Received: 11 February 2021 | Accepted: 20 September 2021

DOI: 10.1111/infa.12441

INFANCY THE OFFICIAL JOURNAL OF THE INTERNATIONAL CONGRESS OF INFANT STUDIES WILEY

RESEARCH ARTICLE

Replacing eye trackers in ongoing studies: A comparison of eye-tracking data quality between the Tobii Pro TX300 and the Tobii Pro Spectrum

Yentl J.R. De Kloe[1] | **Ignace T.C. Hooge**[1] | **Chantal Kemner**[1] | **Diederick C. Niehorster**[2,3] | **Marcus Nyström**[2] | **Roy S. Hessels**[1]

[1]Experimental Psychology, Helmholtz Institute, Utrecht University, Utrecht, The Netherlands

[2]Lund University Humanities Lab, Lund University, Lund, Sweden

[3]Department of Psychology, Lund University, Lund, Sweden

Correspondence
Yentl J.R. De Kloe, Heidelberglaan 1, 3584CS Utrecht, The Netherlands.
Email: y.j.r.dekloe@uu.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 024.001.003

Abstract

The Tobii Pro TX300 is a popular eye tracker in developmental eye-tracking research, yet it is no longer manufactured. If a TX300 breaks down, it may have to be replaced. The data quality of the replacement eye tracker may differ from that of the TX300, which may affect the experimental outcome measures. This is problematic for longitudinal and multi-site studies, and for researchers replacing eye trackers between studies. We, therefore, ask how the TX300 and its successor, the Tobii Pro Spectrum, compare in terms of eye-tracking data quality. Data quality—operationalized through precision, accuracy, and data loss—was compared between eye trackers for three age groups (around 5-months, 10-months, and 3-years). Precision was better for all gaze position signals obtained with the Spectrum in comparison to the TX300. Accuracy of the Spectrum was higher for the 5-month-old and 10-month-old children. For the three-year-old children, accuracy was similar across both eye trackers. Gaze position signals from the Spectrum exhibited lower proportions of data loss, and the duration of the data loss periods tended to be shorter. In conclusion, the Spectrum produces gaze position signals with higher data quality, especially for the younger infants. Implications for data analysis are discussed.

1 | INTRODUCTION

Infant looking behavior is a widely studied subject, pioneered by Robert Fantz in the early sixties (Fantz, 1961). In 1956, Fantz published a technique to study visual preferences in chimpanzee infants (Fantz, 1956), which he later used to study visual preferences in human infants (Fantz, 1964), see Colombo and Mitchell (2009) for a brief history. The study of infant looking behavior has since been prominent in various fields of research, for example, face perception (Cashon & Cohen, 2004), face scanning (Haith et al., 1977), recognition memory (Rose et al., 1982), joint attention and social cognition (Carpenter et al., 1998), and infant learning in general (see Colombo & Mitchell, 2009, for a review). In these studies, gaze direction was coded manually from video or in real-time. This has at least two shortcomings: It is labor intensive, and it is subjective. Modern day video based eye trackers do not have these shortcomings. With an eye tracker, gaze location is determined automatically and objectively. Eye tracking is non-intrusive and can be used to study viewing behavior across the life span. It has been used to study viewing behavior in infants from as young as 2 or 3 months of age (Johnson et al., 2004; Jones & Klin, 2013), up to adulthood.

An often-used eye tracker in developmental eye-tracking research is the Tobii Pro TX300 (Falck-Ytter et al., 2015; Gomez et al., 2017; Hessels et al., 2016, henceforth TX300). This specific eye tracker is no longer being manufactured and sold. In the study we are affiliated with, the YOUth cohort study at Utrecht University[1] (Onland-Moret et al., 2020), this is problematic as one of our eye trackers recently broke down. This is also a potential problem for other cohort studies where the eye tracker is being used, like: EUROSIBS[2] (Jones et al., 2019) and EASE[3] (the EASE Team et al., 2018). The seemingly easy solution to the broken eye tracker may be to replace it with a newer model. For the TX300 the natural successor, as marketed by the manufacturer, is the Tobii Pro Spectrum (henceforth Spectrum). However, eye trackers are not all alike and may differ in, for example, the quality of the eye-tracking data it produces (as e.g., specified by the manufacturers[4]). In Figure 1, an example of differences in data quality between two eye trackers is shown.

Eye-tracking data quality is often characterized through the precision, accuracy, and data loss of the gaze position signal (Holmqvist & Andersson, 2017, p. 159–160). Precision is defined as the reproducibility of a gaze position signal from one sample to the next, assuming a stable gaze position. Precision can be operationalized as the root mean square sample-to-sample (RMS-s2s)

[1] https://www.uu.nl/en/research/youth-cohort-study

[2] https://www.eurosibs.eu

[3] https://ki.se/en/kind/the-project-ease

[4] See for example: https://www.tobiipro.com/siteassets/tobii-pro/brochures/tobii-pro-tx300-brochure.pdf and https://www.tobiipro.com/siteassets/tobii-pro/product-descriptions/tobii-pro-spectrum-product-description.pdf/?v=2.0.5, both accessed 23 November 2020.

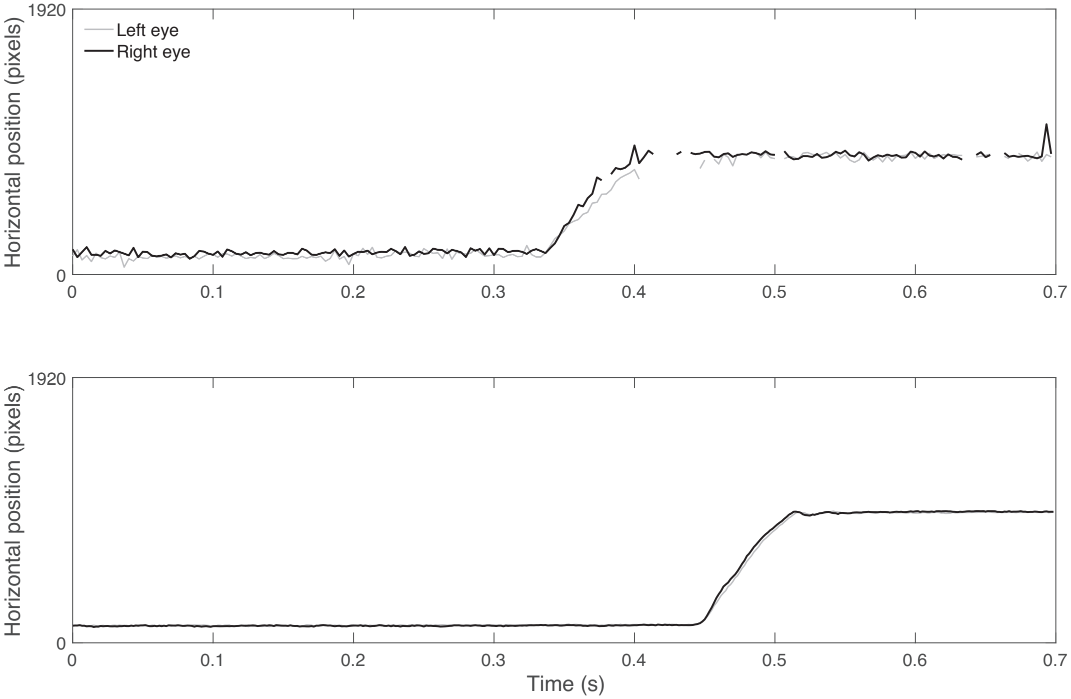

FIGURE 1 Examples of raw gaze position signals from the Tobii Pro TX300 recorded at 300 Hz (top panel) and the Tobii Pro Spectrum recorded at 600 Hz (bottom panel), from two 5-month-old infants. The gray lines are coordinates from the left eye; the black lines are coordinates from the right eye. The gaze position signals from the TX300 are less precise than those from the Spectrum, and contain periods of data loss, indicated by gaps between 0.35 and 0.5 s

deviation in gaze position signals (Holmqvist & Andersson, 2017, p. 179), where lower RMS-s2s deviation values correspond to better precision. Accuracy is defined and operationalized as the distance between the assumed true and recorded gaze position (Holmqvist & Andersson, 2017, p. 160). Smaller distances correspond to better accuracy. Data loss refers to the proportion of samples during which no gaze coordinate was reported relative to the reported sampling frequency of the eye tracker. Thus, a 250 Hz eye tracker is expected to deliver 250 gaze coordinates per second. If only 225 are delivered, this means 10% data loss is observed.

