perception: A computational account. *Vision Research*, *134*, 43–59.

https://doi.org/10.1016/j.visres.2016.10.015

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42. https://doi.org/10.1016/j.visres.2012.10.012

Tobii Studio User's Manual (2016). Version 3.4.5. Copyright Tobii Technology AB.

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. https://doi.org/10.1111/infa.12055

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. https://doi.org/10.3758/s13428-012-0245-6

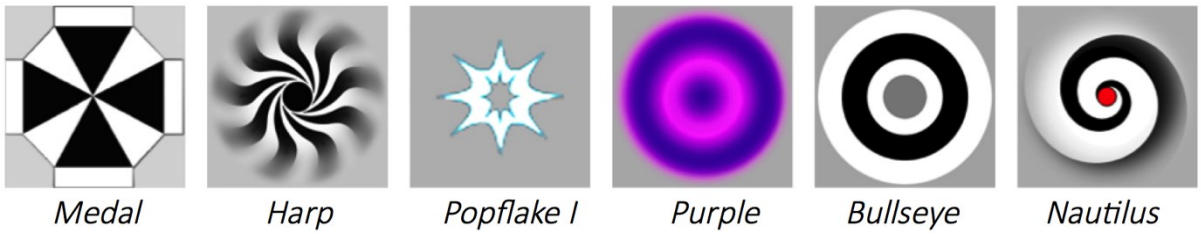
Ygge, J., Aring, E., Han, Y., Bolzani, R., & Hellström, A. (2005). Fixation Stability in Normal Children. *Annals of the New York Academy of Sciences*, *1039*(1), 480–483. https://doi.org/10.1196/annals.1325.049

Yoshida, H., & Smith, L. B. (2008). What's in View for Toddlers? Using a Head Camera to Study Visual Experience. *Infancy : The Official Journal of the International Society on Infant Studies*, *13*(3), 229–248. https://doi.org/10.1080/15250000802004437

Zeileis, A., Kleiber, C., Krämer, W., & Hornik, K. (2003). Testing and dating of structural changes in practice. *Computational Statistics & Data Analysis*, *44*(1), 109–123. https://doi.org/10.1016/S0167-9473(03)00030-6

Zihl, J., & Dutton, G. N. (2015). Development and Neurobiological Foundations of Visual Perception. In *Cerebral Visual Impairment in Children* (pp. 11–49). Springer Vienna. https://doi.org/10.1007/978-3-7091-1815-3_2

a.



b.

