Robustness and Precision: How Data Quality May Influence Key Dependent Variables in Infant Eye-Tracker Analyses

Sam V. Wass
Medical Research Council Cognition and Brain Sciences Unit, Cambridge, UK

Linda Forssman and Jukka Leppänen
School of Medicine
University of Tampere, Finland

In recent years, eye-tracking has become a popular method for drawing conclusions about infant cognition. Relatively little attention has been paid, however, to methodological issues associated with infant eye-tracking. Here, we consider the possibility that systematic differences in the quality of raw eye-tracking data obtained from different populations and individuals might create the impression of differences in gaze behavior, without this actually being the case. First, we show that lower quality eye-tracking data are obtained from populations who are younger and populations who are more fidgety and that data quality declines during the testing session. Second, we assess how these differences in data quality might influence key dependent variables in eye-tracking analyses. We show that lower precision data can appear to suggest a reduced likelihood to look at the eyes in a face relative to the mouth. We also show that less robust tracking may manifest as slower reaction time latencies (e.g., time to first fixation). Finally, we show that less robust data can manifest as shorter first look/visit duration. We argue that data quality should be reported in all analyses of infant eye-tracking data and/or that steps should be taken to control for data quality before performing final analyses.

Correspondence should be sent to Sam V. Wass, MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, United Kingdom. E-mail: sam.wass@mrc-cbu.cam.ac.uk
Eye-tracking has become an increasingly widespread research method in recent years. The number of articles on Web of Science featuring the word “eye-tracking” in the title or abstract increased from 345 in 2008 to 641 in 2013. Data presented in these articles have been used to motivate a number of claims about early cognitive development—for example, that infants with variation in serotonin-system genes show heightened attention to social signals of fear (Forssman et al., 2014a), that 24-month-old infants are capable of making false belief attributions (Southgate et al., 2007), and that children with language impairments show deficits in fine-grained inhibitory control (Kelly et al., 2013). The variety of conceptual challenges involved in making these inferences have been discussed elsewhere (Aslin, 2007, 2011).

Although eye-tracking has countless advantages over other means of measuring infants’ eye movements (e.g., higher temporal and spatial resolution), the use of this technology is also associated with several methodological issues that have been relatively little discussed hitherto (see Aslin, 2011; Gredeback, Johnson, & von Hofsten, 2010; Oakes, 2010a; Shic, Chawarska, & Scassellati, 2008, 2009). One issue, which is the focus of the present paper, is the possibility that systematic differences might be identifiable in the quality of raw eye-tracking data obtained from different individuals and that these differences might cause the appearance of differences in gaze behavior, without this actually being true. Such an explanation might be seen as analogous to the ongoing debate on motion biases in connectivity analyses based on magnetic resonance imaging—where, for example, it has been speculated that systematic differences in the amount of in-scanner movement between typical individuals and those with autism spectrum disorders (ASD) might in fact be the cause of the widely reported apparent differences in functional and structural brain connectivity (Power, Barnes, Snyder, Schlaggar, & Petersen, 2012; Yendiki, Koldewyn, Kakunoori, Kanwisher, & Fischl, 2013).

Our concerns pertain to three types of eye-tracking paradigm in particular. First, instances in which a single paradigm is shown both to typical individuals and an atypical group, as defined either clinically, epidemiologically, or based on some other facet of behavior. (Take, for example, studies comparing infants from high vs. low socioeconomic status backgrounds or studies comparing eye movements in good vs. poor language learners.) Could it be the case that the groups differ on some other aspect of behavior (such as fidgetiness or irritability during testing) which causes differences in data quality—which can cause the appearance of differences in gaze behavior, without this actually being the case? Second, instances in which a single
paradigm is shown to younger and to older individuals. In such a case, could it be that lower quality data obtained from younger individuals might cause apparent differences in gaze behavior? Third, studies in which data quality might vary systematically between conditions of an experiment. This might be the case in instances where one condition is more engaging than another or where one condition is consistently presented before another. Again, this might lead to differences in data quality between conditions, which might misleadingly cause the appearance of actual differences in gaze behavior.

If such differences were to exist, they would go largely undocumented. Even current best practices in infant eye-tracking (Oakes, 2010b) request only that experimenters describe the quality of the calibration obtained prior to recording in terms of the number of calibration points successfully obtained (see Table 1 for a definition of calibration and of other key technical terms described in this paper). They do not request any information about the quality of tracking obtained during recording. Yet this can vary substantially between individuals, as we describe below.

TABLE 1
Glossary

More detailed descriptions of these terms can be found in Holmqvist et al., (2011)

Areas of Interest (AOI) – Participants’ gaze is typically analyzed relative to certain user-defined Areas of Interest (AOIs) within the stimulus presentation screen. See Figures 6a,c and Figure 9a for examples

Calibration – a series of small objects are presented at predetermined locations on the screen; the participant’s position of gaze is recorded during this sequence. All subsequent gaze data have a correction that was calculated during calibration. When calibration data is absent, the eye tracker uses a default calibration based on a prototypical eyeball

Corneal reflection – all commercial eye trackers work by tracking the position of the participant’s pupil (which moves as the eye moves) relative to the position of an infra-red light reflected from the participant’s cornea (which remains still as the eye moves)

Fixation duration – when viewing a visual array we spontaneously manifest a sequence of eye movements in order to ensure that light from objects of interest is projected onto the fovea. Our eyes alternate between periods in which the eye is static and visual processing occurs (fixations) and rapid eye movements (saccades) during which visual processing is suppressed. Within the infant literature, the same term has also sometimes been used more approximately to refer to looks to vs away from the screen

Interpolation – the process of ‘filling in’ missing sections of data

Look/visit durations – these are typically reported as the time interval elapsed between gaze entering an AOI and leaving it. A number of fixations (refoveating eye movements) can therefore be contained within a look

Position of gaze (POG) – eye trackers typically return the participant’s position of gaze (POG) relative to a 2-D stimulus presentation area (usually a computer screen). POG is typically returned in X (horizontal) and Y (vertical) screen coordinates

Precision – the degree to which reporting of POG is consistent between samples
DATA QUALITY IN INFANT EYE-TRACKING—WHY MIGHT IT BE AN ISSUE?