Eye-tracking data quality is known to be related to eye-tracking measures like fixation duration and number of fixations (Wass et al., 2014). Imagine that in a longitudinal study a subset of participants is measured with one eye tracker, and another subset is measured with another eye tracker. If the data quality between the eye trackers differs, this could lead one to draw conclusions that may be attributable to differences in data quality, *not* to any differences in looking behavior. In order to show the severity of this problem, we will give three examples on how data quality may affect experimental outcome measures.

Consider, for example, a researcher interested in visual perception who has conducted a study using an eye tracker. Commonly used eye-tracking measures in visual perception experiments are (average) fixation duration (Holmqvist & Andersson, 2017, p. 527) and number of fixations (Holmqvist & Andersson, 2017, p. 561). Fixations in the gaze position signals are usually determined by fixation and saccade classification algorithms (Nyström & Holmqvist, 2010). Holmqvist et al. (2012) showed that fixation classification algorithms are sensitive to the precision of the gaze position signal. They found that lower precision (higher noise levels in the data) caused average fixation duration to increase and the number of fixations to decrease for their algorithm.

In low accuracy eye-tracking data, the observed gaze position is shifted away from the true gaze position. Inaccurate fixation locations may be problematic for, for example, area of interest (AOI) analysis (Orquin et al., 2016). An example of this can be found in reading research, where the word AOIs are small and close to each other. In these cases, the undesirable situation might occur where fixations are wrongly assigned to the neighboring AOI, or to no AOI at all. As an example, Holmqvist et al. (2012) showed that AOI measures like (total) dwell time and time to first AOI hit may be related to the accuracy.

The effect of data loss on eye-tracking measures is more intricate. When data loss periods are short (e.g., <100 ms) interpolation methods may be applied (Hessels et al., 2017). When data loss periods are longer, the eye-tracking data might be unusable. This may be, for example, because an entire fixation-saccade-fixation sequence is missing, which cannot be interpolated. For such sections, one cannot compute relevant fixation- or saccade-based eye-tracking measures. Wass et al. (2013) showed that the output of corporate fixation-classification algorithms (as delivered by eye-tracker manufacturers), varies with different levels of data loss.

The examples above show that studying potential differences in eye-tracking data quality between eye trackers is crucial for (prospective) longitudinal studies, including our YOUth cohort study. The problem of replacing an eye tracker in a longitudinal study is also representative for replacing an eye tracker in between studies, setting up a multi-site study with different eye trackers, and reproducing a study with a different eye tracker. In this study, we compare eye-tracking data quality from the Spectrum with eye-tracking data quality from the TX300 and discuss implications of the differences in data quality we observe. As eye-tracking data quality is also related to the participants' age (Hessels & Hooge, 2019), we compare eye-tracking data quality between the eye trackers for three separate age groups: 5-month-old infants, 10-month-old infants, and 3-year-old toddlers.

2 | METHODS

2.1 | Participants

For this study, gaze position signals were used from a total of 1189 children around three ages: 5-months-old (ranging from 4 to 7 months), 10-months-old (ranging from 8 to 12 months) and 3-years-old (ranging from 23 to 50 months). Eye-tracking data were obtained from YOUth, a large cohort study based in Utrecht, the Netherlands (Onland-Moret et al., 2020). The participants were recruited through local municipalities within the province of Utrecht. The present study was conducted according to guidelines laid down in the Declaration of Helsinki. The YOUth study was approved by the Medical Research Ethics Committee of the University Medical Center Utrecht (protocol number: NL51465.041.14), and all participants' parents or guardians provided written informed consent prior to any assessment or data collection. The Spectrum was (for a limited time) used for eye-tracking recordings in the YOUth study. Within each age group, participants' gaze was recorded with one of the two eye trackers. Due to the longitudinal nature of the YOUth study, the same participants may occur in more than one of the age groups for the TX300 recordings. The main comparisons in this study are between eye trackers within each age group. With the Spectrum, gaze position signals from 20 five-month-old infants (M = 4.9, SD = 0.9 months, 13 females), 38 ten-month-old infants (M = 10.0, SD = 0.7 months, 23 females), and 24 three-year-old children (M = 33, SD = 7.4 months, 10 females) were collected. These were compared with gaze position signals from 490 five-month-old infants (M = 5.0, SD = 0.8 months, 277 females), 486 ten-month-old infants (M = 9.9, SD = 0.8 months, 246 females), and 131 three-year-old children (M = 31.3, SD = 5.3 months, 73 females) recorded with the TX300.

2.2 | Stimuli and apparatus

Gaze position signals were collected using the TX300 recording at 300 Hz and the Spectrum recording at 600 Hz. The resolution of the monitors attached to the eye trackers was 1920 × 1080 pixels, with physical screen sizes of 51.1 × 28.7 cm for the TX300 screen and 52.7 × 29.6 cm for the Spectrum screen. The distance between the participant and the TX300 screen was approximately 62 cm and the distance to the eye tracker was approximately 65 cm. The distance between the participant and the Spectrum screen was approximately 63 cm and the distance to the eye tracker approximately 67 cm. The distances were chosen such that the stimuli were presented at approximately the same size when expressed in degrees. Communication with the eye trackers was achieved using the Tobii SDK controlled through MATLAB running on Mac OSX Mavericks 10.9 or Linux Ubuntu 18.04. Due to a transition in operating systems from Mac OSX to Linux on 24 December 2018, a total of 633 participants were recorded with MATLAB running on Mac OSX, and 556 with MATLAB running on Linux. All recordings with the Spectrum were done on the Mac OSX operating system. Psychtoolbox (Brainard, 1997) was used for stimulus presentation.

Eye-tracking data quality was estimated for gaze position signals from a gap-overlap experiment as used in Cousijn et al. (2017). The gap-overlap experiment started with a central stimulus (an expanding and contracting clock with an average size of 2.6° × 2.6° and a maximum size of 3.5° × 3.5°). After the clock was gazed at, it started spinning with a speed of 500°/s to maintain the participant's attention. After 600–700 ms, the central stimulus was followed by a peripheral stimulus presented at 19° from the center of the screen on either the left or the right side (a yellow oval (2.6° × 2.6°)). Through varying the moment at which the central stimulus disappeared, attentional disengagement is investigated (Saslow, 1967). The central stimulus (1) remains onscreen during the presentation of the peripheral stimulus (overlap condition), (2) disappears 200 ms before peripheral stimulus onset (gap condition) or (3) peripheral stimulus onset is the same as central stimulus offset (baseline condition). Each condition consisted of 16 trials, adding up to a total of 48 trials. An additional 12 trials (4 of each condition) were added if upon online assessment less than 12 trials from a condition were valid. A trial was considered valid when the participant fixated both the central and peripheral target during their presentation. The operators sometimes quit the gap-overlap experiment before the completion of the experiment when they deemed the child to be tired, irritable or if the child for some other reason was not fit to complete the experiment. Table 1 provides an overview of the number of participants who completed 48 trials and 60 trials, as well as the number of remaining participants and their median number of completed trials per eye tracker. A Bayesian contingency table analysis conducted in JASP 0.14.1 (JASP Team, 2020), revealed that whether participants completed 48, 60, or another number of trials did not seem to differ between eye trackers (BF_{01}=26 for the null hypothesis of no difference between eye trackers). Prior to the experiment, a 5-point-operator-controlled calibration was performed (for details see: Hessels et al., 2015). The operator checked the output of the calibration provided by the Tobii SDK. If deemed satisfactory, the experiment started.