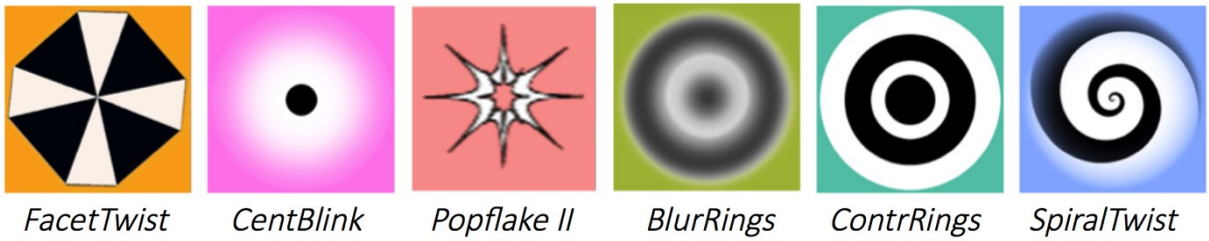


Figure 1. Examples of calibration targets and their variants in their fully developed form of appearance. Top row (a): calibration targets expanding to 5° when presented on the screen. Bottom row (b): Variants of the calibration targets used for the Spread trials, expanding to 17°. The modified calibration targets Popflake II and BlurRings (a variant of Purple) kept their distinct movements. Harp and Nautilus were reduced to CentBlink (a blinking central disc surrounded by a white corona) and SpiralTwist (a twisting spiral). ContrRings resembled Bullseye but lacked the blinking center. FacetTwist was identical to Medal except that it did not show the four white bars. Both ContrRings and FacetTwist kept their contrast in the periphery, while CentBlink and SpiralTwist had a blurred periphery. Two different background colors for each target variant in (b) were equally balanced over the participants.
Video examples are provided online (see https://osf.io/3k8jp/?view_only=e8075dc7bf0e4ab780c5e620b8f4860f). Harp, Nautilus and the modifications of the target variants are developed for the present study by the first author, the other calibration targets were kindly provided by other laboratories. We thank Scott Johnson, Gustav Gredebäck, Elika Bergelson and SR Research.

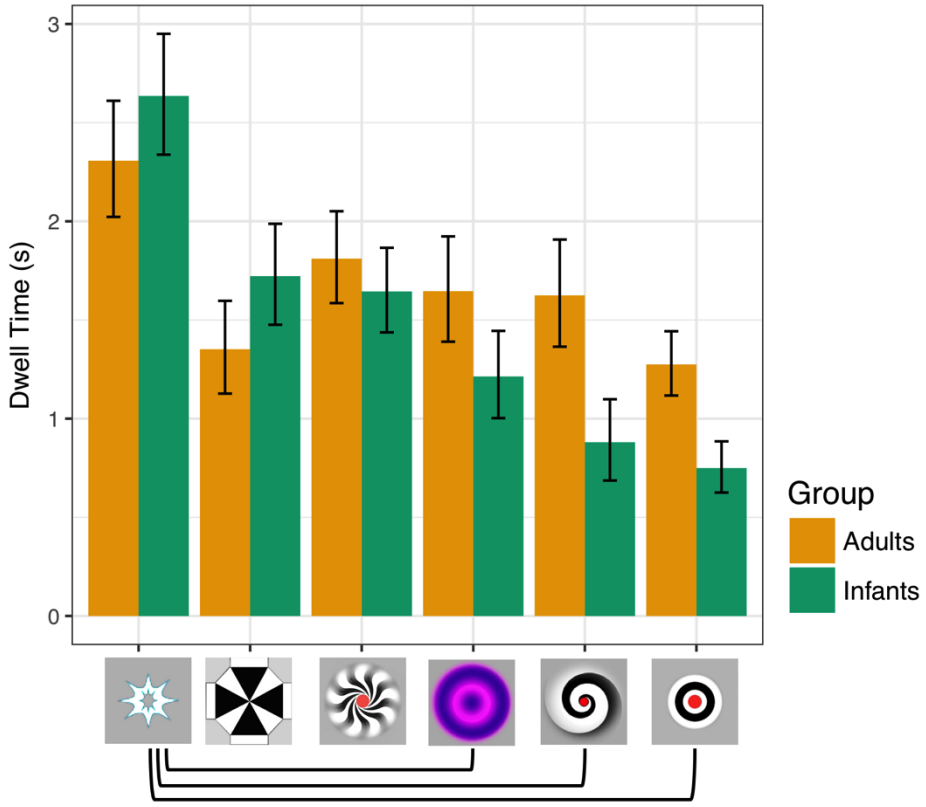


Figure 2. Preference for the calibration targets. Dwell time (ms) and 95% confidence intervals (CI) as a function of calibration target. Brackets depict significant differences ($p > .05$) between the targets. Predicted means in this and in the other plots are estimated and back transformed with the R package predictmeans version 0.99 (Luo, Ganesh & Koolaard, 2014).

