The Official Journal of the International Society on Infant Studies
Infancy

Infancy, ***17***(1), 9–32, 2012

ISSN: 1525-0008 print / 1532-7078 online
DOI: 10.1111/j.1532-7078.2011.00089.x

WILEY-BLACKWELL

A Critical Test of Temporal and Spatial Accuracy of the Tobii T60XL Eye Tracker

James D. Morgante, Rahman Zolfaghari, and Scott P. Johnson
Department of Psychology
University of California

Infant eye tracking is becoming increasingly popular for its presumed precision relative to traditional looking time paradigms and potential to yield new insights into developmental processes. However, there is strong reason to suspect that the temporal and spatial resolution of popular eye tracking systems is not entirely accurate, potentially compromising any data from an infant eye tracking experiment. Moreover, "best practices" for infant eye tracking, such as knowing which software tool enhances experimental flexibility, remain to be determined. The present investigation was designed to evaluate the temporal and spatial accuracy of data from the Tobii T60XL eye tracker through the use of visual latency and spatial accuracy tasks involving adults and infants. Systematic delays and drifts were revealed in oculomotor response times, and the system's spatial accuracy was observed to deviate somewhat in excess of the manufacturer's estimates; the experimental flexibility of the system appears dependent on the chosen software.

Since the pioneering studies of looking behavior in young infants by Fantz in the late 1950s and early 1960s (e.g., Fantz, 1961), a number of experimental paradigms based on infant looking time have been devised to test infants' detection, discrimination, preference, categorization, learning, and expectations of visual and auditory stimuli (Aslin, 2007). Looking time measures have served as a preferred method for infant testing for several decades and

Correspondence should be sent to James D. Morgante or Scott P. Johnson, Department of Psychology, University of California, 3291 Franz Hall, Box 951563, Los Angeles, CA 90095. E-mail:jmorgante@ucla.edu or scott.johnson@ucla.edu

are the methodological basis for the majority of published works in the field. Yet the interpretation of looking times is fraught with difficulty, for several reasons. First, looking times are susceptible to influence by infant state (e.g., fussiness, sleepiness, and general arousal). Second, infants sometimes exhibit preferences for novelty and sometimes for familiarity; it can be difficult to predict which will prevail in any one instance. Preferences are a function of age, stimulus exposure, difficulty of processing, and so forth (Hunter & Ames, 1988), and linking hypotheses between observations and psychological constructs can be weak (Cohen, 2002). Third, looking time data are vulnerable to experimenter bias, even if unintentional, because they are the result of a judgment produced by a human observer. Finally, looking time data are inefficient. Many infant looking time paradigms yield one or a small number of data points, reflecting infants' preference for one of only a limited number of stimuli.

One candidate method that can overcome some of these difficulties is infant eye tracking. Corneal-reflection eye trackers work by comparing a video input of the pupil with the highlight reflected off the cornea, usually from a light source invisible to humans, in the infrared range of the spectrum. The center of the pupil and the corneal reflection are tracked in real time and provide information about the participant's point of gaze (POG) on the stimulus (see Gredebäck, Johnson, & and von Hofsten, 2010, for additional information about infant eye tracking).

Traditionally, recording eye movements in infants was difficult because of the small field of view necessitated by the need for accurate capture of the pupil. If the infant being tested moved out of view of the camera, there would be a loss of data. Of course, infants do not follow instructions to sit still. Chinrests and head-mounted optics (camera and other components) commonly in use with adult participants are impractical for young populations. However, there has been significant progress in the production of lightweight models appropriate for toddlers (Franchak, Kretch, Soska, & Adolph, in press). The experimenter who wished to use an eye tracker was therefore faced with a dilemma: either to risk fussiness owing to the need to hold the infant tightly or to risk small samples of data owing to movement (Hainline, 1981). Some researchers worked around these limitations by recording infants as they were supine (e.g., Haith, 1980), and this method is still in use in some laboratories, but in general, this is not a solution applicable to a diversity of testing situations (e.g., when infants become ambulatory).

In the mid-1990s, Applied Science Laboratories developed an eye tracker (ASL model 504; Bedford, MA) with "remote optics" employing a table-mounted camera with a pan-tilt mechanism. The eye tracker software incorporated an algorithm for adjusting the camera's direction and focus to

accommodate participant movement, keeping the image of the eye within view automatically. If the infant moved too fast or too far and overloaded the system's capability, the experimenter could take control of the camera and change its direction with a remote control, then switch back to the automatic mode. The next major advance in eye tracking technology most relevant to studying infants came in 2003 when Tobii Technology released an eye tracker (Tobii ET-17; Falls Church, VA) with several key advantages; these advantages are characteristic of the more recent Tobii models as well. First, reacquisition of the eye's image following a tracking loss is facilitated by an enlarged field of view. Second, improvements in video processing allow the Tobii system to track both eyes, not just one. Third, the system provides feedback to the experimenter meant to convey the quality of calibration of the participant's POG; if calibration is determined to be poor, the calibration routine can be repeated. Fourth, there is software built into the Tobii for designing and implementing experiments.

Eye tracking is emerging as an exciting new method in infant studies (Gredebäck et al., 2010). The first paper to use an infant-friendly eye tracker appeared in 2000 (Johnson & Johnson, 2000), and the number of papers that have appeared that use eye tracking methodology has since snowballed, beginning with a special issue of *Infancy* in 2004 (Gredebäck & von Hofsten, 2004; Hunnius & Geuze, 2004; Johnson, Slemmer, & Amso, 2004; McMurray & Aslin, 2004). Eye tracking has been used to examine a variety of perceptual and cognitive phenomena, including categorization (Quinn, Doran, Reiss, & Hoffman, 2009), event prediction (Johnson, Amso, & Slemmer, 2003; Johnson & Shuwairi, 2009), sequence learning (Kirkham, Slemmer, Richardson, & Johnson, 2007), memory (Richmond & Nelson, 2009), object perception (Amso & Johnson, 2006; von Hofsten, Kochukhova, & Rosander, 2007; Johnson, Davidow, Hall-Haro, & Frank, 2008), motion perception (Kato, de Wit, Stasiewicz, & von Hofsten, 2008), face perception (Frank, Vul, & Johnson, 2009; Turati, Valenza, Leo, & Simion, 2005), and social cognition (Falck-Ytter, Gredebäck, & von Hofsten, 2006).