TABLE 1 Number of participants who completed 48 trials, 60 trials, or a different number of trials (due to premature termination of the experiment), per eye tracker

	48 trials	60 trials	Different number of trials (median number of trials)
Spectrum	23	43	16 *(36)*
TX300	334	596	177 *(40)*

Note: For participants with a different number of trials, the median number of trials is given in parentheses.

TABLE 2 Percentage of disregarded trials for the accuracy analysis, per eye tracker, for each age group

	5-month-old	10-month-old	3-year-old
Spectrum	7%	0.5%	5%
TX300	26%	25%	14%

Note: Trials were disregarded when no fixations were available during the presentation of the central stimulus.

2.3 | Operationalizations of eye-tracking data quality

2.3.1 | Precision

Precision was computed with a moving-window method applied to the gaze position signals. The RMS deviation was calculated for each window (100 ms, 30 samples for the TX300, 60 samples for the Spectrum).[5] The step size of the moving window was 20 ms (6 samples for the TX300, 12 samples for the Spectrum). As the moving window was slid across the gaze position signal, the RMS-s2s deviation was also calculated for windows including saccades. This may seem problematic because during saccades, the RMS-s2s deviation may be higher. As the distributions were positively skewed, we computed the median because the median is less sensitive to skewness. These medians were averaged over all the trials for every participant. The horizontal and vertical RMS-s2s deviations were summed up through Pythagoras' theorem to acquire the precision for the left and for the right eye. These two values were subsequently averaged to acquire one final estimate of precision of the gaze position signal for each participant.

2.3.2 | Accuracy

Accuracy was estimated through the distance between the central target and the closest fixation which occurred during the presentation of the central target. For each participant, these accuracy values were averaged over all trials, resulting in one accuracy estimate for each participant. When no valid fixations were available, the trial was disregarded. Table 2 provides an overview of the percentage of disregarded trials in the accuracy analysis, per eye tracker, for each age group. Fixation(s) in the gaze position signal were classified with the I2MC algorithm, version 2.0 (Hessels et al., 2017). We chose I2MC because it was designed to classify fixations in gaze

[5] Using a moving window of 30 samples (100 ms for the TX300, 50ms for the Spectrum) resulted in similar findings.

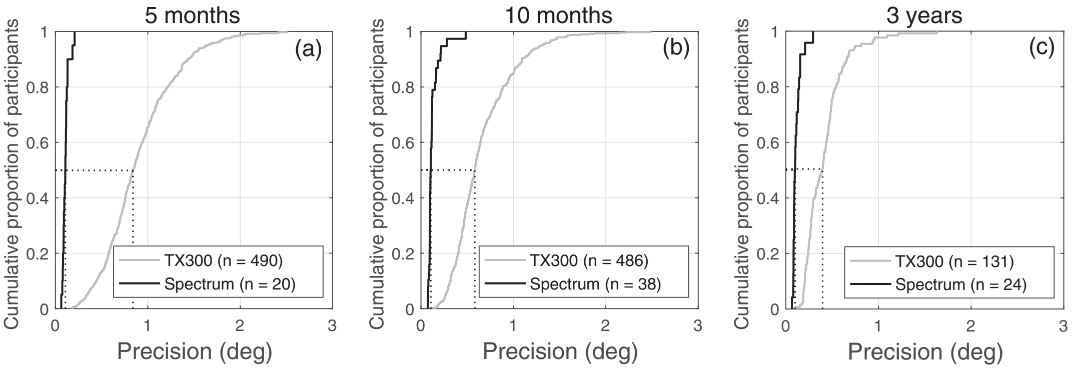

FIGURE 2 Precision for three age groups and two eye trackers. Panels depict the cumulative probability functions of RMS-s2s deviation as a measure for precision for the TX300 and Spectrum, separately for the 5-month-olds (panel a), 10-month-olds (panel b), and 3-year-olds (panel c). The number of steps in each line correspond to the number of participants. The dotted lines indicate the median RMS-s2s for each eye tracker. The gray lines depict the functions for the TX300, the black lines for the Spectrum. For each age group, precision of the gaze position signal is overall better when obtained with the Spectrum than with the TX300

position signals prone to a wide range of noise levels and when periods of data loss may occur, which is likely in eye-tracking data from infants (Dalrymple et al., 2018; Hessels & Hooge, 2019; Wass et al., 2014).

2.3.3 | Data loss

For data loss, the proportion of samples where no gaze position was reported by the eye tracker was determined separately for the left eye and the right eye signal, for each trial. These proportions were subsequently averaged for each participant, and finally averaged over eyes, resulting in one proportion of data loss for each participant.

2.4 | Eye-tracking data quality compared between two eye trackers

We compared the eye-tracking data quality between the Spectrum and the TX300 through empirical cumulative distribution functions (ECDFs). We chose ECDFs as they allow us to compare data quality between different group sizes. Moreover, they allow the reader to easily determine what proportion of participants yielded precision, accuracy, or data loss under a specific value.

3 | RESULTS

3.1 | Precision

How does the distribution of precision compare between the eye trackers? As is visible in Figure 2, the black line (Spectrum) is to the left of the gray line (TX300) for all three age groups. This means that the median precision value was lower and, therefore, better when recorded with the Spectrum compared to the TX300 (see dotted lines Figure 2). Moreover, as is visible from

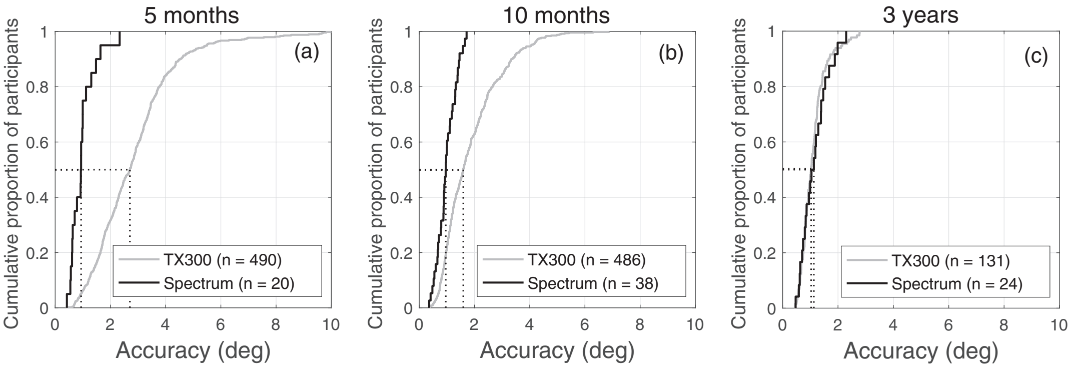

FIGURE 3 Accuracy for three age groups and two eye trackers. Panels depict the cumulative probability functions of accuracy for the TX300 and Spectrum, separately for the 5-month-olds (panel a), 10-month-olds (panel b), and 3-year-olds (panel c). The number of steps in each line correspond to the number of participants. The dotted lines indicate the median accuracy for each eye tracker. The gray lines depict the functions for the TX300, the black lines for the Spectrum. For the 5- and 10-month-old group, accuracy of the gaze position signal is overall better when obtained with the Spectrum. For the 3-year-old group, accuracy of the gaze position signal is similar between the two eye trackers

the fact that the black lines (Spectrum) are much steeper than the gray lines (TX300), variance in precision values is smaller for the Spectrum than for the TX300. For example, for the 5-month-old group measured with the Spectrum, 90% of the precision values range between 0.06° and 0.13°. In contrast, the range for the TX300 group is much larger, ranging from 0.19° to 1.43°. Furthermore, we observe that the gaze position signals for the TX300 were more precise for the older groups than for the younger groups. This is visible from the medians (dotted lines), where for the 5-month-old infants (panel a) the median RMS-s2s deviation is around 0.8°, for the 10-month-old infants (panel b) approximately 0.6° and for the 3-year-olds (panel c) around 0.4°. For the Spectrum median RMS-s2s deviation is approximately 0.1° for each age group.