Most eye-trackers used in infant eye-tracking require three separate elements to be identified to estimate where an infant is looking: the position of the pupil, the reflection of a point-light that is projected from the eye-tracker on the infant’s eyeball, and the position of the infant’s head in 3D space (Aslin & McMurray, 2004; Duchowski, 2007; Holmqvist et al., 2011; Kolakowski & Pelz, 2006). Figure 1a shows an example of a typical eye image in which the pupil and corneal reflection have been identified. Infants, unlike adults, often fidget and move in sitting position during recording. This can lead to changes both in the position of the head relative to the tracker and in the angle of the eyes relative to the tracker (see Figure 1b,c). These can disrupt the accuracy of eye-tracking in a number of ways (Holmqvist et al., 2011). In instances where the child is leaning back, or viewing the eye-tracker at an angle, the corneal reflection may be obscured by a “droopy” eyelid. Similarly, the pupil can be inaccurately identified if either the pupil or iris is partially obscured by the eyelid, or due to shadows on the iris. Positioning of the face nonperpendicular to the camera can disrupt the algorithms used to identify the position of the face in 3D space. Infants also present a number of further challenges during tracking. For example, increased wateriness in infants’ eyes (even in the absence of crying) can lead to the identification of multiple corneal reflections, causing inaccurate gaze tracking. Large pupils (common in infants) can be hard to identify by pupil detection algorithms designed for adult eyes. And finally, infant eye-tracking experiments often include displays with abrupt changes in luminance and participant movement, which can lead to inaccuracies in the automated techniques used for
finding and maintaining the correct threshold for identification of the pupil and/or corneal reflection.

These problems can affect the quality of data obtained during eye-tracking in four ways. The first problem, known as low precision data (Blignaut & Beelders, 2012; Blignaut & Wium, 2014; Holmqvist et al., 2011), occurs when one of the elements (pupil, corneal reflection, or head position) is incorrectly identified by the eye-tracking software. In cases where this error is random and varies from one gaze sample to the next, this can lead to increased sampling error (i.e., increased variance, akin to a “noisier signal”) in the reporting of the infant’s position of gaze. See Figure 2, Sample 2 for an example.

A second problem, known as low robustness, can occur if any one of the three elements is unavailable during tracking (e.g., a corneal reflection being obscured by an eyelid). This can lead to the tracker failing to report on the position of gaze at all. Inspection of the raw data obtained during tracking suggests that data often “flicker” off for periods ranging from a few milliseconds to several seconds. See Figure 2, Sample 3 for an example.

A third parameter of data quality, which we do not discuss in the present article, is that of spatial accuracy during tracking. This is a disparity between where an infant is actually looking and the reported position of gaze returned by the eye-tracker; often this is a consistent error and manifests as the eye-tracker recording gaze accurately but “throwing” every measurement off by a certain amount. Figure 3 illustrates the difference between spatial accuracy and precision. Frank, Vul, and Saxe (2011) and Morgante, Zolfaghari, and Johnson (2012) have both previously examined spatial accuracy in eye-tracking and reported substantial inaccu-
Good quality sample

Fixation durations (time intervals between eye movements)

Low precision sample

Low robustness sample

Figure 2  Samples of viewing data from a Tobii 1750 tracker recording at 50 Hz. Data are from 11-month-old typically developing infants. In eyetracking data four datapoints are available at each time point — showing where on the X dimension of the screen the left and right eyes are looking, and where on the Y dimension the left and right eyes are looking. The first sample illustrates high quality data, with continuous data and individual fixations and saccades clearly visible. Fixations (time intervals between eye movements) are clearly visible and are marked in red along the top. The position of gaze at each instant can be analyzed with a high degree of confidence (see sample frames). The second sample illustrates low precision data, in which reported POG is not stable. The third sample illustrates low robustness with periods of missing data ranging from one iteration to several 100 msec. Figure based on Wass et al. (2013). See Figure S1 for further samples of raw data, obtained from a Tobii TX300.
racies that can become more severe over time (“drift”). They assess spatial accuracy by adding an additional, user-defined calibration check during data recording (e.g., using scripts downloaded from: http://langcog.stanford.edu/materials/calib.html).

Based on these data, Frank and colleagues calculated two regressions for the $X$ and $Y$ coordinates independently and used these to correct the gaze estimates returned by the eye-tracker (Frank et al., 2011).

In addition, there is a fourth parameter of data quality, which we also do not discuss in depth in this article. This is temporal delay—that is,
the latency between an eye gaze event taking place and the eye gaze event being reported by the eye-tracker. Morgante et al. (2012) examined this parameter by comparing the POG overlay and the stimulus in a video exported from a Tobii Studio recording. The latency was measured by comparing the exported POG video with the Tobii Studio combined file (Text Export file). They found that the eye movement latency measured from the exported video might deviate up to 54 msec when compared with the Tobii Studio combined file. When the stimuli were presented in E-prime and recorded simultaneously in Tobii Studio as a video, they also observed a drift—that is, an increase in recording discrepancy with increasing time, between the exported POG video from Tobii Studio and the E-prime gaze data file. Similarly, Shukla, Wen, White, and Aslin (2011) used a separate video camera to analyze the time delay between an eye movement taking place and being recorded by the eye-tracker. They compared results from a high-speed camera running at 300 frames per second directed at their participant's eye with that of a Tobii 1750 interfacing with Matlab via Talk2Tobii and Smart-T. Their results suggested that more than 95% of samples showed a discrepancy of under 100 msec. Although important, this error is less of interest in the present case, as there is no reason to believe that this error might vary systematically between individuals, when the hardware setup used is kept constant.

THE PRESENT STUDY

In the present study we concentrate on the first two parameters of data quality, as these have received little attention hitherto in the literature. The first is precision, the consistency in the reported POG between samples. And the second is robustness, how broken or fragmented contact with the tracker is during recording.

In Wass, Smith, and Johnson (2013; see also Shic et al., 2008, 2009), we have previously examined how precision and robustness can affect the accuracy of the automated parsing techniques used to identify fixation durations—that is, the time intervals between saccadic eye movements (see Table 1). Our analyses suggested that robustness strongly influences fixation durations as returned by standard, commercially available fixation duration parsing algorithms ($r = .66/.47/.19$ in the 6-/12-month-old/adult samples we examined): More robust data are associated with apparently longer fixation durations. We also identified strong relationships between precision and the fixation durations returned by standard fixation parsing algorithms ($r = -.67/- .14/- .47$ in the 6-/12-month-old/adult samples we
examined): Less precise data are associated with apparently shorter fixation durations. Through simulations and hand-coding, we identified possible underlying causes for these relationships (Wass et al., 2013). To our knowledge, however, no previous research has investigated whether similar confounding influences can be identified between data quality and other key dependent variables in eye-tracking analyses, such as proportion looking to areas of interest (AOIs) and look/visit durations and saccadic reaction times (see Table 1).