In any burgeoning new field, initial excitement yields to harsh realities. In the case of infant eye tracking, limitations in eye tracking paradigms and equipment are becoming apparent, and they center around two issues: accuracy (temporal and spatial) and usability (calibration and experimental flexibility). *Temporal accuracy* refers to the timing of eye movements relative to stimulus events and is a function of the computer's processing capacity and software. This is relatively unimportant if the dependent measure is an estimate of infant interest in one or more locations on the screen, but vital if the experimenter needs to know about anticipations of, or reactions to, stimulus events. *Spatial accuracy* refers to the relation of the POG as recorded by the eye tracker and the location on the stimulus where the participant is actually

looking. It is an issue of paramount significance in any eye tracking paradigm. If estimates of POG are inaccurate, the data may not be usable. Spatial accuracy is a function of calibration of the participant's POG by the eye tracker, which is accomplished by having him or her look at several known points on the screen, one at a time. Typically, these points consist of small, colorful, visual–auditory stimuli.

Usability, in our view, can be considered to be a function of ease of calibration and experimental flexibility. As previously mentioned, calibration of the POG is accomplished by moving a small attention-getting stimulus to specific locations on the monitor with known *x*-*y* coordinates as the eye tracker computer records the spatial relation of the pupil and corneal reflection. Each participant's POG must be calibrated because each eye is shaped differently. An infant-friendly calibration routine, developed by the third author, was first described in Johnson et al. (2003) and is now provided with every Tobii eye tracker. Similar routines are available on eye trackers made by other manufacturers. Although this calibration procedure is in widespread use, the extent to which it may yield a well-calibrated POG remains unknown. This is an issue we address in experiments described subsequently.

Experimental flexibility varies from eye tracker to eye tracker. Tobii, for example, includes proprietary software (Studio) with which to design and implement experimental protocols, and it works well for some designs, such as fixed-duration trial presentation. Other designs, however, such as infant-controlled habituation, cannot be implemented using Studio. This has led some investigators (including our own laboratory) to turn to outside party software, integrated with the eye tracker, to run experiments. Two common means of doing so are E-Prime (made by Psychology Software Tools Inc., Sharpsburg PA) and MATLAB (made by MathWorks, Natick, MA). In principle, any custom coding program can be integrated with the Tobii software development kit.

The extent to which these software tools might be considered by infant researchers and the accuracy of experimental results that stem from their interface with Tobii Studio are unknown at present. Because of the popularity of Tobii eye trackers in the investigations of cognitive, perceptual, and social development in human infants, we decided to evaluate the temporal and spatial accuracy of the Tobii T60XL using Tobii Studio (version 2.2.8) and E-Prime (version 2.0) software.

OVERVIEW OF METHODS

In the four evaluations that follow, we describe the tests of temporal and spatial accuracy with adult and infant participants. Participants were tested

in a darkened room, seated alone (adults) or in a caregiver's lap (infants), with the eyes approximately 65 cm from the eye tracker optics (per manufacturer instructions). The optics are built into the frame that supports the monitor viewed by the participant. Infrared lights are also built into this frame to provide illumination producing a corneal reflection. The Tobii monitor is supported by an Ergotron model LX LCD Mount Arm (Ergotron Inc., St. Paul, MN) bolted to a table. The mount arm allows the experimenter to manipulate the position of the monitor to ensure an ideal field of view for the eye tracker camera, by bringing the monitor to the participant's face.

Evaluation 1 (temporal accuracy) consisted of a comparison of two kinds of data output from Tobii Studio: (a) a *combined* file that contains data about *x*-*y* coordinates of each participant's POG at each time stamp (60 Hz) as well as the stimulus being shown and other information and (b) a video recording of the experimental session consisting of the POG superimposed on the stimulus, exported as an audio–video interleaved file, henceforth referred to as *.avi*. Evaluation 2 (temporal accuracy) was similar to Evaluation 1 except that E-Prime was used to control stimulus presentation and collation of data output from the Tobii server on the same computer. Data (POG coordinates) were output to a second computer running Tobii Studio which was then used to produce the .avi for subsequent coding. Frame-by-frame coding of the .avi for Evaluations 1 and 2 was completed by the third author. Evaluations 3 and 4 (spatial accuracy) entailed a *calibration check* consisting of a series of concentric circles, moving inward so as to direct the observer's POG toward the center; adults and infants, respectively, participated in the two studies.

Calibration

Each participant's POG was calibrated using the five-point "infant calibration" routine developed by the third author and provided by the manufacturer. Tobii Studio provides the option of a nine-point calibration, but we elected to use the five-point routine because we have found that it works well for infant participants (the calibration stimulus itself is highly attractive) and we presumed that most laboratories using Tobii eye trackers with infants have followed suit. In addition, we wished to directly compare populations of adults and infants in terms of calibration accuracy (Evaluations 3 and 4). The POG was calibrated by comparing it to known coordinates on the screen as the participant viewed a target-patterned "attention-getter," looming/contracting in synchrony with a rhythmic sound. The attention-getter was presented at five locations on the monitor, and the participant looked at each in turn. Generally, the calibration routine was completed in

less than a minute. Completion of the routine ranged from several seconds to a few minutes, say if the experimenter needed to readjust the participant's positioning relative to the eye tracker or deemed a calibration unacceptable such that recalibration was necessary.

Tobii eye trackers provide a pictorial representation of the quality of calibration after a participant looks at the five stimuli presented during the calibration routine. If calibration quality at any single location is high, a small dot appears in the center of a circle corresponding to that position on the monitor (Figure 1). Ideally, therefore, a five-point calibration routine results in five single dots, each within their respective circles. Less-than-ideal calibrations are represented by missing points and/or by colored lines that extend from one or more points. Presumably the length, number, and dispersion of these lines bear some relation to a mismatch between the "true" POG that was recorded at that moment during the calibration and the actual location of the calibration point, but we are aware of no empirical verification of this possibility. The acceptability of the calibration is determined by visual inspec-