3.2 | Accuracy

As is clearly visible in Figure 3, the black line (Spectrum) is to the left of the gray line (TX300) for the 5-month-old group (panel a) and the 10-month-old group (panel b). This indicates that measurements done with the Spectrum show higher accuracy than measurements with the TX300. For the 3-year-old group (panel c in Figure 3) the black and gray lines overlap, indicating similar accuracy of the gaze position signals for both eye trackers. Moreover, as is visible through the steepness of the lines, variance in accuracy levels is smaller for the 5- and 10-month-olds measured with the Spectrum, than for the TX300. For example, for the 5-month-olds (panel a in Figure 3) 90% of the accuracy values ranged from 0.42° to 1.64° for the group measured with the Spectrum, while for the group measured with the TX300 these values ranged from 0.6° to 4.48°. Furthermore, accuracy for the TX300 recordings was best for the 3-year-olds, followed by the 10-month-old and 5-month-old infants, respectively. This is visible from the median (indicated by the dotted lines), where for the 5-month-old infants (panel a), the median is approximately 2.7°, for the 10-month-olds (panel b) approximately 1.6°, and for the 3-year-olds (panel c)

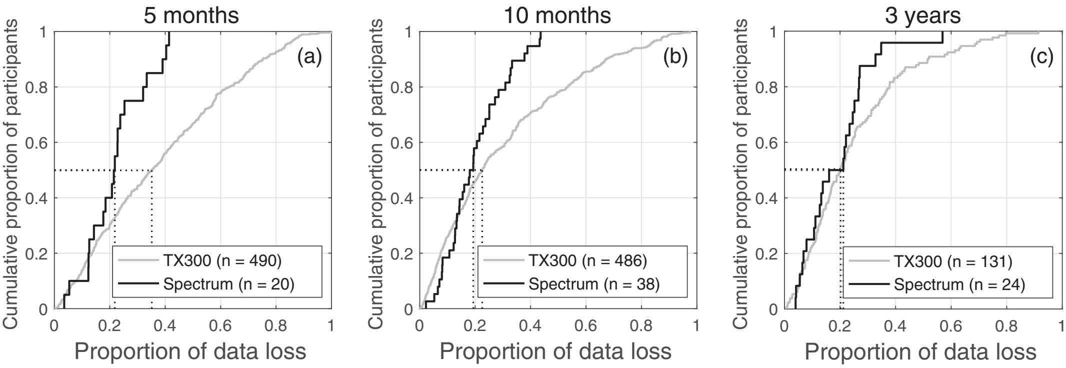

FIGURE 4 Proportion of data loss for three age groups and two eye trackers. Panels depict the cumulative probability functions of proportion of data loss for the TX300 and Spectrum, separately for the 5-month-olds (panel a), 10-month-old (panel b), and 3-year-olds (panel c). The number of steps in each line correspond to the number of participants. The dotted lines indicate the median proportion of data loss for each eye tracker. The gray lines depict the functions for the TX300, the black lines for the Spectrum. For the 5- and 10-month-olds, the Spectrum typically exhibited lower proportions of data loss compared to the TX300. For the 3-year-olds, the proportion of data loss is more similar across eye trackers

approximately 1°. For the Spectrum, median accuracy is approximately 1° for all age groups. We found similar results for the left and right peripheral target.

3.3 | Data loss

Figure 4 depicts data loss for the TX300 and the Spectrum for three age groups. For the 5- and 10-month-old groups, the gaze position signals from the Spectrum show lower proportions of data loss. The proportion of data loss is similar for both eye trackers in the 3-year-old group. Moreover, the variance of the proportion of data loss was smaller for the Spectrum than for the TX300. For example, for the 5-month-olds (panel a) 90% of the proportion of data loss ranged from 0.04 to 0.4 for the group measured with the Spectrum, while for the group measured with the TX300 these values ranged from 0.01 to 0.75. For the 5-month-olds, the median proportion of data loss (indicated by the dotted line in Figure 4) was approximately 0.21 for the infants measured with the Spectrum, and 0.35 for the infants measured with the TX300. For the 10-month-olds, the median proportion of data loss was 0.18 when obtained with the Spectrum, and 0.23 for the TX300. For the 3-year-olds, the median proportion of data loss was more similar: around 0.17 for the Spectrum and 0.2 for the TX300.

3.4 | Can differences in eye-tracking data quality be attributed to selection bias?

The results thus far indicate that eye-tracking data quality is generally better for the Spectrum than for the TX300. However, it could be the case that the differences in eye-tracking data quality did not originate from differences between the eye trackers. Can we rule out the possibility that the better data quality of the Spectrum is caused by an accidental selection bias, given the relatively small group recorded with the Spectrum ($N = 82$) compared to the TX300 ($N = 1107$). It

TABLE 3 Percentages of children with non-blue eyes per eye tracker, for each age group

	5-month-old	10-month-old	3-year-old
Spectrum	15.0%	32.4%	43.5%
TX300	19.6%	21.7%	28.7%

TABLE 4 Child's movement as indicated by the operators per eye tracker, and for each age group

Age group; eye tracker	Never	Seldom	Often	Always
5-month-olds Spectrum	15.0%	25.0%	45.0%	15.0%
5-month-olds TX300	20.8%	45.3%	21.8%	12.1%
10-month-olds Spectrum	13.5%	18.9%	51.4%	16.2%
10-month-olds TX300	17.3%	45.3%	25.1%	12.3%
3-year-olds Spectrum	21.4%	43.5%	26.1%	8.7%
3-year-olds TX300	19.4%	49.6%	27.1%	3.9%

could be that the group of children recorded with the Spectrum happened to produce better eye-tracking data quality. For example, eye color and amount of movement of the child are relevant for eye-tracking data quality (see Hessels, Andersson, et al., 2015; Hessels et al., 2015; Niehorster et al., 2018).

In our sample we scored eye color as "blueish" (blue, green, and/or gray) or "non-blue" to match previous research (Nyström et al., 2013). The percentages of children per age and eye tracker with non-blue eyes can be found in Table 3. A lower percentage of 5-month-olds with non-blue eyes was recorded with the Spectrum than with the TX300. For the 10-month- and 3-year-olds, a higher percentage of children with non-blue eyes was recorded with the Spectrum than with the TX300. Since it has been reported that data quality is usually better for people with non-blue eyes (Hessels, Andersson, et al., 2015; Nyström et al., 2013), the better quality of the Spectrum eye-tracking data may be partly due to group differences in eye color. However, we found the largest differences between the eye trackers for all measures of data quality for the 5-month-olds, and in this group the percentage of children that had non-blue eyes was lower for the Spectrum than the TX300.

Furthermore, the previous studies reporting better data quality for people with non-blue eyes have exclusively used eye trackers that employ the dark-pupil technique, such as the TX300. The Spectrum can, however, use both dark and bright pupil tracking modes, and it is unknown whether systematic differences in data quality between eye colors may also be expected when using the bright pupil technique. We address the consequences of the Spectrum being able to track in both dark and bright pupil modes in the discussion.

The amount of movement of the children was indicated by the operator on a 4-point-scale: *never*, *seldom*, *often,* and *always*. In Table 4, the amount of movement as indicated by the operator is presented per eye tracker, for each age group. The 5- and 10-month-olds measured with the Spectrum were reported to move more often in comparison with the TX300. As movement may have a negative effect on data quality, one may expect worse data quality as the amount of movement increases. However, we found data quality to be better for the Spectrum. We, therefore, conclude that the differences in data quality cannot be attributed to the group differences in movement. Reported movement for the 3-year-olds was similar for both eye trackers.