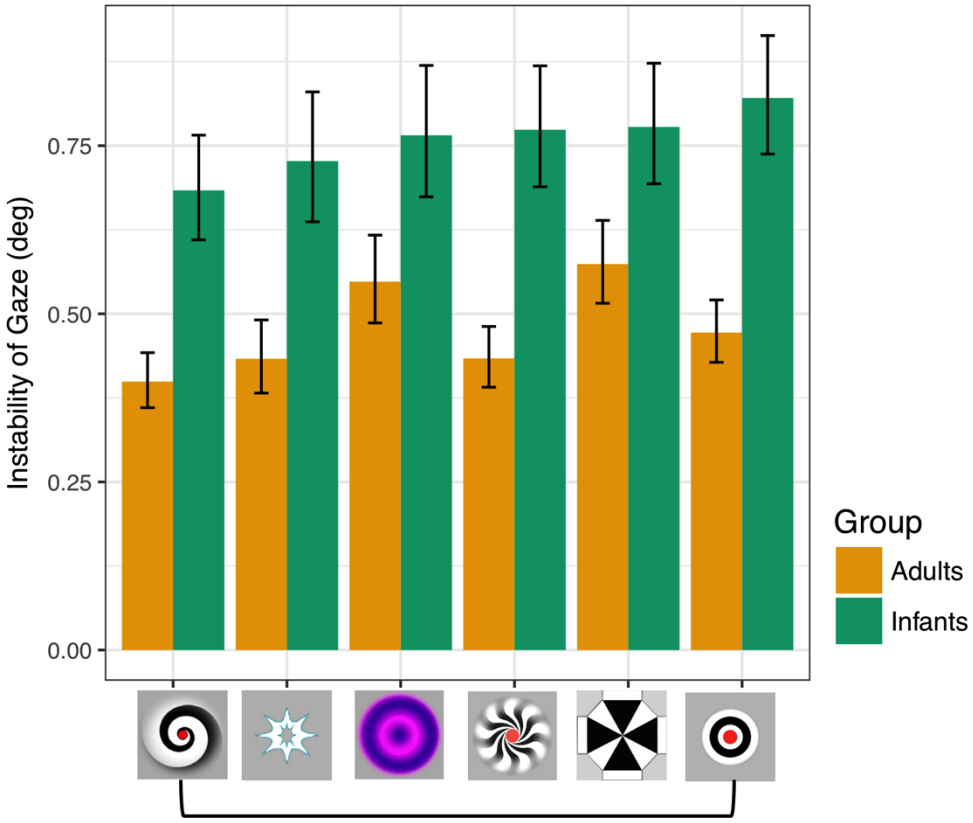
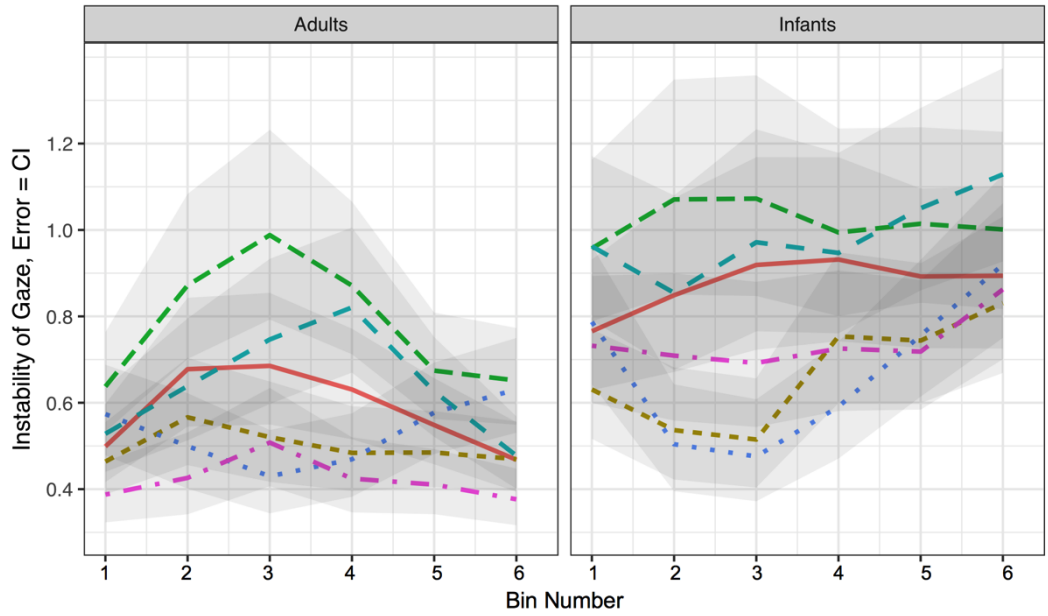