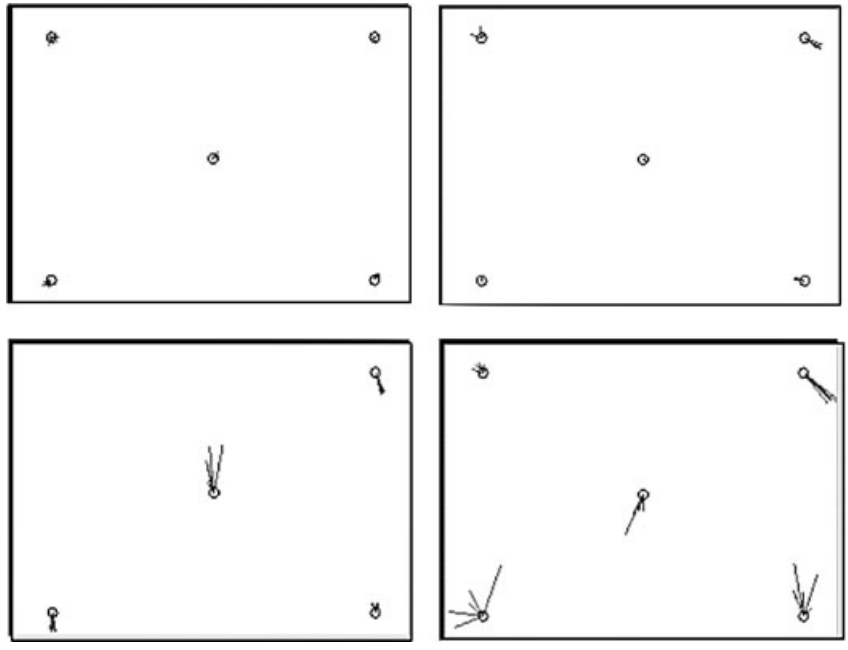

Figure 1 Representations of calibration quality provided by the Tobii eye tracker. The top two panels depict the best possible calibrations; the bottom two panels reflect less-than-ideal calibrations (missing points and lines extending from different points). These representations were taken from saved calibrations performed with a Tobii model 1750 because Tobii Studio does not save them.

tion. Even if lines grossly protrude or points are missing, the system will not alert the user. Irrespective of its acceptability, the experimenter can proceed with the obtained calibration file and begin the experiment. In the following evaluations, all adult and infant participants had a pictorially represented calibration file that was complete, with all five points containing dots within or nearly within the circles denoting spatial location (as shown in Figure 1).

EVALUATION 1: TEMPORAL ACCURACY WITH TOBII STUDIO

The goal of the first evaluation was to compare latencies in the combined file and the .avi, each exported from Tobii Studio. We used a *visual latency* task designed to assess oculomotor reaction time, implemented in Tobii Studio.

Method

Participants

Twenty-seven UCLA undergraduates were observed, recruited from a pool of participants managed by the Department of Psychology; they received course credit for participation. We excluded data from 12 adults (five women, seven men; *M* age = 21 years, 6 months) because of low data yield (data were obtained from < 80% of the time that stimuli were presented), leaving 15 participants (12 women, three men; *M* age = 21 years, 4 months) with a data yield of 80% or greater for the visual latency task. Seven adults had a yield in the 1–50% range, and five adults had a yield in the 51–79% range.

System configuration

Tobii Studio was installed on a Dell PC running Windows XP (SP 3) with a 2.83-GHz Intel core 2 Quad processor, a 500-GB (7,200 rpm) hard drive, 4 GB of RAM, and an NVIDIA GeForce 9500GT 1-GB graphics card (Dell Inc., Round Rock, TX). The firewall was off and there were no other operations running in the background. Studio received gaze coordinates from the eye tracker server when recording commenced. The monitor was a 24-inch thin film transistor liquid crystal display, 1,920 × 1,200 pixels.

Procedure

Oculomotor reaction time was assessed with a visual latency task modified from a paradigm previously used to investigate the development of visual attention (Amso & Johnson, 2005). An attention-getting object

moving in synchrony with a sound was presented in one of four locations for 1 sec, followed 500 msec later by a second object appearing in one of the other three locations (Figure 2). Object locations were determined randomly for each trial, with the stipulation that the second object always appeared in a different location than the first. Each location was contained by a 9.7 × 9.5 cm (8.5 × 7.6° visual angle at the participant's 65 cm viewing distance) white rectangle in a grid-like arrangement. The timing of the appearance of the second object was therefore predictable, but the location was not, which we reasoned would minimize the possibility of oculomotor anticipations (although some anticipations would be expected by chance). There was a temporal gap of 1 sec between trials, and 36 trials were presented, for a total task duration of about 2 min 5 sec.

Results and discussion

Data consisted of oculomotor latencies to move the POG into the location of the second attention-getter presented on each trial, from a position outside this location. Only a single latency was entered into the data set for any given trial—the first time at which the POG changed position from outside the location to inside. The location was defined as the square whose boundary surrounded the object (see Figure 3). The latency was defined as the difference in msec between the onset of the second attention-getter and the end of the previous fixation, before the shift of gaze into the relevant location.

Data for each participant (S2, S4, etc.) and each trial (1–36) are shown in Figure 3. Most latencies are in the 200- to 400-msec range, which is to be expected given that it takes a minimum 150 msec to program an eye movement under most circumstances (Collewijn & Tamminga, 1984; Fischer & Weber, 1993). Latencies from the Studio combined file (open circles) and .avis (black diamonds) were closely matched on a trial-by-trial basis, yet for the majority of the trials, latencies as output by Studio in the combined file

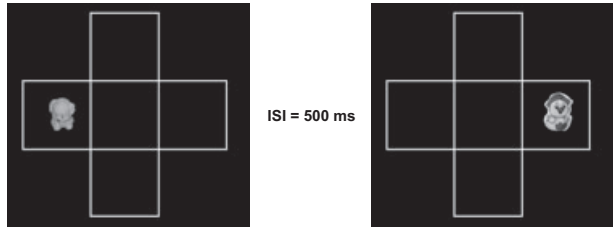

Figure 2 The visual latency task. An attention-getter appears in one of four locations for 1 sec, followed 500 msec later by a second attention-getter in one of the other three locations. The second location cannot be predicted.