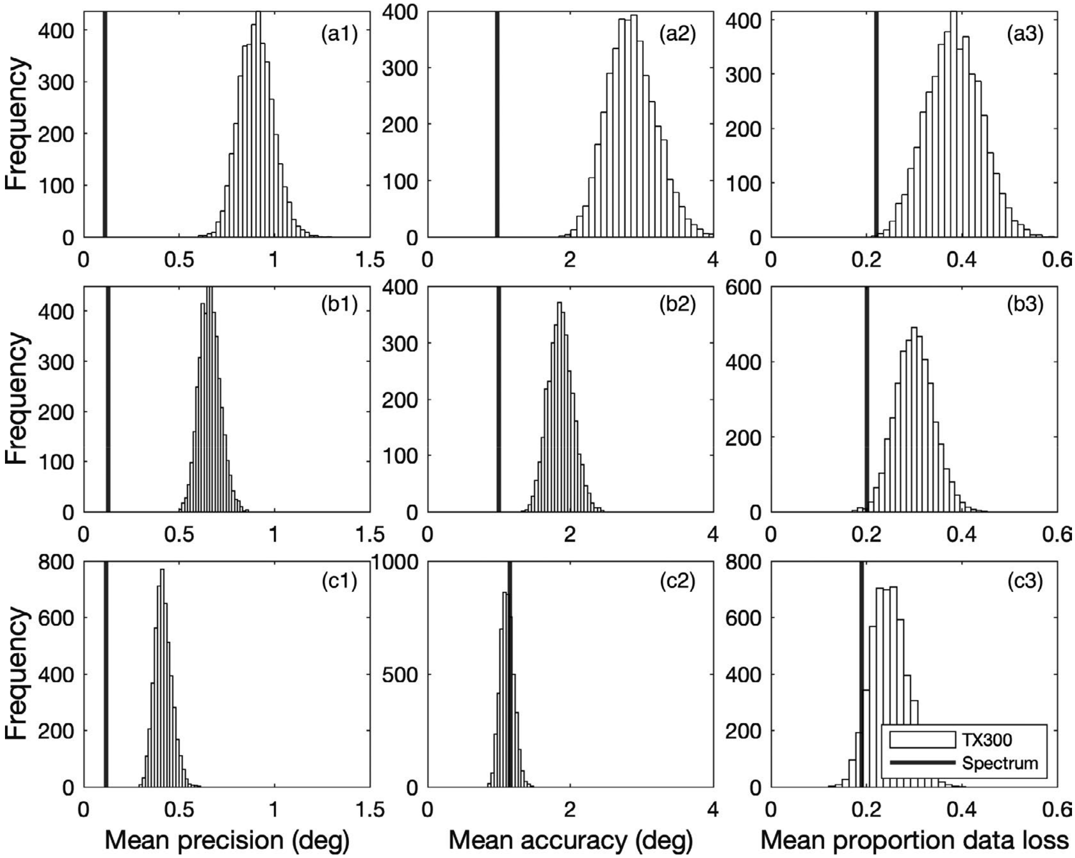

FIGURE 5 Distributions of means for the estimates of the precision, accuracy and data loss. Panels labeled a depict data from the 5-month-old infants, panels b data from the 10-month-old infants, and panels c data from the 3-year-old children. Panels 1 exhibit the distribution of means for precision (RMS-s2s deviation) of the samples from the TX300. Panels 2 depict the distribution of means of accuracy, estimated through the mean distance between assumed true and recorded gaze position, per age group. Panels 3 depict data loss, estimated through the proportion of samples without a gaze coordinate, for each age group. The black vertical lines represent the means of the data quality measures of all measurements with the Spectrum

To investigate whether the results in our study could be explained through accidental selection bias, we now assume that there is no difference in eye-tracking data quality between the two eye trackers, but that the differences in eye-tracking data quality are due to differences between the groups.

We drew a large number (5000) of random subsets of participants from the TX300 data and computed average eye-tracking data quality. Each sample was as large as the group measured with the Spectrum. We thereby produced a distribution of means for each eye-tracking data quality measure per age. This allowed us to state the eye-tracking data quality obtained with the Spectrum as a percentage of that obtained with the TX300. If the percentages are higher than or equal to 97.5% (corresponding to two-tailed significance level of .05) we conclude that the differences in data quality are *not* due to selection bias but to differences between the eye trackers.

The mean precision values (panels a1, b1 and c1 in Figure 5) of the gaze position signals obtained with the Spectrum were always lower (indicating better precision) than the mean precision of the TX300 subsets. The mean accuracy values (panels a2, b2 and c2 Figure 5) from the

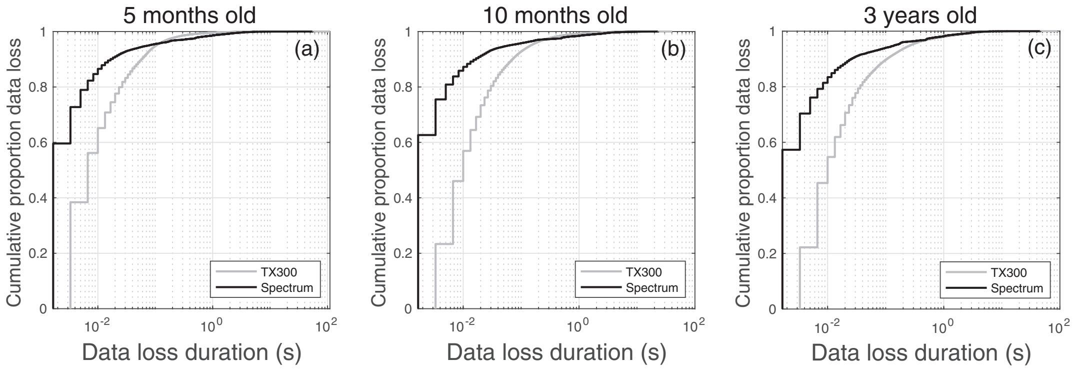

FIGURE 6 Empirical cumulative distribution functions of the durations of data loss for the 5-month-old (panel a), 10-month-old (panel b) and 3-year-old children (panel c). The gray lines depict the functions for the TX300, the black lines for the Spectrum. Note that the *x* axis is logarithmic

5- and 10-month-old infants measured with the Spectrum were always lower (therefore better) than the mean accuracy values of the TX300 subsets. For the 3-year-old group, the mean accuracy of the Spectrum was similar to or better than 32.9% of the TX300 subset. The mean proportion of data loss (panels a3, b3 and c3 in Figure 5) from the Spectrum was similar to or better than 99.4% of the TX300 subsets for the5-month-old group, 99.72% for the 10-month-old group and 94.4% for the 3-year-old group. Thus, we conclude that our findings cannot be explained through selection bias.

3.5 | Duration of data loss

Thus far we have operationalized data loss as the proportion of lost samples. However, data loss may have at least two causes. Data loss may occur when the participant is blinking or not looking at the screen, or when the participant is looking at the screen but the eye tracker fails to produce a gaze coordinate. The latter may occur when downward facing eye lashes obscure the pupil in the eye image, or when dark makeup hinders the determination of the pupil center (Holmqvist & Andersson, 2017, p. 207). Blinking or looking away usually cause data loss periods to be longer than 100 ms. For such periods we cannot determine whether the cause of the data loss is by human behavior solely, or in the determination of the gaze position by the eye tracker. However, shorter data loss periods, that is, those under a 100 ms, are unlikely due to blinking or looking away and are thus indicative of the quality of the measurement. Fortunately, shorter periods (e.g., under 100 ms) are potential candidates for interpolation. We thus ask how the distributions of data loss periods compare between the eye trackers.