Figure 3. Instability as a function of calibration target and participant group. Values are back transformed and estimated for a head distance of 13.2 mm, with a CI of 95%. Instability of gaze in the adult group differed from the infant group in that adults attended the videos Medal and Purple with lower accuracy than the infants (all $t < 2.1$, $p < .05$). The bracket indicates the difference found in the infant group ($p < .01$).

a.



b.

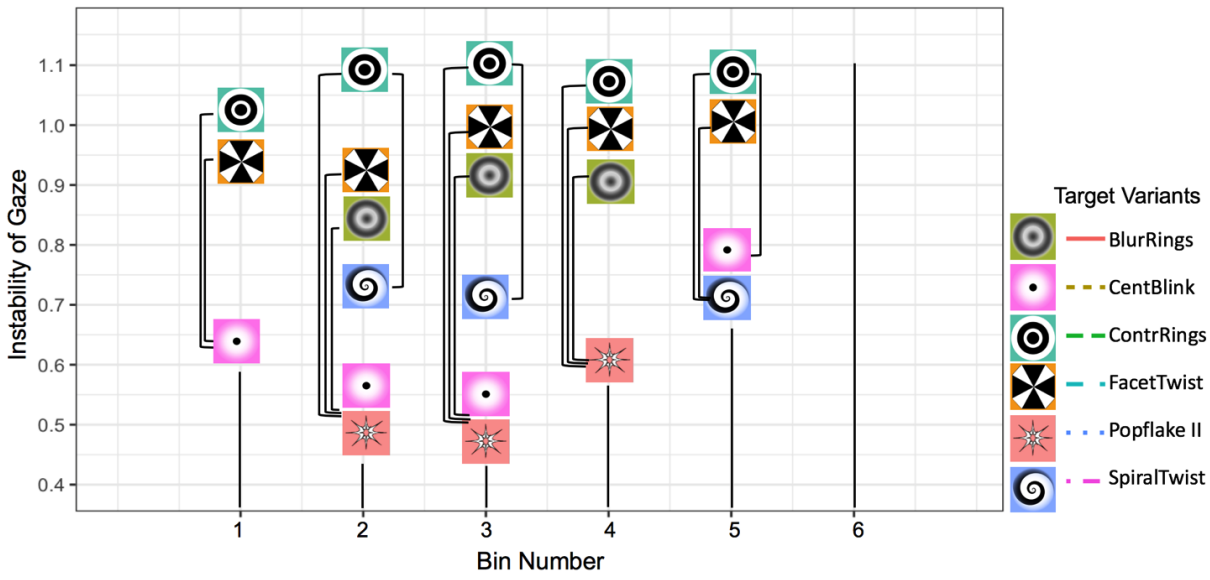


Figure 4. Changes in accuracy during the presentation period of the Spread trials. (a) Predicted Instability as a function of target variant and participant group, with 95% CI. (b) Infants' Instability as a function of target variants and bins. Brackets indicate differences with significance level $p <$ .01. For coefficients and standard errors see supplementary Table S5. The target variants were most expanded at 2s. Numbers on the x axis represent the 6 time bins.