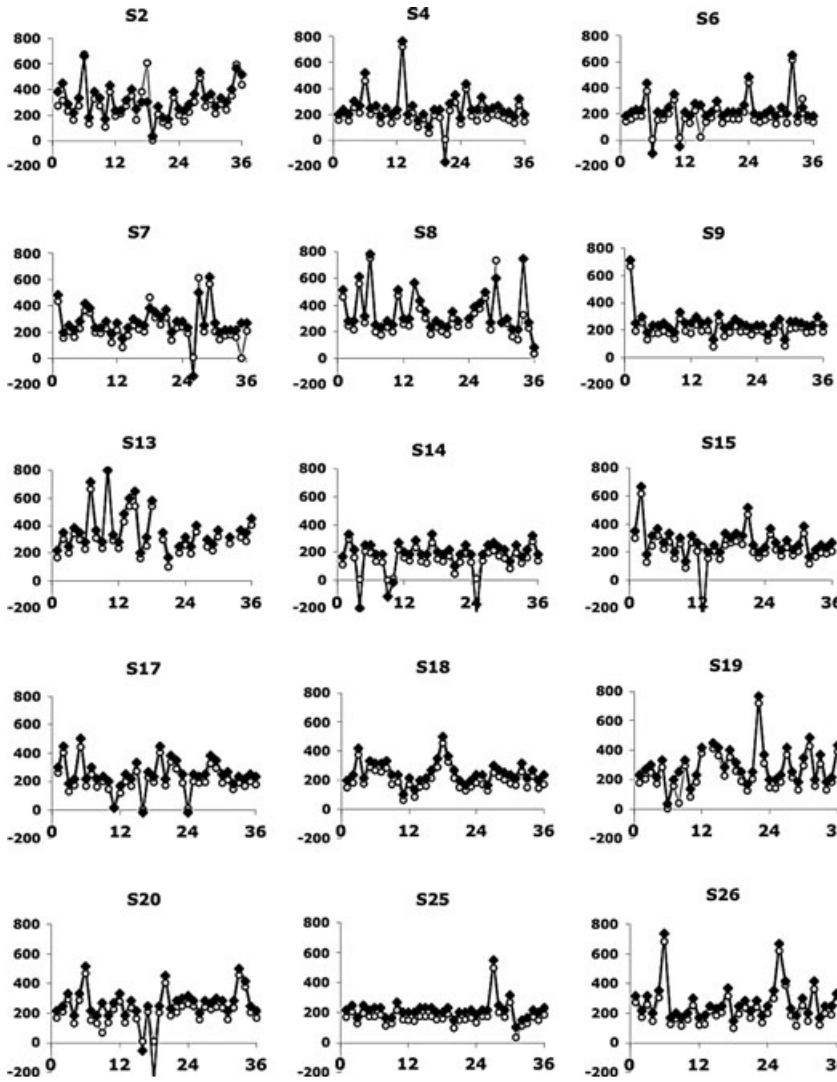

Figure 3 Results from the visual latency task for each of the 15 participants in Evaluation 1. The *x*-axis shows trial (1–36), and the *y*-axis shows oculomotor latency in msec. Open circles: data from Tobii Studio, coded automatically. Filled diamonds: data from the .avi exported from Tobii Studio, coded by the third author.

were consistently longer than those in the .avis (M offset across participants = 44.5 msec, SD = 7.3, range = 27.0–54.5 msec). The mean offset was reliably different than zero, $t(14) = 23.6$, $p < .0001$.

In summary, latencies in the Tobii Studio combined file and in the .avi recording were similar but not identical, with a systematic delay in data from the combined file.

EVALUATION 2: TEMPORAL ACCURACY WITH E-PRIME

The goal of the second evaluation was to compare latencies in the output file produced by E-Prime and the .avi produced by Tobii Studio. We used the same visual latency task to assess oculomotor reaction time as described previously, designed in and implemented by E-Prime.

Methods

Participants

Thirty-four UCLA undergraduates were observed, recruited from the same participant pool as described previously. We excluded data from 19 adults (15 women, four men; *M* age = 21 years, 2 months), leaving 15 participants (13 women, two men; *M* age = 21 years, 4 months) with a data yield of 80% or greater for the visual latency task. One adult could not be calibrated after several unsuccessful attempts and, subsequently, was not tested in the visual attention task. Two adults were excluded because their calibration could not be validated with our spatial accuracy task, which is described in Evaluation 3. Eight adults had a yield in the 1–50% range, and six adults had a yield in the 51–79% range. One adult had both an acceptable calibration and a visual behavior data yield >80%, but her viewing behavior made it difficult to score for latencies because she never attended to the targets. The final adult was excluded because of a problem with the stimulus display. Adults who participated in Evaluation 2 also participated in Evaluation 3, described subsequently.

System configuration

Two computers were linked together via an Ethernet hub. The first computer, the Dell PC (previously described), ran E-Prime, displayed the stimuli, and obtained gaze coordinate data from the Tobii server. It sent commands to a second computer, a Mac Pro with a Bootcamp partition running Windows XP (SP 3) (Apple Inc., Cupertino, CA), to begin and end the Studio recording and insert event markers into the recording. The firewall was off and there were no other operations running in the background. Computer 2 ran Studio and received gaze coordinates from the eye tracker

with a Datapath Vision RGB DVI Capture Card. That is, the second computer captured the stimulus that E-Prime presented and mapped gaze coordinates onto it within Tobii Studio. The system had a 2.66-GHz Xenon processor with a 100-GB (7,200 rpm) hard drive, 2 GB of RAM, and an NVIDIA GeForce 7300 256-MB graphics card.

Procedure

The design and procedure were the same as those in Evaluation 1 except that the task was implemented in E-Prime.

Results and discussion

Data again consisted of oculomotor latencies to move the POG into the location of the second attention-getter presented on each trial, from a position outside this location (see Figure 2). Data for each participant and each trial are shown in Figure 4. Latencies from E-Prime (open circles) appear to show similar response times to those from Evaluation 1, but latencies from the .avis (black diamonds) show a consistent decrease in response times with trials for each subject.

Because these data were completely nested within participant, we used generalized estimating equations to quantify the discrepancy between E-Prime and .avi data by analyzing latency as a function of condition (E-Prime versus .avi), trial (1–36), and their interaction. We effect coded and centered condition (E-Prime = −.5, .avi = .5) and centered trial. We report unstandardized regression coefficients (B) and Wald Zs for each parameter, and we employ standard regression vernacular to describe the effects (Diggle, Liang, & Zeger, 1994; Fitzmaurice, Laird, & Ware, 2004; Liang & Zeger, 1986).