Figure 6 depicts the ECDFs of the data loss durations for the Spectrum and the TX300 in three age groups. As is visible in panel a in Figure 6, roughly 85% of all periods of data loss observed in the Spectrum from the 5-month-old infants had durations of 0.01 s or less, while for the TX300 roughly 55% of all periods of data loss observed had durations of 0.01 s or less. For the 10-month-old infants (panel b in Figure 6) roughly 85% of all periods of data loss observed in the Spectrum gaze position signals had durations of 0.01 s or less, while for the TX300 roughly 45% of all periods of data loss observed had durations of 0.01 s or less. For the 3-year-olds (panel c in Figure 6), roughly 81% of all periods of data loss observed in the Spectrum gaze position signals had durations of 0.01 s or less, while for the TX300 roughly 45% of all periods of data loss observed had durations of 0.01 s or less.

When combining these findings with the observed differences in the proportion of data loss between the Spectrum and the TX300, we may draw the following conclusions. Not only is the overall data loss lower for the Spectrum than for the TX300, but the proportion of shorter data loss periods (those that may be candidates for interpolation) is much higher. Thus, the Spectrum yields more analyzable eye-tracking data.

3.6 | Saccadic latency

So far, we have observed differences in eye-tracking data quality between the two eye trackers. As we argued in the introduction, data quality may affect outcome measures of an eye-tracking experiment. To observe whether this was the case in the present experiment, we compared the saccadic latency—an important outcome measure of the gap-overlap experiment—for the gap and overlap conditions between the two eye trackers per age group. For participants who executed at least 48 trials, the saccadic latency was determined for the first 48 trials. Participants who completed less than 48 trials were excluded from the present analysis. For the Spectrum, 17 five-month-olds, 36 ten-month-olds, and 16 three-year-olds completed at least 48 trials. With the TX300, 453 five-month-olds, 440 ten-month-olds, and 135 three-year-olds completed at least 48 trials. Saccadic latency was operationalized as the difference between peripheral stimulus onset time and fixation offset time from the central stimulus. Fixations in the gaze position signals were classified with the noise-robust I2MC algorithm (Hessels et al., 2017). Trials were excluded when:

1. There was more than 75% data loss in the trial.
2. When the last fixation on the center was followed by data loss.
3. A saccade was made to the wrong side of the screen.
4. When the saccadic latency was shorter than 80 ms, as this indicates that saccade programming was probably initiated before the onset of the peripheral target (Fischer & Ramsperger, 1984).

The mean number of trials excluded for each criterium is reported in Table 5. As visible in the last row of this table, the number of remaining trials for which the saccadic latency could be computed, was lower for the TX300 than for the Spectrum. This was mainly due to exclusion based on data loss criteria (row 1–4 in Table 5), rather than criteria related to participant behavior (rows 5–8 in Table 5).

The median saccadic latencies and interquartile ranges for the gap and overlap conditions are presented in Table 6, per eye tracker, for each age group. As is visible from this table, the interquartile ranges show substantial overlap, and there does not seem to be a clear systematic difference in median saccadic latencies between the two eye trackers. This may seem counterintuitive, given that we argued in the introduction that data quality may affect experimental outcome measures. However, it is crucial to note that we considered the variation in precision in our eye-tracking data when choosing the fixation-classification algorithm used in deriving the outcome measures (saccadic latencies in this case). We applied a noise-robust fixation classification algorithm (I2MC), which was found to be robust to variability in precision between 0° and 2° RMS-s2s deviation (Hessels et al., 2017). In our study, the RMS-s2s deviation of the

TABLE 5 Exclusion criteria and mean number of trials excluded across participants

	5-month-old	10-month-old	3-year-old
More than 75% data loss - Spectrum	1.5	2.2	1.0
More than 75% data loss - TX300	6.8	5.1	3.6
Last central fixation followed by data loss – Spectrum	0.8	0.75	0.5
Last central fixation followed by data loss - TX300	9.2	6.3	3.0
Saccade to wrong side Spectrum	5.9	3.1	2.7
Saccade to wrong side TX300	4.9	2.9	2.5
Saccadic latency <80 ms Spectrum	0.8	1.2	1.6
Saccadic latency <80 ms TX300	0.9	0.9	1.2
Remaining trials Spectrum	29.6	33.9	37.4
Remaining trials TX300	23.2	29.7	34.4

TABLE 6 Median saccadic latencies in milliseconds (and interquartile ranges) for the gap and overlap condition for each age group

	Gap condition	Overlap condition
5-month-olds Spectrum	282.2 (234.8–402.6)	1159.1 (506.8–2152.5)
5-month-olds TX300	320.2 (286.4–378.3)	918.7 (543.9–2079.0)
10-month-olds Spectrum	222.7 (203.8–243.2)	635.4 (508.8–1306.9)
10-month-olds TX300	254.5 (233.9–277.0)	586.4 (465.8–1258.9)
3-year-olds Spectrum	213.2 (184.9–241.2)	434.2 (352.8–801.3)
3-year-olds TX300	234.0 (210.0–263.7)	454.6 (379.7–691.2)

TX300 rarely exceeded 2° for the 5- and 10-month-old infants, and never for the 3-year-old children. We also excluded trials with more than 75% data loss and trials in which the last fixation on the central target was followed by data loss. More trials were excluded based on these criteria for the TX300 than the Spectrum (see Table 5). We are, therefore, hesitant to conclude that data quality does not affect our outcome measures (e.g., based on a statistical analysis of the saccadic latencies), given the choice for analysis tools and exclusion criteria in our study. Whereas the problem of different levels of precision may be minimized by choosing a suitable fixation classification algorithm, there is no solution to the problem of data loss. We address this further in the discussion.

4 | DISCUSSION

The duration of longitudinal studies may exceed the lifespan of an eye tracker. When an eye tracker breaks down, it has to be replaced. However, the data quality of the replacement eye tracker may differ from the data quality of the replaced eye tracker. As data quality may affect experimental outcome measures, this may have consequences for the interpretation of the eye-tracking data. This problem may also occur when replacing an eye tracker in between studies and setting up a multi-site study with different eye trackers. In the present study, we compared the data quality of gaze position signals obtained with the Tobii Pro TX300 and Tobii Pro Spectrum. The TX300 is out of production and the Spectrum is promoted by Tobii Pro as its successor. Data quality of the TX300 and the Spectrum were operationalized through precision, accuracy and data loss. Data quality was compared between the TX300 and the Spectrum for several age groups: 5-month-old infants, 10-month-old-infants and 3-year-old children.

Overall, the Spectrum produced gaze position signals with better data quality than the TX300. The precision of the Spectrum was better for all age groups, especially for the 5- and 10-months-old infants. The accuracy of the Spectrum was better for the 5- and 10-month-olds, and similar to the TX300 for the 3-year-olds. The proportion of data loss was lower for the Spectrum, especially for the 5-months-old. Moreover, the duration of data loss periods was shorter when obtained with the Spectrum, for all age groups.

4.1 | Implications of replacement

Upon replacement of the TX300 with the Spectrum in a longitudinal cohort study, data quality will improve. Intuitively, collecting eye-tracking data with better data quality may seem unproblematic. However, as mentioned in the introduction, outcome measures of eye-tracking experiments may be affected by the increase in data quality. This is especially the case for research with younger infants, where the difference in data quality between the two eye trackers is the largest. If one were to compare experimental outcome measures from groups recorded with different eye trackers, one may conclude there to be a difference in looking behavior which is actually attributable to data quality differences. Several things should, therefore, be considered when replacing an eye tracker with a better one.

4.1.1 | Precision

When replacing the TX300 with the Spectrum, precision will increase, especially for younger infants. How this increase in precision would affect the experimental outcome measures, depends on the specific analysis tool used. The most common investigated outcome measures of eye-tracking experiments are the number and duration of fixations (p. 527 and 561 Holmqvist & Andersson, 2017). Fixations in the gaze position signal are often automatically classified by algorithms. Depending on the fixation classification algorithm used, lower precision could lead to either an increased number of fixations and decreased average fixation duration or to a decreased number of fixations and increased average fixation duration (see Hessels et al., 2017; Wass et al., 2013).