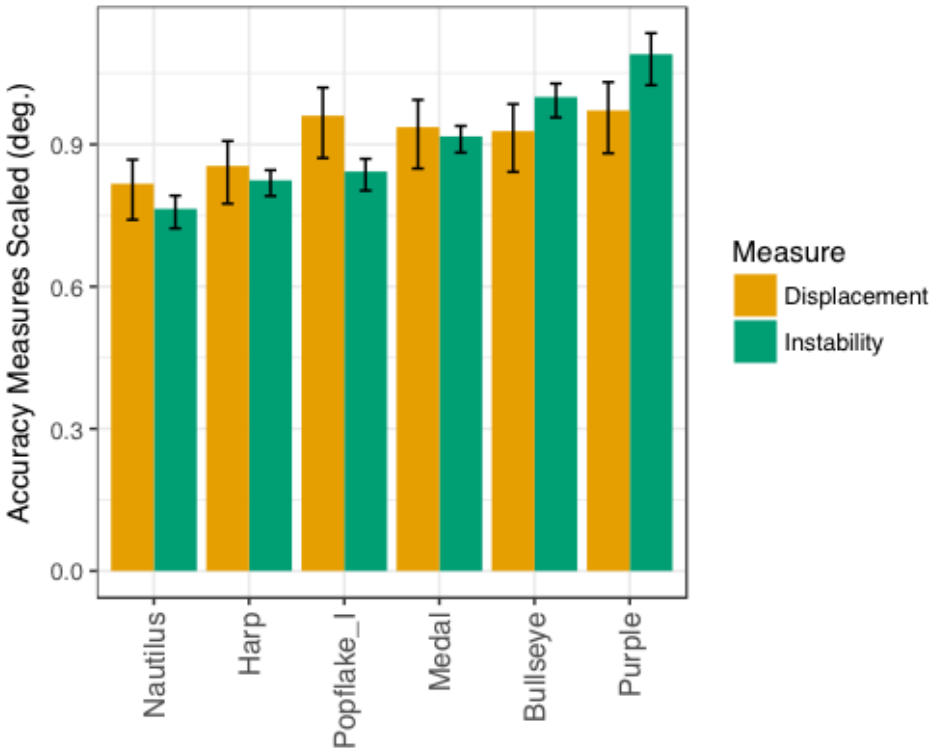


Figure 5. Comparison of the two accuracy measures. Scaled accuracy scores of the measures Displacement and Instability as a function of calibration target during the adult task Calibration-Repetition.