Based on inspection of Figure 4, our goal for this analysis was to assess the likelihood that latency would vary as a function of condition across trials. We regressed latency onto condition, trial, and their interaction to test for this possibility. Overall, each trial corresponded to a 7.59-msec lower latency, which resulted in a significant effect for trial, $B = -7.60$, $SE = .47$, $z = -16.28$, $p < .0001$. Latencies in the .avi condition were lower than in the E-Prime condition, resulting in a significant effect for condition, $B = -325.76$, $SE = 11.43$, $z = -28.49$, $s < .0001$. The interaction was strong and statistically significant, $B = -14.97$, $SE = .84$, $z = -17.80$, $p < .0001$. We examined the nature of this interaction by testing the simple slopes for each condition. Within the E-Prime condition, latency was unrelated to trial number, Simple $B = -0.11$, $SE = .29$, $z = -0.38$, $p = .70$. Within the .avi condition, however, each trial corresponded to a 15.08-msec lower latency, Simple $B = -15.08$, $SE = .84$, $z = -17.94$, $p < .0001$.

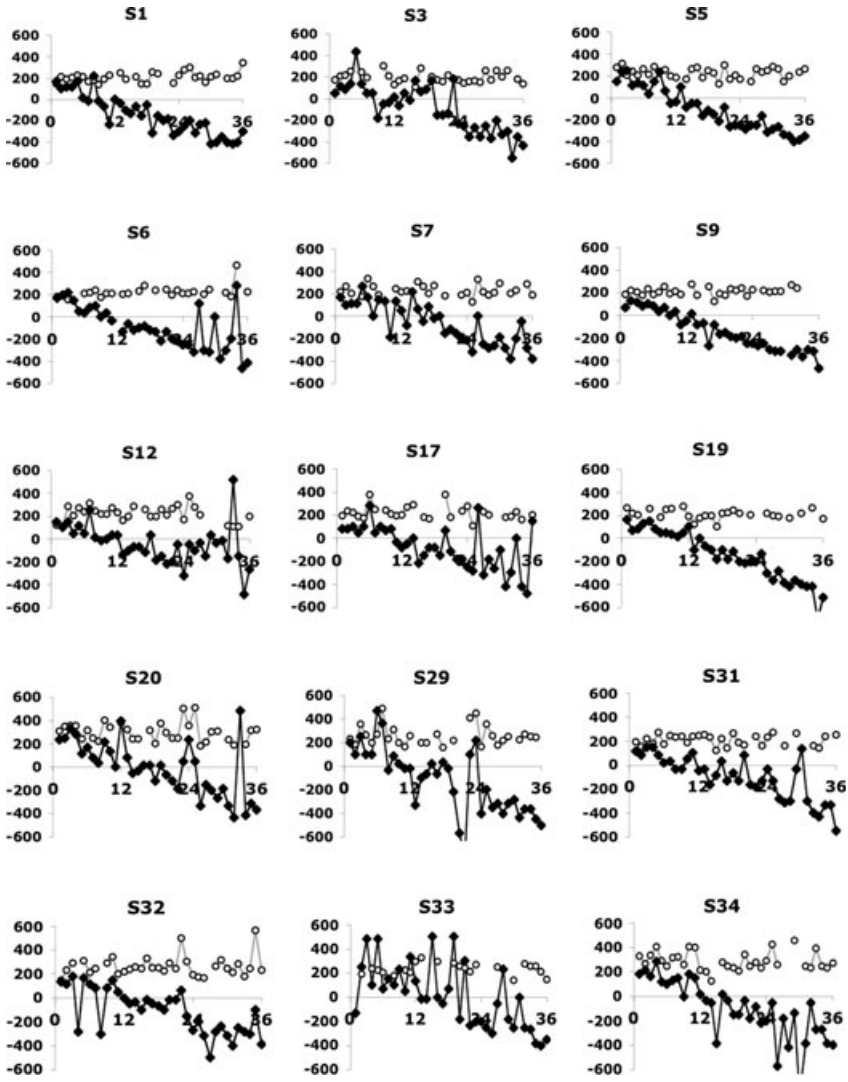

Figure 4 Results from the visual latency task for each of the 15 participants in Evaluation 2. The *x*-axis shows trial (1–36), and the *y*-axis shows oculomotor latency in msec. Open circles: data from E-Prime, coded automatically. Filled diamonds: data from the .avi exported from Tobii Studio, coded by the third author. Note the systematic reduction in latencies across trials in each participants' data from the .avi. This is unlikely to reflect their actual behavior, because it would indicate the ability to predict the object's location, which was determined randomly.

In summary, data from the E-Prime output file and the Tobii Studio .avi were significantly different. There was a clear, systematic bias in the .avi that appears to reflect a reduction across trials in oculomotor latencies, so much so that for most participants, it seems that they came to predict the location of the second object's appearance after a dozen or so trials. In our view, this does not provide an accurate reflection of participants' behavior, because it is not reasonable to assume that they could know this location in advance. Instead, this effect is artifactual and stems from error in the system.

EVALUATION 3: SPATIAL ACCURACY WITH ADULTS

The goal of the third evaluation was to assess the spatial accuracy of the Tobii T60XL. We compared the spatial location of the POG to nine known positions on the screen, presented one at a time, as each participant viewed them in turn.

Method

Participants

Participants were the same 15 adults who participated in Evaluation 2.

System configuration

The system configuration was identical to that described for Evaluation 2.