4.1.2 | Accuracy

When using the Spectrum compared to the TX300, accuracy will be better, especially for younger children. If the accuracy is of the same magnitude as the size of or distance between AOIs, AOI based measures are unreliable. Thus, one should take into account the expected accuracy values of the target group when designing an experiment. If this is done properly, we foresee no obvious problem with the increased accuracy of the Spectrum.

4.1.3 | Data loss

Upon replacement, data loss is likely to improve in two ways: The proportion of data loss is likely to become smaller and the duration of the data loss periods is likely to be shorter. This affects the computation of experimental outcome measures. Experimental outcome measures may vary with different levels of data loss. During fixation classification, fixation candidates may be broken up by short periods of data loss (Holmqvist et al., 2012), in which case the number of fixations increases and the average fixation duration decreases. Whether this happens or whether the number of fixations decreases and the average fixation duration increases, depends on the fixation classification algorithm used (Hessels et al., 2017).

Another implication follows from the fact that the proportion of data loss is often used for including trials and participants. When the proportion of data loss is lower, more trials per participant are likely to be included, as well as the number of participants in the experiment. Also, as data loss durations tend to be shorter, more of these periods are likely to be candidates for interpolation. Together this could mean that more subjects and trials are included when the TX300 is replaced with the Spectrum.

4.2 | Suggestions for solutions

In longitudinal studies, researchers should be aware of the implications of replacing an eye tracker with a better one. We advise to always check how data quality measures relate to the outcome measures of an experiment. We provide some suggestions for solutions here to the implications mentioned above.

4.2.1 | Precision

For precision, one might wish to use a fixation-classification algorithm robust to a wide range of noise levels in the gaze position signals. Hessels et al. (2017) and van Renswoude et al. (2018) developed such algorithms: Identification by two-means clustering (I2MC) and GazePath. For the I2MC algorithm, fixation duration and number of fixations were found to be robust to a range in precision between 0° and 2° RMS-s2s deviation. In our study, the RMS-s2s deviation of the TX300 rarely exceeded 2° for the 5- and 10-month-old infants, and never for the 3-year-old children. This is in line with Hessels, Andersson, et al. (2015) who reported that the RMS-s2s deviation rarely exceeds 3° in infant research. For GazePath, the authors show that the median fixation duration computed with their method is not correlated to precision in their participant group. The median fixation duration is correlated to precision for the EyeLink and Tobii analysis methods. For a detailed description of the algorithms and their application we refer to the respective papers.

4.2.2 | Data loss

For the effect of the amount data loss on experimental outcome measures, we again advice to choose a fixation classification algorithm robust to different levels of data loss. A suggestion is the previously mentioned I2MC algorithm (Hessels et al., 2017). Compared to seven commonly used fixation classification algorithms, I2MC was shown to be most robust to different levels of data loss under 100 ms.

4.3 | To replace or not to replace

For our participant groups, we found that the Spectrum generally produces gaze position signals with better data quality. The question may, therefore, arise whether the TX300 should be replaced anyway, without it breaking down. There is no easy answer to this question, but we would like to discuss some considerations. Variability in data quality does not only pose a problem when switching eye trackers. For example, as Hessels and Hooge (2019) showed, data quality also varies with age of the participant even when measured with the same eye tracker. Moreover there is a large variability in data quality between subjects. Therefore, most arguments made in this paper about differences in data quality and their effect on outcome measures could already be made for the infant gaze position signals in the first place, even without changing the eye tracker. If we look at the Spectrum, we see that the data quality measures are similar for every age group. For example, precision is approximately 0.1° for each age group (black lines in Figure 2). Also, variance in all data quality measures is much smaller. What this means is that the Spectrum decreases the problem of differences in data quality between subjects and across age groups. Another benefit is the lower proportion of data loss which may allow inclusion of more trials per subject and more subjects.

4.4 | Other notable differences between the TX300 and the spectrum

Besides eye-tracking data quality, sampling frequency might also affect the eye-tracking measures derived from an experimental set up. The Spectrum and TX300 collect eye movement data at different sampling frequencies. The TX300 collects gaze positions at 300 Hz, while the Spectrum can collect gaze positions at up to 1200 Hz. Andersson et al. (2010) used simulations to study the effect of sampling frequency on temporal eye-tracking measures. They distinguish between latency measures (defined by one given and one sampled time point) and duration measures (defined by two sampled time points). Consider, for example, a latency measure such as saccadic reaction time. The given point could be the onset time of a certain target, while the sampled time point would be the start of the fixation on the target. The "true" start of the fixation on the target could be just before the gaze position is sampled (and therefore picked up immediately), or just after sampling (the fixation will be registered at the next sample). The temporal sampling error may, therefore, be near zero (in the former case) or as large as one sampling interval (3.33 ms for the TX300; 1.67 ms for the Spectrum) in the latter case. One should be aware of this when comparing, for example, standard deviations of saccadic reaction times between groups. Should one want to exclude the difference in temporal sampling frequency as a potential explanation in an analysis, one could downsample the data of the higher frequency to the lower frequency.

Another notable difference between the two eye trackers is the illumination method. The TX300 uses dark-pupil eye tracking, where an illuminator (the infrared light source) is located away from the optical axis of the camera. When filmed, the pupil appears black, and these dark pupils in combination with the corneal reflection of the infrared light are used to estimate gaze position. The Spectrum is also able to use the bright pupil eye-tracking technique. In bright pupil tracking, the infrared light is located on the optical axis. In the eye image, this results in bright pupils which, in combination with the corneal reflection, are used to estimate gaze position. During calibration, all participants are subjected to dark and bright pupil methods and the method that is found to provide the best signal is chosen by the eye tracker for the actual recording. In this study, the gaze position signals of 4 out of the 82 children measured with the Spectrum were recorded through dark-pupil eye tracking,[6] the other 78 children were recorded through bright pupil eye tracking. The illumination method may affect the contrast between the pupil and the rest of the eye. This could potentially be a factor in differences in accuracy and precision between the two eye trackers in this study.

Finally, the following observation was made when operating both systems. The design of the TX300 and the Spectrum differ in several ways, one of which is the plastic sheet which covers the camera and infrared lights. This cover is more reflective for the Spectrum. Several research assistants in our cohort study reported that for some older baby's and 3-year-olds, the reflective cover functioned as a mirror in which the children made funny faces at themselves, or at least seemed to stare at it more frequently in comparison to the TX300.

4.5 | Limitations

In this study, we compared the data quality of the TX300 and Spectrum using a between subject-design. Ideally, we would have done this with a fully within-subjects design, that is, having children perform the gap-overlap experiment with both eye trackers. However, this could not be accommodated as part of the already running cohort study and within the constraints of the protocol approved by the medical ethics committee. The children are subjected to a battery of experiments on the day that they are in the research center. However, there is no time left for additional measurements. Conducting a within-subjects design would have meant excluding one of the other measurements (e.g., EEG or observations of parent-child interaction), which we deemed unwanted.

It should be noted that our estimates for accuracy were based on one spatial location. Ideally, one would estimate accuracy at the various relevant spatial locations used in the experiment. If one conducts, for example, free-viewing experiments with infants and young children, it makes sense to estimate accuracy for multiple locations, not only the center of the screen.

5 | CONCLUSION

The Tobii Pro Spectrum produces eye-tracking data with better data quality than the Tobii Pro TX300. If, for any reason, an eye tracker has to be replaced in an ongoing study, a newer version might seem unproblematic (as it often has better specifications). However, we want to stress that this may have implications for the outcome measures of the study. We advise any ongoing study, especially longitudinal studies where chances are that the duration of the study exceeds the lifespan of an eye tracker, to evaluate how eye-tracking data quality relates to their experimental outcome measures (e.g., when comparing groups or recordings from different time points). Finally, we advise applying noise-robust fixation classification algorithms

[6] This number was found upon manual inspection of the eye images. Eye images were saved for all the measurements with Spectrum. The identities of the children were not linked to the images.

when appropriate, to diminish the effect of differences in eye-tracking data quality on experimental outcome measures.