We have also investigated whether systematic differences in precision and robustness can be identified between clinical and nonclinical populations. One study compared eye-tracking quality in data from 6- to 11-month-old infants at high vs. low familial risk of developing ASD (Wass, Jones, Gliga, Smith, Charman, Johnson, & BASIS team, under review). Although precision did not vary between populations, non-significantly (but consistently across two independent cohorts) lower robustness was identified in data from high-risk (ASD) relative to low-risk infants. However, to our knowledge, no previous research has examined whether eye-tracker data quality varies systematically as a result of factors such as age and testing time, as well as other aspects of behavior, such as fidgetiness during tracking.

Our aims for the present study were therefore twofold. First, we wished to examine whether systematic differences could be identified in data quality from different individuals—for example, whether younger individuals tend to show lower quality tracking. Second, we wished to explore how data quality might influence other key dependent variables in eye-tracker analyses. Previously, we have shown that lower quality data can create the impression of shorter fixation durations. Here, we examine whether data quality can relate to three other dependent variables, namely proportion looking to particular AOI (e.g., to the eyes relative to the mouth in a face), to reaction times and visit/look duration.

To address these questions we have conducted a number of analyses based on preexisting datasets, collected from different populations, on different eye-trackers and during the administration of different eye-tracking paradigms. For each analysis, we have laid out our specific predictions and hypotheses, as we describe case by case below.

The analyses presented are in six parts. In analyses 1–3, we evaluate how data quality varies systematically between and within individuals. We considered this important because, whereas some degree of random measurement error is inevitable, it leads to a risk of false-negative but not of false-positive findings—whereas systematic measurement error leads to a risk of false-positive findings. The three specific factors we examined are age (analysis 1), testing time (early vs. late in the testing session’ analysis
2), and noncognitive factors such as the amount of movement during recording (analysis 3).

In analyses 4–6, we evaluate how data quality may systematically influence a number of key dependent variables in eye-tracking analyses. Previously we have shown that participants from whom worse quality tracking is obtained give the impression of showing shorter fixation durations. Here, we examine how data quality might lead to systematic biases on other parameters, namely proportion looking reported to AOI (analysis 4), reaction time latencies (analysis 5), and first look/visit duration (analysis 6).

**METHODS**

**Participants**

Data included in these analyses are taken from two sources. The first is an ongoing cross-sectional study that comprises a series of laboratory assessments administered to 9-, 12- and 15-month-olds typically developing infants (unpublished data). The second is an ongoing randomized-controlled study examining the training of attentional control in 9-month-old infants (Forssman, Wass, & Leppänen, 2014b). All of the data reported here are taken from the first visit in this repeat-visit study, before any training was conducted. For both studies, the eye-tracking data included in this paper were recorded concurrently with electrocardiogram measurements, which have not been included in this report. Analysis 1 also includes a comparison sample of data from typical young adults that was collected specially for this study.

**Stimuli**

Analyses 1, 3, 5, and 6 are based on a gaze following experiment that uses stimuli taken from Condition 1 (the eye contact condition) of Senju and Csi-bra (2008). Twelve 8-sec clips were presented. In each video, an actress seated at a table began each trial with her head lowered. 2,400 msec after the start of each trial she raised her head and looked directly at the camera. 4,200 msec after the start of each trial, she looked at one of the two objects; her gaze remained on that object until the end of the trial. Figure 9a illustrates the screen layout and shows the AOI on which analyses were based.

Analyses 2 and 4 are based on a visual paired comparison task featuring a number of different faces (taken from Lundqvist, Flykt, & Öhman, 1998; see Figure 6a,c, e.g.). All data included in this study are taken from the familiarization phase of this experiment. In this phase, a single face was presented on-screen continuously until 10 sec of accumulated looking
time data had been obtained. Figure 6 illustrates the AOI used in our analyses.

**Stimulus presentation and data recording**

All stimuli were presented on a 23-inch monitor integrated with a Tobii TX300 eye-tracker (Tobii Technology, Stockholm, Sweden). Infants were seated on their caregiver’s lap at circa 60 cm viewing distance from the monitor. Stimuli were presented using Matlab, Psychtoolbox, and the Talk2Tobii toolbox.\(^1\) This toolbox allows for a live gaze-contingent interface via Matlab during stimulus presentation, but uses the default manufacturer-supplied algorithms for pupil, corneal reflection, and face identification during tracking. Data quality issues should be identical to those encountered using Tobii Studio.

Tobii have also developed an infant illumination mode,\(^2\) which incorporates a number of changes to the illumination method as well as to the image processing techniques used for infant data. This was not used in this sample. However, data shown in Figure S1 were recorded with this feature and suggest that similar challenges remain even with this improved illumination.

A typical calibration procedure was used (see Table 1). This involved showing the infant an attractive figure sequentially in five locations on the screen. If the first calibration was not successful (i.e., one or more calibration points were missing), the calibration was repeated at least two times to attain satisfactory calibration for all five locations. If one or more calibration points were missing after three attempts at recalibration, the final calibration outcome was accepted and the experiment was started. In other research, we have analyzed the number of missing calibration points relates to the degree of correspondence between eye-tracking and video-based analyses (Leppänen, Forssman, Kaatia, Yrttiaho, and Wass, 2014). In some systems, it is additionally possible to quantify the accuracy of the calibration points obtained during recording, but this has not been applied in the present study.

**Algorithms for analyzing eye-tracking data quality**

To analyze data quality, we wrote new data analysis algorithms in Matlab. These algorithms, which are amenable for use with data obtained from

---

\(^1\)[http://psy.ck.sissa.it/t2t/index.html]

any eye-tracker, are available for download.\textsuperscript{3} Below is a detailed description of how the two parameters of data quality on which we report in this study, precision and robustness, were calculated.

**Precision**

Previous research on both adults (Blignaut & Beelders, 2012; Holmqvist et al., 2011) and infants (Frank et al., 2011; Morgante et al., 2012) has analyzed eye-tracker accuracy by recording eye gaze position as the eye views a series of stationary targets presented at different places on the screen (akin to a post hoc calibration). If we assume that the infant is looking at the calibration point presented, any difference between the position where the target is presented and the position of gaze reported by the eye-tracker is the error involved in eye-tracking. From this error, two measures are calculated—spatial accuracy (or offset) and precision (or consistency). Figure 3 illustrates the difference between the two. It shows two samples recorded during a post hoc calibration check. In each case, the position of the calibration target has been plotted in red. As in all calibration sequences, it is assumed that the position of this target (normally a small, salient target on an otherwise blank screen) corresponds to the actual POG of the participant. The POG reported by the eye-tracker is, in each case, drawn in blue. Sample (a) shows high consistency (stability) in reported POG but a large stable offset between reported POG and assumed actual POG. It therefore suggests low spatial accuracy but high precision (consistency). Sample (b), in contrast, shows markedly more inconsistent reporting of POG between samples. It therefore suggests low precision.