Procedure

Adults' POG was calibrated as described previously. The *spatial accuracy* task involved a set of brightly colored annuli that shrank to a center point (see Figure 5). This stimulus was devised by Michael C. Frank at Stanford University and is available at http://langcog.stanford.edu/materials/calib_check.avi. Adults' visual behavior was recorded as the annulus moved to nine points on the screen. The annulus appeared first at point 5 (center), then to points 1, 7, 3, and 9, then back to point 5, then to points 8, 6, 2, and 4, and finally back to point 5 for a third time. The center of each point was separated by 5.8 cm (5.1°) (point 4 was displaced vertically by about .3°). The spatial accuracy task was completed twice, once immediately following

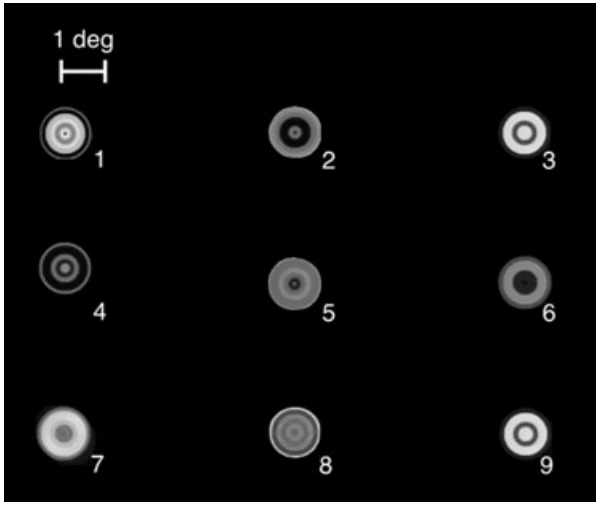

Figure 5 Schematic depiction of the spatial accuracy task. An annulus consisting of concentric circles shrinking to a central point was presented at nine positions on the screen one at a time. All points are shown in the figure for clarity. The scale (top left) and numbers next to each point were not present in the stimulus. See text for details.

calibration and again following the temporal accuracy task described in Evaluation 2. Each task lasted 30 sec.

Results and discussion

Data consisted of the *x*-*y* coordinates of the single longest fixation that was recorded while the annulus stimulus was in each of the nine positions. Accuracy was operationalized as the deviation in degrees of visual angle of the locations of the POG and the annulus, with greater accuracy characterized by smaller deviations.

Accuracy data are presented in Figure 6. Each participant is represented by a different symbol (open triangle, filled triangle, x, cross, and so forth), and the locations of the annulus are represented by the large open circles. In most cases, the POG is offset relative to the annulus, with many tending to cluster below it. Accuracy data were analyzed with an 11 (location) $\times$ 2 (presentation: first or second) repeated-measures analysis of variance (ANOVA); the dependent measure was deviation of the POG expressed in degrees of visual angle. There were no reliable effects, indicating that there was not a systematic tendency for some

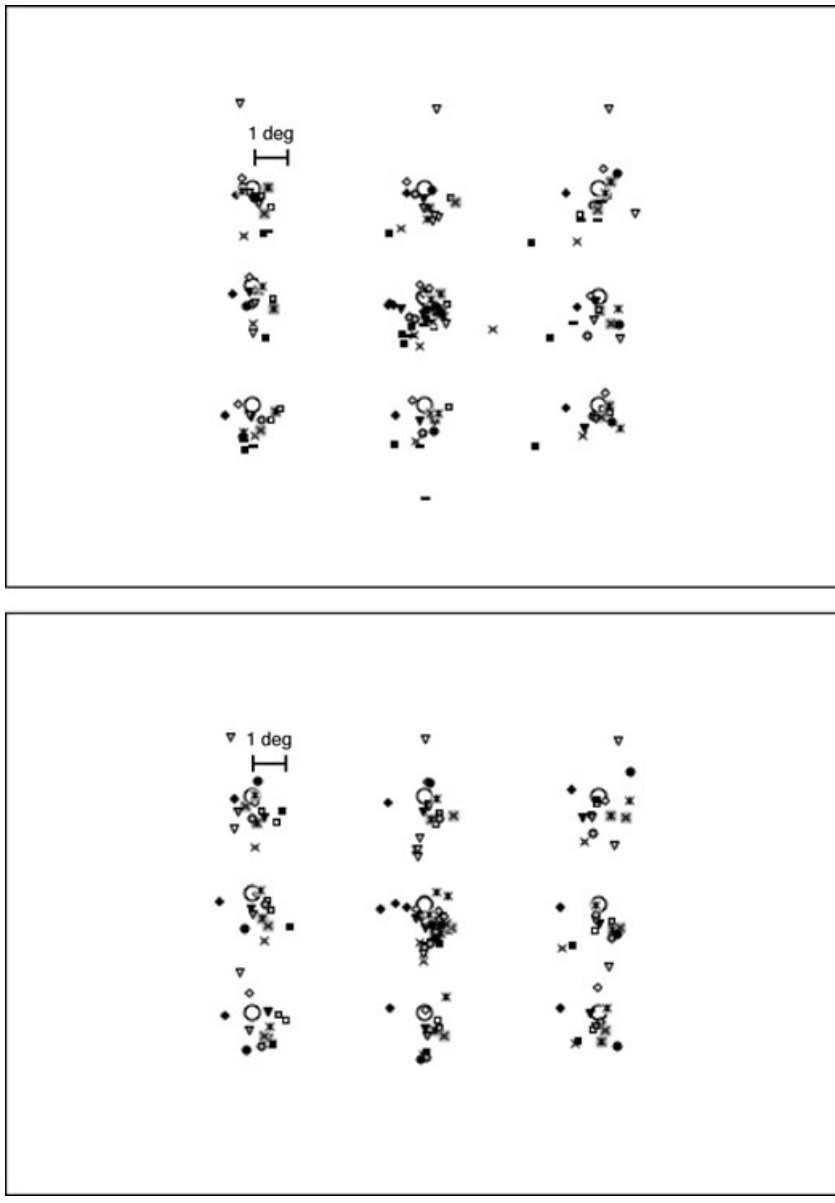

Figure 6 Results from Evaluation 3 (spatial accuracy) with adults. The location of each adult's longest single fixation during the presentation of each annulus is shown. Each of the 15 participants is represented by a unique symbol, 11 in all (three in the center location and one in the other locations). The locations of the annuli are shown as large open circles.

locations to be calibrated more accurately than others, and calibration quality was not reliably different during the second presentation (i.e., no evidence of drift or deterioration of calibration quality). The *M* deviation per adult participant was 1.27° visual angle (*SD* = .73), range = .51–3.28°.

EVALUATION 4: SPATIAL ACCURACY WITH INFANTS

The goal of the fourth evaluation was to assess the spatial accuracy of the POG in a sample of infant participants.

Method

Participants

Thirty-seven infants were observed, recruited from the greater Los Angeles area with a letter sent to new parents who then returned a postcard indicating interest in participation in infant research. We included data from the 15 infants (11 girls, four boys; *M* age = 10 months, 17 days, *SD* = 110 days) whose data yield was the most comparable to the observed adults. Twelve of the infants had both a spatial accuracy and visual latency task data yield of 80% or greater; they met the established adult criteria. Four had a data yield of 73% or greater for the spatial accuracy task and 81% or greater for the visual latency task. However, the three infants included had a visual latency data yield of 85% or greater and, overall, were the closest to meeting the established criteria; the fourth was subsequently excluded.