ACKNOWLEDGMENTS

Y.J.R. de Kloe and R.S. Hessels were supported by the Consortium on Individual Development (CID). CID is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003, to C. Kemner). The authors declare no conflicts of interest with regard to the funding source for this study.

CONFLICT OF INTEREST

The Tobii Pro Spectrum was provided by Tobii Pro for a loan period of several months. Tobii Pro had no involvement in this study.

ORCID

Yentl J.R. De Kloe https://orcid.org/0000-0001-5950-4973

REFERENCES

Andersson, R., Nyström, M., & Holmqvist, K. (2010). Sampling frequency and eye-tracking measures: How speed affects durations, latencies, and more. *Journal of Eye Movement Research*, *3*(3), 1–12.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Carpenter, M., Nagell, K., Tomasello, M., Butterworth, G., & Moore, C. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, *63*(4), i. https://doi.org/10.2307/1166214

Cashon, C. H., & Cohen, L. B. (2004). Beyond U-shaped development in infants' processing of faces: An information-processing account. *Journal of Cognition and Development*, *5*(1), 59–80.

Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of Learning and Memory*, *92*(2), 225–234. https://doi.org/10.1016/j.nlm.2008.06.002

Cousijn, J., Hessels, R. S., Van der Stigchel, S., & Kemner, C. (2017). Evaluation of the psychometric properties of the gap-overlap task in 10-month-old infants. *Infancy*, *22*(4), 571–579. https://doi.org/10.1111/infa.12185

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*, 803. https://doi.org/10.3389/fpsyg.2018.00803

Falck-Ytter, T., Carlström, C., & Johansson, M. (2015). Eye contact modulates cognitive processing differently in children with autism. *Child Development*, *86*(1), 37–47. https://doi.org/10.1111/cdev.12273

Fantz, R. L. (1956). A method for studying early visual development. *Perceptual and Motor Skills*, *6*(1), 13–15. https://doi.org/10.2466/pms.1956.6.g.13

Fantz, R. L. (1961). A method for studying depth perception in infants under six months of age. *The Psychological Record*, *11*, 27–32.

Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668–670. https://doi.org/10.1126/science.146.3644.668

Fischer, B., & Ramsperger, E. (1984). Human express saccades: Extremely short reaction times of goal directed eye movements. *Experimental Brain Research*, *57*(1), 191–195. ISSN, 0014-4819, 1432-1106. https://doi.org/10.1007/BF00231145

Gomez, A., Piazza, M., Jobert, A., Dehaene-Lambertz, G., & Huron, C. (2017). Numerical abilities of school-age children with Developmental Coordination Disorder (DCD): A behavioral and eye-tracking study. *Human Movement Science*, *55*, 315–326. https://doi.org/10.1016/j.humov.2016.08.008

Haith, M., Bergman, T., & Moore, M. (1977). Eye contact and face scanning in early infancy. *Science*, *198*(4319), 853–855. https://doi.org/10.1126/science.918670

Hessels, R. S., Andersson, R., Hooge, I. T. C., Nyström, M., & Kemner, C. (2015). Consequences of eye color, positioning, and head movement for eye-tracking data quality in infant research. *Infancy*, *20*(6), 601–633. https://doi.org/10.1111/infa.12093

Hessels, R. S., Cornelissen, T. H. W., Kemner, C., & Hooge, I. T. C. (2015). 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

Hessels, R. S., & Hooge, I. T. (2019). Eye tracking in developmental cognitive neuroscience – The good, the bad and the ugly. *Developmental Cognitive Neuroscience*, *40*, 100710. https://doi.org/10.1016/j.dcn.2019.100710

Hessels, R. S., Hooge, I. T. C., & Kemner, C. (2016). An in-depth look at saccadic search in infancy. *Journal of Vision*, *16*(8), 1534–7362. https://doi.org/10.1167/16.8.10

Hessels, R. S., Niehorster, D. C., Kemner, C., & Hooge, I. T. C. (2017). Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). *Behavior Research Methods*, *49*(5), 1802–1823. https://doi.org/10.3758/s13428-016-0822-1

Holmqvist, K., & Andersson, R. (2017). *Eye tracking: A comprehensive guide to methods, paradigms, and measures* (2nd ed.). CreateSpace.

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 – ETRA '12* (p. 45). ACM Press. https://doi.org/10.1145/2168556.2168563

JASP Team. (2020). JASP (version 0.14.1) [Computer software].

Johnson, S. P., Slemmer, J. A., & Amso, D. (2004). Where infants look determines how they see: Eye movements and object perception performance in 3-month-olds. *Infancy*, *6*(2), 185–201.

Jones, E., Mason, L., Begum Ali, J., van den Boomen, C., Braukmann, R., Cauvet, E., Demurie, E., Hessels, R., Ward, E., Hunnius, S., Bolte, S., Tomalski, P., Kemner, C., Warreyn, P., Roeyers, H., Buitelaar, J., Falck-Ytter, T., Charman, T., & Johnson, M. (2019). Eurosibs: Towards robust measurement of infant neurocognitive predictors of autism across Europe. *Infant Behavior and Development*, *57*, 101316. https://doi.org/10.1016/j.infbeh.2019.03.007

Jones, W., & Klin, A. (2013). Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism. *Nature*, *504*(7480), 427–431. https://doi.org/10.1038/nature12715

Niehorster, D. C., Cornelissen, T. H. W., Holmqvist, K., Hooge, I. T. C., & Hessels, R. S. (2018). What to expect from your remote eye-tracker when participants are unrestrained. *Behavior Research Methods*, *50*(1), 213–227. https://doi.org/10.3758/s13428-017-0863-0

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

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*(1), 188–204. https://doi.org/10.3758/BRM.42.1.188

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E., Brouwer, R. M., Buimer, E. E., Hessels, R. S., de Heus, R., Huijding, J., Junge, C. M., Mandl, R. C., Pas, P., Vink, M., van der Wal, J. J., Hulshoff Pol, H. E., & Kemner, C. (2020). The YOUth study: Rationale, design, and study procedures. *Developmental Cognitive Neuroscience*, *46*, 100868. https://doi.org/10.1016/j.dcn.2020.100868

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

Rose, S. A., Gottfried, A. W., Melloy-Carminar, P., & Bridger, W. H. (1982). Familiarity and novelty preferences in infant recognition memory: Implications for information processing. *Developmental Psychology*, *18*(5), 704–713. https://doi.org/10.1037/0012-1649.18.5.704

Saslow, M. G. (1967). Effects of components of displacement-step stimuli upon latency for saccadic eye movement. *Journal of the Optical Society of America*, *57*(8), 1024–1029. https://doi.org/10.1364/JOSA.57.001024

the EASE Team, Thorup, E., Nyström, P., Gredebäck, G., Bölte, S., & Falck-Ytter, T. (2018). Reduced alternating gaze during social interaction in infancy is associated with elevated symptoms of autism in toddlerhood. *Journal of Abnormal Child Psychology*, *46*(7), 1547–1561. https://doi.org/10.1007/s10802-017-0388-0 van Renswoude, D. R., 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

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

How to cite this article: De Kloe, Y. J. R., Hooge, I. T. C., Kemner, C., Niehorster, D. C., Nyström, M., & Hessels, R. S. (2021). Replacing eye trackers in ongoing studies: A comparison of eye-tracking data quality between the Tobii Pro TX300 and the Tobii Pro Spectrum. *Infancy*, 27, 25–45. https://doi.org/10.1111/infa.12441