We wished to develop a technique for measuring precision that can be applied post hoc to any previously recorded data—even that for which post hoc calibration data are unavailable. In Wass et al. (2013), we used within-fixation variance to estimate precision. However, this technique requires fixation parsing first to be performed; in contrast, the technique presented here requires no prior calculations.

Precision was calculated in three stages. First, the raw sample-by-sample \(X\) and \(Y\) screen coordinates obtained during tracking were exported from the recording software (e.g., Tobii Studio or E-prime) originally used and imported into Matlab. Second, data were smoothed using a simple down-sampling procedure: They were chunked into consecutive window-sized segments using a 100 msec window size, and a single median average was calculated per window. For windows in which fewer than 50\% of samples (i.e., 15 samples in 100 msec window with 300 Hz data) were

\textsuperscript{3}https://www.mrc-cbu.cam.ac.uk/people/sam.wass/
Figure 4. Demonstration of precision and robustness calculations. For precision calculations (a, b), the red line shows the rough data, the blue line shows the down-sampled data, and the black line shows the difference between the two, calculated iteration by iteration. The median difference was taken as our precision measure. For robustness calculations (c, d), two 30-sec samples of 50 Hz data are shown. The blue diamonds show the raw data; the red circles show the data for which a successful interpolation could be performed. The yellow bars, along the x axis, show the fragments of usable data obtained.
available, the entire window was returned as blank and excluded from further calculations. Third, the precision of the raw data was calculated by analyzing the average difference between the down-sampled and the unfiltered data across all samples obtained.

Figures 4a,b illustrates these analyses. Figure 4a shows an example of “high precision” data and 4b an example of “low precision” data. For each, the rough (unsmoothed) and the smoother gaze data have been drawn; the black lines (with a separate \( y \) axis) show the difference between the smoothed and unsmoothed data. It can be seen that the “low precision” data show more “jitter” or variation from sample to sample; this corresponds to a larger differences (drawn black) between the smoothed and unsmoothed data. (Video-based analyses confirm that this does not arise from oculomotor instability in the participant, but rather from sampling inaccuracy in the eye-tracking device.) The median difference between filtered and unfiltered data was calculated as an estimate of precision. A higher value of the precision metric therefore represents less precise tracking. The calculation was performed separately for the \( x \) and \( y \) gaze coordinates and then averaged.

**Robustness**

Robustness is calculated as the relative proportion of periods of data presence vs. absence during recording. There are two causes of data absence: First, the participant is not looking at the screen; second, the participant is looking at the screen but the eye-tracker is failing to detect it. As previously described, the eye-tracker aims to perform three separate identifications per frame to perform POG parsing: pupil, corneal reflection, and head position. An inability to perform any of these identifications accurately can lead to the eye-tracker returning null values for that frame, leading to periods of data loss ranging from a single sample through to periods of several seconds (see Figures 2, 4 and S1).

Previous discussions of robustness have reported the total proportion of data within the testing session for which data are available (Holmqvist et al., 2011). Following previous research (Wass et al., 2013), we have taken a different approach and reported the duration of mean usable data fragments obtained. This allows us to differentiate between (1) cases in which the participant showed unbroken looking during the first half of a trial, followed by completely absent data for the second half, and (2) instances in which the infant was looking continuously throughout the trial but contact with the eye-tracker flickered on and off throughout the trial. Figure 4c,d illustrates the difference between the two. Both of these cases might manifest as, say, 50% of data available, although the effect
of robustness on key dependent variables in eye-tracking data would be very different between the two cases. Calculating average fragment duration as a measure of robustness allows us to differentiate between the two. See Figure S2 for a detailed comparison of how the two measures interrelate.

Average fragment duration was calculated in the following manner. First, raw data were exported from the recording software originally used (e.g., Tobii Studio) and imported into Matlab. Second, interpolation (see Table 1) was performed to give a more accurate estimate of the amount of usable data available for each individual. The interpolation algorithm became active once a gap in the data was located. If the gap in the data was more than 150 msec, no interpolation was attempted. If the data gap was <150 msec, data were linearly interpolated until the first usable data sample was reached. The figure of 150 msec was selected as our experience suggests that this is the minimum approximate time beneath which no complete saccade-fixation-saccade sequences will be present (e.g., Wass et al., 2013). However, it is possible that blinks will last <150 msec (Holmqvist et al., 2011).

Following interpolation, the velocity was calculated between the last interpolated sample and the first sample after interpolation; if the velocity change was above a threshold of 35° sec\(^{-1}\), it was judged that the data before or after the interpolated sample were insufficiently accurate or that a saccade may have taken place during the interpolated sample, and the interpolation was rejected (see Wass et al., 2013 for detailed motivation). Third, the mean duration of the usable data fragments was calculated, as well as frequency distributions of all usable data fragments obtained.

Our algorithms also return two other robustness measures: First, the total proportion of available data and second a frequency distribution showing how the available data fragments were distributed. Figure S2 illustrates the results of these analyses in a sample dataset. It includes a scatterplot showing how average fragment duration relates to proportion of available data samples.

The strength of the relationship between these two parameters of data quality may vary between eye-trackers. In previous research, based on data from a Tobii 1750, we conducted similar (not identical) calculations to estimate how robustness and precision covary within a dataset and reported low correlations (average \(r = -.12;\) Wass et al., 2013). In the present dataset (based on data from a Tobii TX300), however, these relationships were found to be markedly higher. In one sample we analyzed, for example, they were found to be as high as \(r (49) = -.58, p < .001.\) The reasons for these differences may be to do with different tracking method-
ologies used by different eye-trackers. We have reported both dimensions of data quality separately in the analyses that follow.

RESULTS

Our results are presented in six parts. In analyses 1–3, we evaluate how data quality varies between and within individuals as a function of age, testing time, and gross behaviors during tracking. In analyses 4–6, we evaluate how data quality relates to a range of common eye-tracking variables, namely proportion looking to AOI, reaction time latencies, and first look/visit duration.

In the analyses that follow, unless otherwise stated any group mean comparison reported was calculated using an independent samples t-test (two-tailed), and any bivariate correlation was calculated using a Pearson’s product moment correlation.

Analysis 1—How does data quality vary between infants and adults?