We excluded data from an additional 22 infants (six girls, 16 boys; *M* age = 6 months, 22 days, *SD* = 122 days) because of low data yield or fussiness. Five infants had an overall yield in the 1–50% range, and 10 infants had a yield in the 51–79% range. Six were excluded because they either cried or fussed and did not complete the session. It should be noted that 11 of the excluded infants were between 3 and 4 months of age. Although we recruited infants from 3 to 18 months, we had little success ($n = 1$) with infants < 6 months old with these particular methods and criteria.

System configuration and procedure

The system configuration and procedure were identical to those described for Evaluation 3. Data were collected during the visual latency task but are not reported here.

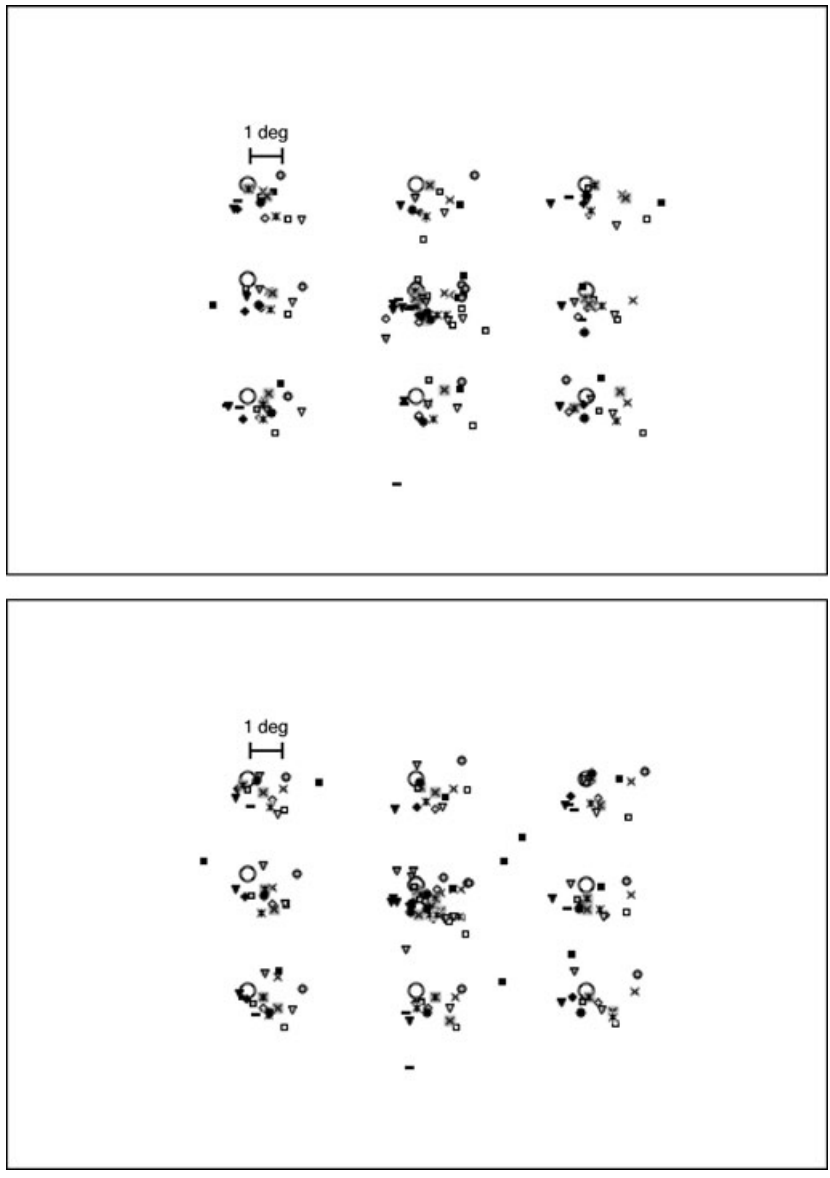

Figure 7 Results from Evaluation 4 (spatial accuracy) with infants. The location of each infant's longest single fixation during the presentation of each annulus is shown. Each of the 15 participants is represented by a unique symbol, 11 in all (three in the center location and one in the other locations). The locations of the annuli are shown as large open circles.

Results and discussion

Accuracy data are presented in Figure 7. As in Evaluation 3, each participant is represented by a different symbol (open triangle, filled triangle, x, cross, and so forth), and the locations of the annulus are represented by the large open circles. Accuracy data (deviation of the POG) were again analyzed with an 11 (location) × 2 (presentation: first or second) repeated-measures ANOVA, and as with adults, there were no reliable effects. The *M* deviation per infant participant was 1.22° visual angle (*SD* = .44), range = .52–2.17°.

Accuracy data from adults (Evaluation 3) and infants (Evaluation 4) were compared with a 2 (age group) × 11 (location) × 2 (presentation) mixed ANOVA, with repeated measures on the second and third factors. There were no significant main effects or interactions, indicating that calibration accuracy for adults and infants did not differ reliably.

GENERAL DISCUSSION

Our goal in the present investigation was to evaluate the temporal and spatial accuracy of data from the Tobii T60XL eye tracker, which we accomplished with visual latency and spatial accuracy tasks involving adults and infants. A subsidiary goal was to evaluate usability, specifically in terms of pragmatic testing issues—ease of calibration and versatility of the system for use with multiple software platforms and experimental designs. Each issue is discussed in turn subsequently. We have done our best to provide an objective means of evaluation of the eye tracking system, but a part of our evaluation is necessarily subjective and reflects our own views.