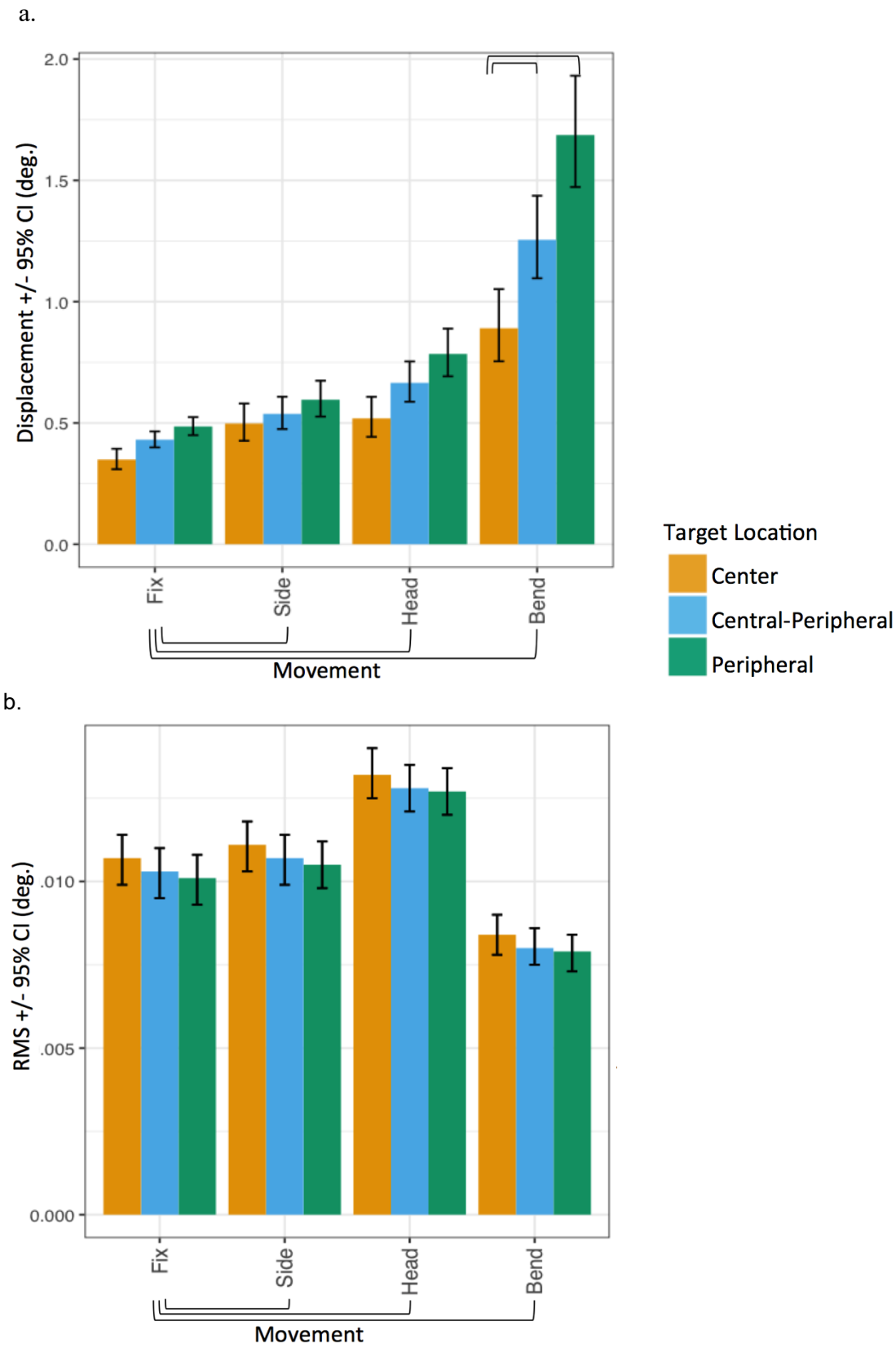
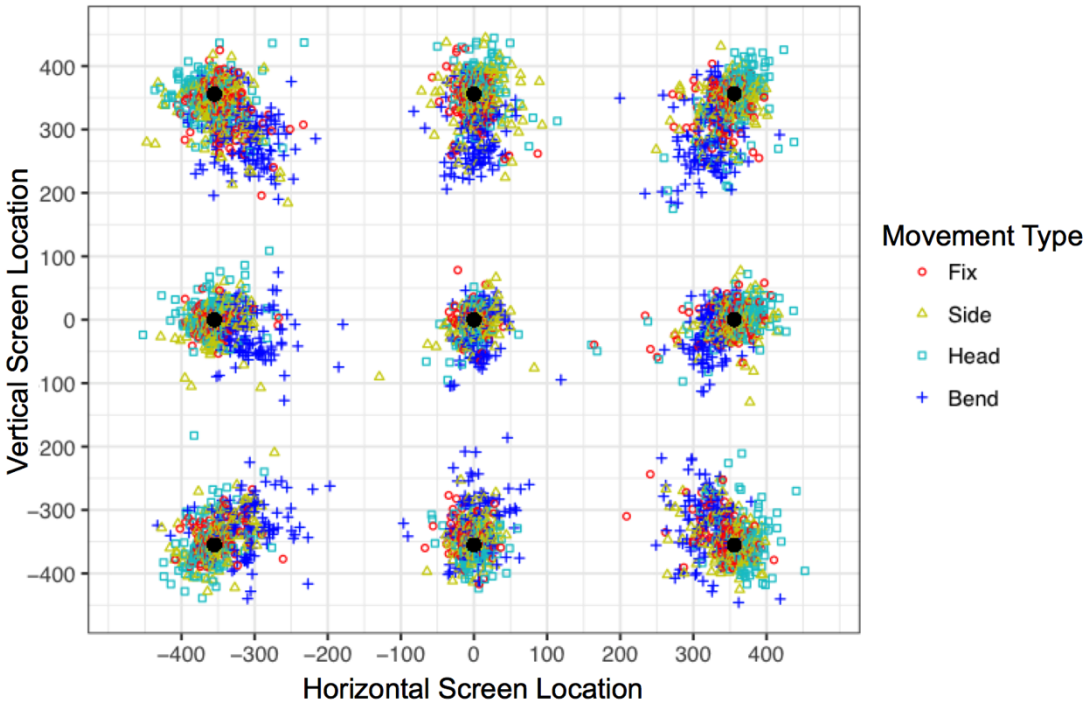