To assess how data quality varies between infants and adults, we analyzed data obtained during the presentation of an identical series of short (8-sec) video clips to 43 typically developing 9-month-old infants and six adults. We predicted that tracking data would be less precise and robust in infants than in adults.

1. Precision: Figure 5a shows a scatterplot of how precision was distributed in our sample (higher value = less precise tracking, units \( \times 10^{-3} \)). As can be seen, infants showed less precision \((M = 7.60, SEM = 4.1)\) than adults \((M = 4.30, SEM = 3.4)\), although the difference was only marginally significant, \(t(47) = 1.89, p = .066.\)

2. Robustness: Figure 5b shows a scatterplot of our robustness metric. As can be seen, infants also had less robust data, with fragment duration (period of sustained contact) being shorter in infants \((M = 1.99, SEM = 1.3)\) than adults \((M = 3.81, SEM = 1.75)\), \(t(47) = 3.06, p = .004.\) (The Levene’s test for equality of variance was met, \(p = .56.\))
Analysis 2—How does data quality vary between early and late in a testing session?

In addition to analyzing how eye-tracker data quality varies between individuals as a function of factors such as age (Analysis 1) and risk factors, we also examined how data quality varied between early and late in a testing session. Figure 5 shows scatterplots illustrating these variations:

- Top—Scatterplots showing how data quality varies between infants and adults. Each data point represents the mean values obtained for one individual; (a) precision, (b) robustness.
- Middle—Scatterplots showing how data quality varies between individuals over time. Two datapoints are obtained per individual, one recorded early and one late in the session; (c) precision, (d) robustness.
- Bottom—Scatterplots showing the relationship between head movement and data quality; (e) precision, (f) robustness.
group status (clinical vs. nonclinical; Wass, Jones, Gliga, Smith, Charman, Johnson, & BASIS team, under review), we also wished to assess how data quality varies within individuals over time. To assess this, we analyzed data obtained during the presentation of a visual paired comparison task to 16 typically developing 9-month-old infants. Figure 5c,d shows scatterplots illustrating the results of these analyses. This task was presented in different blocks, interspersed with other tasks in a heterogeneous testing battery that lasted circa 15 min in duration. We compared eye-tracking data obtained during the presentation of the first block, which was presented typically 4–5 min into the testing session, with the last block, which was presented typically 12–13 min into the testing session. The data plotted are averages of the trials within each block. We predicted that lower quality data would be obtained later in the testing session.

1. Precision: As can be seen in Figure 5c, and as expected, data obtained from infants were significantly less precise early in the session ($M = 4.21, \ SEM = 0.34$) than later ($M = 4.80, \ SEM = 0.38$), $t(15) = 3.00, p = .008$). It is noteworthy that, despite this fall-off in performance as testing continued, test–retest reliability was high ($r = .80$), suggesting that precision is stable as an index of individual differences.

2. Robustness: As the data shown in Figure 5d for robustness were not normally distributed, as indicated by a one-sample Kolmogorov–Smirnov test, $Z(18) = .24, p = .006$), nonparametric tests were used. The results indicated that the median fragment durations from early in the session ($M = 1.5, \ SEM = 0.6$) and later ($M = 0.9, \ SEM = 0.2$) were not appreciably different, Wilcoxon signed-rank test, $Z = -1.1, p = .26$. Thus, the robustness of tracking was not found to decline across the session.

Analysis 3—How does data quality vary contingent on head movement?

To explore how data quality relates to other parameters of behavior, such as subject movement during data recording, we analyzed head movement velocity. All Tobii trackers record the positions of the two eyes independently in 3D space relative to the screen at all times during recording. These data were exported from the data analysis software into Matlab where the following processing steps were applied. First, because visual inspection of the data suggested they were vulnerable to occasional egregious artifact, the velocity of all samples was calculated separately for the
three dimensions (x, y, and z) and samples showing a shift in position of >1.25 cm between individual iterations (in 300 Hz data) were excluded on the basis that such values were above the maximum velocity with which an infant could move their head and were therefore likely to be artifactual. Second, data were down sampled to 30 Hz using a median moving window. Third, three-dimensional head velocity data were collapsed into one dimension. Fourth, the median head velocity throughout the whole trial was calculated.

Our analyses used the same infant data as that used in Analysis 1, in which 43 typically developing 9-month-olds viewed a series of short video clips. Head velocity and data quality were calculated independently for each trial, and then, a single per-participant average was calculated. Figure 5e,f shows the results of these analyses. A Kolmogorov–Smirnov test showed that head movement data were not normally distributed (Z (42) = .28, p < .001), and so nonparametric statistics (Spearman’s ϱ) were used.

We hypothesized that increased head velocity would be associated with lower eye-tracking data quality, due to an increased likelihood of problems being encountered at various stages of the data processing. Consistent with our hypothesis, a significant negative correlation was observed between head velocity and robustness, ϱ (42) = −.41, p = .006, suggesting that increased head movement was associated with less robust data. For head movement and precision, the relationship was not significant, ϱ (42) = .15, p = .32.

We considered the possibility that, as both head movement and eye gaze data were obtained from the same eye-tracker during recording, pervasive data quality issues might influence all aspects of data quality and therefore lead to a circular relationship. However, head movement data were only calculated (because they were only available) for those samples for which eye-tracking data were also obtained. Furthermore, although it is conceivable that less precise estimation of head position and less precise estimate of POG could have a shared cause (imprecise face estimation—as opposed either to imprecise pupil or glint estimation which would affect POG but not head position estimates), we found that the relationship of head movement to precision was less strong than that of head movement to robustness, which is seemingly free of this possible shared artifact. This suggests that shared artifact is probably not the cause of the relationship we observed between head movement and data quality.

To assess the degree to which the differences observed in Analysis 2 (on change in precision early vs. late in the session) were attributable to differences in head movement, the change in head velocity early vs. late in the testing session was calculated for these data using an identical procedure to that used for the data quality analyses described above. A Kol-
mogorov–Smirnov test suggested that all data were not normally distributed ($Z(18) = .33$, $p < .001$) and so nonparametric tests were used. Although median ($SEM$) head velocity was marginally lower at the early testing session ($2.2 \times 10^3 [1.0 \times 10^3]$) relative to the later testing session ($2.4 \times 10^3 [0.5 \times 10^3]$), a Wilcoxon signed-rank test suggested that this difference was not significant ($Z = -.50$, $p = .62$). This suggests that the decrease in precision observed in Analysis 2 between the earlier and later testing sessions was not attributable to an increase in head movement.