Figure 6. Effects of movement type on adult accuracy and precision. Accuracy (a: displacement) and precision (b: rms) as functions of movement type and target location. note the converse effects for accuracy and precision when approaching the screen during Bend Movement, changing the viewing angle during Head Movement, and when attending peripheral target locations.

a.



b.

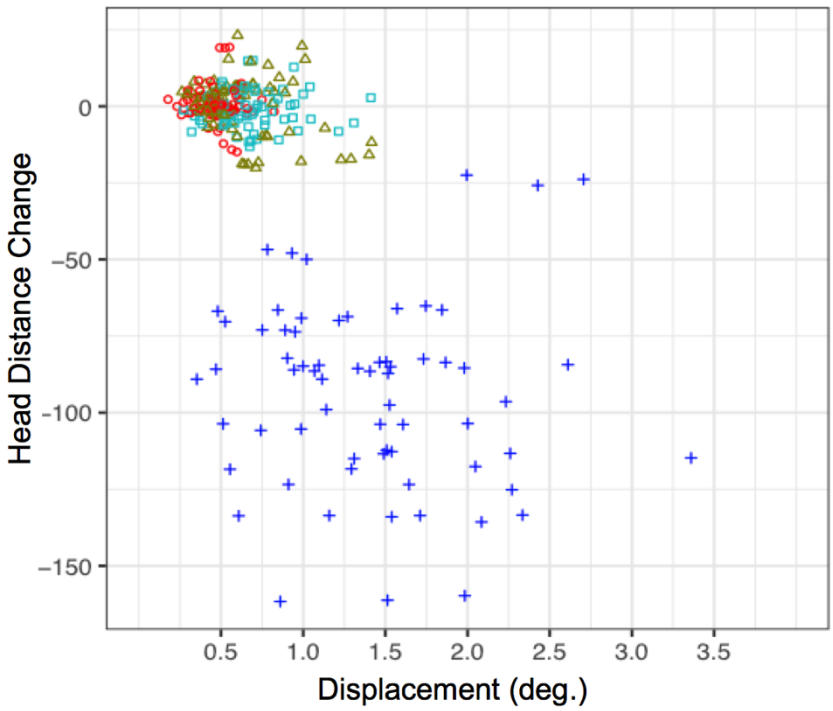


Figure 7. Movement type affects the registration of gaze points. Gaze points during the adult movement tasks plotted on their measured screen locations in pixel coordinates. Black discs indicate the actual target positions, inter target distance was 9° of visual angle (a). Accuracy (x axis) plotted against head - camera distance change after calibration. Negative values indicated reduced distance to the eye-tracking camera in mm (b).

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix 1. Supplementary methods.
Appendix 2. Supplementary results.
Appendix 3. Supplementary discussion.
Figure S1. Example of the visual target for the adult participants during the movement tasks.
Figure S2. Arrangements of the calibration target locations within one Preference trial.
Figure S3. Arrangements of the target locations in the Verification trials.
Figure S4. Boxplots for the participant groups in the joint part of the experiment.
Figure S5. Histogram of fixation duration for the two participant groups after exclusion of invalid trials.
Figure S6. Dispersion of the participant groups as a function of head distance change in the Verification trials.
Table S1. The sizes of areas of interest (AOI) and periods of interest (POI).
Table S2. Comparison of the adult and infant sample.
Table S3. Attempts to validate calibration success.
Table S4. Completed trials, means and standard deviation of the accuracy and precision scores during the Verification trials.
Table S5. Differences of gaze instability between target variants during the Spread trials.