**Interim summary**

The results presented in analyses 1–3 suggest that data quality varies systematically between individuals, as a function of age, and within individuals, as a function of testing time and amount of movement during recording. In the following three analyses, we examine how differences in data quality may relate to a range of commonly reported dependent variables in eye-tracking analyses.

**Analysis 4—Data quality and proportion looking to Areas of Interest**

Probably the most frequently reported dependent variable in eye-tracker analyses is proportion looking to AOIs. For example, it is common to report on individual differences in looking behavior toward faces: Static faces are shown, and the different features (eye, nose, and mouth) are marked using rectangular AOIs, to assess whether participants look more on to one feature than to another. This analysis is used, for example, to examine differences in gaze behavior associated with ASD. We were concerned to evaluate possible relationships between data quality and proportion looking to AOIs.

In the following examples, the AOIs used are defined narrowly and are immediately contiguous, to maximize the possible confounding influence of data quality on results. It is important to note that, although widespread, designing experiments with AOIs defined in this way is not in accordance with the practices recommended by the eye-tracker manufacturers, who recommend defining AOI sizes based on the expected accuracy performance of the eye-tracker and the study design.\(^4\)

Our analysis was in two parts. First, we conducted a simulation in which a sample of eye-tracking data was artificially manipulated to simulate the effects of lowering tracking precision, to assess how this manipulation affected results. As a result of this simulation, we generated a prediction for the relationship that we would find in a sample of genuine

---

\(^4\)http://youtu.be/67HaqnmQ7Pw?t=18m54s
eye-tracking data. Second, we then evaluated this prediction in a analysis of real eye-tracking data.

**Simulation and prediction**

To generate a prediction as to how raw data quality might relate to proportion looking to AOIs we first conducted a simulation to assess the effect of adding inaccuracies of up to 2° to our data. Although this level of noise is greater than that commonly claimed by eye-tracker manufacturers for adults, our own observations (see, e.g., Figs 2, 3 and S1) suggest that it is not uncommon in infants. For example, Sample 2 in Figure 2 shows a continuous measurement error of approximately 15% on the Y dimension, corresponding to 3.6° in a monitor subtending 24°.

Figure 6 shows a graphical overview of this simulation, which is based on a single trial that was excerpted from a typically developing 9-month-old infant viewing a static picture of a face for 10 sec. First, the raw gaze sample was plotted (Figure 6a), and the results were analyzed for proportion looking time to eyes, nose, and mouth (Figure 6b). These were found to be 0.77 (eyes), 0.03 (nose), and 0 (mouth). Then, the precision of the data was artificially manipulated by adding, iteration by iteration, random noise up to 2° to the X and Y gaze coordinates, to simulate the effects of low precision tracking (Figure 6c). Following addition of the simulated noise, the data were reanalyzed for proportion looking time to eyes, nose, and mouth (see Figure 6d). The new proportion looking times observed were 0.31 (eyes), 0.12 (nose), and 0.04 (mouth).

As a result of this simulation, we predicted therefore that the general effect of this artifact would be that of a flattening out of the distribution of gaze. In cases, such as face viewing, where gaze is naturally concentrated in one area (the eyes; Johnson, 2010), the effect of low precision would be to create the appearance of a reduced concentration of gaze in that area and an apparently increased concentration in other areas.

**Evaluation of prediction from simulation with real data**

To evaluate this prediction, we analyzed a corpus of viewing data obtained from 22 typically developing 9-month-old infants. Sixty-five usable trials were available for analysis. Figure 7 shows scatterplots illustrating the results of these analyses. Consistent with our predictions, we

---

found that less precise data were associated with less looking time to eyes: $r (64) = -0.28, p = .03$. A Kolmogorov–Smirnov test suggested that results for proportion looking to nose and mouth were not normally distributed ($Z (64) = .25, p < .001$), and so Spearman’s $\rho$ was calculated to assess the relationship between precision and proportion looking to nose and

![Figure 6](image_url) Ten seconds of viewing data from a 9-month infant. (a, b) Shows the original data sample, (c, d) shows the same gaze data sample, but subjected to a “low precision” simulation by adding Gaussian noise. In each case, (a, c) shows the gaze data superimposed on the image that was being viewed. Looking time to different areas of interests (AOIs) has been calculated; gaze coded as to the eye region is drawn red, to the nose region is drawn green, to the mouth region is drawn blue, and to other areas is drawn black. (b, d) Shows the proportion looking times recorded, subdivided by AOI.
A positive relationship was observed: \( \rho (64) = .29, p = .02 \). This suggests that, consistent with our hypothesis, lower quality data tend to be associated with a lower proportion of looking time recorded to the eyes and a higher proportion of looking time recorded to the nose and mouth. The analysis was repeated with number of samples recorded to eyes and nose/mouth (rather than proportion of samples) as the dependent variable. Figure S3 shows the results. Outcomes were similar, although the relationship between precision and number of samples to the nose/mouth was not significant (\( p = .11 \)).

**Analysis 5—Data quality and reaction time latencies**

Another common variable in eye-tracker analyses is look latency—that is, the time delay between an event taking place and the participant first looking toward a particular area of the screen.

Again, our analysis was in two parts. First, we conducted a simulation in which a sample of eye-tracking data was artificially manipulated to simulate the effects of lowering tracking robustness, to assess how this manipulation affected results. As a result of this simulation, we generated a prediction for the relationship that we would find in a sample of genuine eye-tracking data. We then evaluated this prediction in a full cohort analysis.
Simulation and prediction

Figure 8 shows a simulation we conducted to predict how data quality might influence results recorded using reaction time measures. Reaction time is calculated as the latency between an event taking place and the first recorded POG within the response window—which is the technique used in Tobii Studio, for example, to calculate time to first fixation (Tobii, 2012). Figure 8a shows a single trial of a typically developing infant’s looking behavior during a 1-sec time window. The black line shows the position of the target, and the red line the position of gaze of the infant viewing the target. The latency between the shift in the target location and the first valid position of gaze recorded within the areas of interest is the oculomotor reaction time (in this instance 300 msec). (b) Shows the same data sample, but subjected to a robustness simulation.

**Figure 8** (a) Gaze data from a typically developing 12-month-old infant during a 1-sec gaze window. The black line shows the position of the target, and the red line the position of gaze of the infant viewing the target. The latency between the shift in the target location and the first valid position of gaze recorded within the areas of interest is the oculomotor reaction time (in this instance 300 msec). (b) Shows the same data sample, but subjected to a robustness simulation.
position of a target, which at one point during the trial moves from one position on the screen to another. Figure 8b shows exactly the same data, but subjected to a stimulation in which random segments of the data have been removed to vary robustness. Exactly the same coding criteria were applied: Reaction time is recorded as the latency between the event taking place and the first valid sample looking to the target. In the complete data sample, the RT is recorded as 300 msec; in the second (identical but less robust) sample, it is 480 msec.

As a result of this simulation we predicted that less robust eye-tracker data would be associated with increased reaction time latencies and a decreased likelihood of a look being recorded within a particular time window.

**Evaluation of prediction from simulation with real data**

To assess these predictions we analyzed data from a corpus of 92 typically developing 9- to 15-month-old infants viewing a gaze following paradigm (Senju & Csibra, 2008). The dependent variable analyzed was latency to respond to direct gaze. We predicted that robustness would relate negatively with response latency. To assess this we calculated the robustness of data quality obtained, trial by trial, during the direct gaze period (starting 2,400 msec and ending 4,200 msec into each trial). We also calculated the latency between the start of the direct gaze period and the first POG within a rectangular window drawn around the face. Trials in which the position of gaze was within the response area at the start of
Figure 10  (a, b) Shows 15 sec of viewing material from a 6-month-old infant viewing a static image. Six different areas of interests (AOIs) have been drawn onto the screen in different colors. Gaze data obtained during that trial are also drawn onto the screen. Raw gaze data (gray) have been corrected, and gaze classified as being within a particular AOI is drawn in the color of that AOI. (b) Shows the same data with time on the x dimension. Sections during which the recorded POG was within the pink AOI have been cultered pink, and so on. The bars above the plot show the results of two coding schemes applied to the data. In “Coding A”, the algorithm treats a look to an AOI as ending when the first valid sample is obtained outside of that AOI. In “Coding B”, the algorithm treats a look to an AOI as continuing only while gaze data are being recorded within that AOI. The instances of marked discrepancy between the two coding schemes are marked with red dashed rectangles.
the window were excluded. Figure 9 contains a scatterplot illustrating the results of this analysis. The relationship was found to be negative as predicted ($r (91) = -0.22, p = .035$), suggesting that shorter mean fragment duration (i.e., less robust data) was associated with longer response time latencies.

Analysis 6—Data quality and first look/visit duration

Another common dependent variable in eye-tracker analyses is “visit duration” or “look duration”—the time spent by the infant looking toward each of the different areas (AOIs) on the screen. One particularly popular dependent variable is “first look duration”—the duration of the first look recorded toward a particular AOI.

A particular challenge with these analyses is how best to cope with the frequent, and often lengthy, periods of data loss encountered in infants. Figure 10 illustrates this problem. Is a look considered to have ended when the first valid sample is obtained outside that AOI (“Coding A” in Figure 10)? Or is a look registered only as long as gaze data are recorded within that AOI (“Coding B”)? Or is a look during which any data samples are not present to be rejected and not stored? These are similar to the problems discussed in Wass et al. (2013) regarding the relationship between robustness and fixation duration.

In this experiment, we evaluated “Coding A”, which is the technique used to compute “Visit duration” in Tobii Studio (Tobii, 2012). We predicted that we would observe a positive relationship between first look
duration and fragment duration (i.e., longer look durations associated with more robust data).

**Evaluation of prediction with real data**

To evaluate this prediction, we analyzed data from a corpus of 49 typically developing 9- to 15-month-old infants viewing the same gaze following experiment as in Analysis 5. Analyses were conducted using the AOI illustrated in Figure 9. Robustness was calculated on a per-participant basis as reported previously. Prior to calculating look duration, data were first smoothed using a 100-msec median moving window.

The first dependent variable analyzed was first look duration. This was coded as the duration (in seconds) of the first look to either of the objects after the actress had looked to one of the two objects. The end of a look was coded as the first reported POG outside the AOI, including instances of data loss. Figure 11a shows the results of this analysis. Longer fragment duration (more robust data) was found to be associated with longer first look duration ($r(49) = .30, p = .04$).

The second dependent variable analyzed was the average duration of all looks recorded to AOIs. Three AOIs were recorded: the two objects and the face of the actress. Looks to other areas of the screen were excluded from this analysis. Figure 11b contains a scatterplot illustrating the results of this analysis. The average (SEM) total number of looks recorded per participant was 65.9 (23.4). The average (SEM) duration in milliseconds of all looks to AOIs was 810 (403) msec. Longer fragment duration (more robust data) was associated with longer average look duration to AOIs ($r(49) = .85, p < .001$).

**DISCUSSION**

Systematic variations in data quality between and within individuals

From our analyses we believe that two conclusions can be drawn. First, test–retest reliability for most measures was high, suggesting that data quality is a relatively stable metric. Second, data quality varies systematically both between and within individuals, as a function of age, testing time, and fidgetiness. For age, our analyses suggest that longer usable fragment durations are obtained from adults relative to infants. Although only marginally significant (with a small sample size ($N = 6$) in our adult sample), our findings also suggest that less precise tracking data are obtained from infants relative to adults. We also examined how data quality varies early vs. late in the testing session. We found
that the precision of eye-tracker data is markedly lower later in the session (see also Blignaut & Beelders, 2012). We also examined how data quality relates to fidgetiness, as measured via head velocity. We found that greater head velocity was associated with lower robustness, although no significant relationship with precision was identified. Although we did not examine systematic differences between clinical and nonclinical populations here, in previous research we have compared data obtained from infants at high risk of ASD with infants at low risk (Wass, Jones, Gliga, Smith, Charman, Johnson, & BASIS team, under review). We found that although precision did not vary between populations, there was nonsignificantly (but consistently across two independent cohorts) lower robustness in data from high-risk (ASD) relative to low-risk infants.

**Relationship of data quality to key dependent variables in eye-tracking analyses**

In the analyses presented here, we have assessed whether data quality may influence a number of key dependent variables in infant eye-tracking experiments. From analysis 4, we concluded that lower precision data can appear to suggest a reduced likelihood to look at the eyes in a face relative to the mouth. We predict that the same pattern ought to be present in any analysis in which infants naturally look to one area of the screen relative to another. The effect of lower precision would be that of a “flattening out” of the distribution across the different AOIs.

From analysis 5, we concluded that less robust tracking may manifest as slower reaction time latencies (e.g., time to first fixation). In Leppänen et al. (2014), we examine this issue in more detail. From analysis 6, we concluded that less robust data can manifest as shorter first look/visit duration. It should also be noted that the effect sizes observed varied markedly between analyses. The effect identified in analysis 5 was significant but relatively weaker ($r = -.22$), whereas those in analysis 6 were extremely strong ($r = .85$). Our findings here can be compared with our previous findings on how data quality relates to fixation durations as parsed using the standard dispersal-based parsing algorithms supplied with most eye-trackers (Wass et al., 2013). There we found that participants with poorer quality data were returned as showing markedly shorter fixation durations than those who did not.
Suggestions for further work

There are a number of limitations to these analyses. As with all correlations, the relationships we have documented are vulnerable to the possibility that the observed relationship may be mediated by some unobserved third factor. For example, we cannot entirely distinguish between two explanations: (1) Frequent movement during tracking causes the appearance of decreased looking to the eye region due to increased artifact and (2) poorer quality tracking data are associated with a genuinely reduced likelihood to look to the eye region because both parameters are attributable to some third factor that has not been tracked in the current analysis. However, we also conducted simulations throughout the paper, by experimentally manipulating parameters of eye-tracker quality to measure the effects on eye-tracker dependent variables. The fact that the simulations were consistent with the results of our correlational analyses goes some way to precluding possibility #2.

What can be done to address these issues? First, it should be standard to report on how data quality varies between populations and between conditions—particularly if the experiment is based on examining behavior during response windows. The Matlab scripts we have used in this study are available for download (see link above). Calculating data quality also allows for the inclusion of data quality as a covariate in analyses. Future work should also investigate in more detail how differences in data quality relate to extrinsic factors such as the accuracy of the calibration obtained prior to recording.

Second, our analyses suggesting that data quality (both robustness and precision) varies as a function of time suggest that possible order effects should be tracked with care. Different conditions of experiments should be presented interleaved in blocks rather than consecutively. It is also useful to interleave short, engaging movie clips within experiments, which helps to maintain engagement.

Third, more care should be taken in devising data processing techniques that are specifically designed to be independent of variation in data quality. Unfortunately many of the “off-the-peg” analysis programs available with commercial eye-trackers do not do this. In user-defined data processing, judicious use of techniques such as interpolating and smoothing can help here, although no “one-size-fits-all” solution is available: Optimal parsing techniques may vary between different analyses. For the analysis we conducted in Analysis 4, examining the relationship between precision and proportion looking to AOI, filtering data using a median moving window would be effective for reducing error, as would designing test stimuli with noncontiguous AOI (Holmqvist et al., 2011). For the
analyses in Analysis 5, looking at robustness and reaction time latencies, it is possible apply interpolation to those periods during which the position of gaze remains within either target area, while rejecting sections in which the position of gaze changes during the lost data segment. Those sections with missing data segments should then be excluded from the reaction time analyses. Matlab algorithms that perform these functions, and that are available for download, are described in Leppänen et al. (2014). However, for the analyses in Analysis 6, looking at the relationship between robustness and first look/observation duration, no easy solution exists (see Figure 10).

Similar developments are already ongoing within the field of fixation duration parsing. Tobii have developed the IV-T filter which allows users to vary a number of key parameters associated with fixation parsing—including interpolation across gaps, options for dealing with loss of one eye, noise reduction (median or mean of a three-sample moving window), variable velocity thresholds (30° sec^{-1} over a 20-msec period), and the option to merge adjacent fixations (<0.5°) separated by a brief period of lost data (<75 msec; Olsen, 2012; Olsen & Matos, 2012). Features such as the Velocity Chart afford an easy means of visualizing the data, and researchers can process their data multiple times post hoc while varying the settings to identify the (often substantial) way in which these manipulations can influence results. In Wass et al. (2013), we also presented Matlab fixation parsing algorithms that we designed to perform fixation parsing independent of data quality—although in subsequent papers we have found that relationships with data quality remain (Wass, Jones, Gliga, Smith, Charman, Johnson, & BASIS team, under review). Alternative methods include semi-automated techniques for identifying fixations by hand (Rodriguez Saez de Urabain, Johnson, & Smith, 2013).

**CONCLUSION**

We believe that the analyses presented here open the possibility that differences in eye-tracking data quality between and within individuals may cause the appearance of differences in gaze behavior, without this actually being the case.

A number of different experimental designs may be vulnerable to this confound. First, instances in which a paradigm is applied to individuals of different ages. In this case, eye-tracking data obtained from younger individuals may be of lower quality, which in turn may cause the false

---

6http://youtu.be/kgBtLAwDFRY?t=41m6s
appearance of observed differences in dependent variables. Second, instances in which gaze behavior is compared between two populations. This applies to studies comparing clinical and typical populations, to studies comparing “high-risk” and “low-risk” groups (such as infants from low socioeconomic status backgrounds, or infants with particular familiar characteristics), as well as to studies comparing groups defined by some other behavioral characteristic, such as good and poor learners. In each case, the danger is that between-group differences in factors such fidgetiness or irritability may lead to differences in eye-tracking data quality, which can cause the false appearance of differences in gaze behavior. A third possibility is that systematic differences in data quality may exist between conditions within particular experiments. This applies, for example, in instances in which certain conditions were more novel or engaging than others or presented later in the testing session. Above we have described a number of steps that we consider can help to address these possibilities.

The increased spatial and temporal resolution afforded by eye-tracking is vital in allowing researchers to address new and unanswered questions within infant psychology. The same relationships that we have identified here, between data quality to gross behavioral measures such as fidgetiness, are likely to apply to other techniques widely used with infants—including electroencephalography (EEG), event-related potentials (ERPs), and near infrared spectroscopy (NIRS). It seems also possible that differences in data quality obtained from these other recording techniques might also cause the appearance of differences on key dependent variables without this actually being the case. Every method is flawed, but the upside is that many of the data quality issues we have discussed here are relatively easy to assess, and systematic artifact can potentially be minimized through careful experimental design and/or appropriate data processing techniques.

ACKNOWLEDGMENTS

Thanks to Tim Smith, Tibor Auer, Nic Farhi, Dennis Norris, Emily Jones for advice and useful discussions. Thanks to Jason Babcock for permission to use the eye image used in Figure 1, taken from the Yarbus software. Thanks to Tommy Strandvall and Ricardo Matos from Tobii and to Susan Rose, Hayley White, and Sarah Smith for reading earlier versions of the manuscript and giving comments. This work was supported by a British Academy Postdoctoral Fellowship to the first author and by grants from the Academy of Finland (#218284) and the European Research Council (#283763) to J. M. L.


**SUPPORTING INFORMATION**

Additional supporting information may be found in the online version of this article:

**Figure S1** Further raw data samples.

**Figure S2** Supplementary analyses for our robustness measures.

**Figure S3** An identical analysis to that shown in Figure 7, but examining the number of samples recorded to eyes and nose/mouth instead of the proportion of available samples.