Temporal accuracy

Evaluations 1 and 2 provided two different means of comparing data from an output file (Tobii Studio and E-Prime, respectively) with the .avi recording exported from Studio. It is our opinion that a video record is a vital means of interpreting infant behavior, often yielding insights that are not possible when data are coded automatically from an output file. Two examples from our own laboratory illustrate this point. First, systematic investigations of 4- and 6-month-olds' oculomotor anticipation (Johnson et al., 2003) were motivated by reviews of eye movement recordings as young infants were presented with repetitive object trajectory events; few infants provided evidence of consistent predictive behavior. This seemed surprising in light of extensive claims of early object concepts (e.g., Baillargeon, 1995),

but these informal observations were confirmed by empirical testing (Johnson et al., 2003). Second, systematic investigations of 3-, 6-, and 9-month-olds' attention to faces in complex scenes (Frank et al., 2009) were motivated by reviews of eye movement recordings as young infants watched a *Charlie Brown Christmas* cartoon originally intended to keep infants' interest while participating in an unrelated paradigm; infants seemed largely uninterested in faces under these conditions. This seemed surprising in light of previous evidence for early preferences for faces and schematic, face-like stimuli (e.g., Morton & Johnson, 1991), but these informal observations too were corroborated by empirical testing and we were able to quantify visual attention in terms of low-level salience and age-related entropy in the focus of attention (Frank et al., 2009).

Evaluation 1 revealed a systematic delay in oculomotor response times in the Tobii Studio combined file when compared with the .avi export file. Evaluation 2 revealed a systematic drift in oculomotor response times that was introduced into the .avi when data from E-Prime served as input to Studio. It is not clear what might account for these discrepancies, especially for Evaluation 1, because the .avi is presumably generated from the same information that is contained in the combined file. For Evaluation 2, we considered the possibility that errors in updating information from E-Prime were introduced by limits in processing capacity of the second computer's CPU and other components (viz., the computer running Studio, which generated the .avi). Notably, however, the CPU speed (2.66 GHz) exceeded the manufacturer's recommendations (2.0 GHz, as specified in http://www.tobii.com/Global/Analysis/Downloads/Product_Descriptions/Tobii_System_Recommendations.pdf) and, in our view, therefore, the capacity should have been more than adequate to receive and process gaze coordinate data as input to Tobii Studio. Regardless of the source of the error, our results indicate the need for caution when viewing and coding video records from Tobii Studio. These records can be useful for the interpretation of infant behavior in many circumstances, but latency data from Studio may not be valid. Even in Evaluation 1, in which data from the combined file and the .avi were similar across trials (Figure 3), there was a reliable offset such that the combined file latencies were longer.

Evaluation 1 suggests that eye movement latency measures might deviate up to 54 msec depending on the method of analysis. Clearly, this deviation presents a challenge for the interpretation of results from studies for which the timing of anticipatory or reactive eye movements is an important dependent variable, for two reasons. First, the absolute deviation cannot be known and second, a deviation of this magnitude risks inaccurate classification of some eye movements as either anticipatory or reactive, if they are close to the "cutoffs" of 150 msec (Amso & Johnson, 2005; Canfield, Smith,

Brezsnyak, & Snow, 1997; Johnson et al., 2003) or 200 msec (Bertenthal, Longo, & Kenny, 2007; Gredebäck & von Hofsten, 2004; Kochukhova & Gredebäck, 2007) that have been used in previous coding schemes. That is, an eye movement produced prior to or within 150 msec of stimulus onset would be classified as an anticipation, because it takes about 150 msec to program a saccade, and an eye movement that takes longer would be classified as a reaction. However, the deviation is quite consistent, both within and across Evaluations 1 and 2, and this suggests that statistical comparisons between conditions (using the same analysis method) may be valid.

The disconnect in latencies taken from E-prime versus .avi becomes especially acute when multiple computers and programs control stimuli and data collection, as in our Evaluation 2. In our view, .avi analysis should not be used when the Tobii is connected with E-Prime. At a more general level, the results suggest that researchers must make sure that they know (or at least investigate) the effects of a particular hardware configuration before running and publishing studies.

Spatial accuracy

Evaluations 3 and 4 tested the accuracy of calibration of the POG in adult and infant participants whose calibrations, according to the pictorial representation of calibration quality provided by Tobii Studio, were as optimal as could be achieved (see Figure 1). The deviation between recorded POGs and points in the calibration check stimulus averaged 1.27° for adults and 1.22° for infant participants, with no age differences in accuracy. This deviation is somewhat in excess of the manufacturer's estimates of approximately .95°, a combination of accuracy (what we tried to measure), drift (caused by lighting changes), and spatial resolution (noise), as specified in http://www.tobii.com/Global/Analysis/Downloads/Product_Descriptions/Tobii_TX_Product_description.pdf, but not by much. These three sources of deviation are listed in the product description as contributing a "typical" .5°, .1°, and .35° of error, respectively. It might be that using the nine-point routine yields more accurate calibrations, a possibility that awaits empirical testing, but this might also have the unfortunate effect of reducing the number of infant participants, given the requirement of attending to more points during calibration. Notably, the *best* (smallest) average deviations we observed for any one participant were .51° and .52° for an adult and an infant participant, respectively. The largest average deviations exceeded several degrees.

It is not clear from the present procedure how one could predict the magnitude or angular direction of calibration deviation at any one location for any particular participant, but it is apparent that most deviations were downward from the calibration points (see Figures 6 and 7). As noted

previously, there was no indication in the calibration assessment provided by Studio (Figure 1) of any systematic discrepancy at any of the five points for any participant, because we commenced data collection only after achieving ideal calibration representations. We are left to conclude that calibration quality is close to the manufacturer's estimates, but the extent to which any individual participant can be expected to conform to these estimates cannot be known with great certainty and the consistent spatial error implies that interpreting data from relatively small "areas of interest" (regions in the stimulus within which the focus of attention is a principal dependent measure) is best undertaken with caution.

Usability

The Tobii eye tracker is popular with infant researchers because it is easy to set up and calibrate, and many of its features, such as the analysis tools packaged with Tobii Studio, are intuitive and user-friendly. Previously, we noted that two important aspects of usability are ease of calibration and flexibility. We have found that calibrations are generally quick and straightforward, and it is not difficult to achieve calibration representations that appear to be ideal or nearly so with most populations (Figure 1). An exception is infants younger than 6 months, for whom the available testing window can be brief, and who often will not tolerate repeated calibration attempts. However, a substantial portion of the participants we observed had poor data yields following calibration. Twelve (44%) of the 27 adults we observed in Evaluation 1 provided $<80\%$ data, and seven (26%) provided $<50\%$ data. The task lasted just over 2 min, and all adults were cooperative and appeared to pay attention. Similar proportions of low data yield were found in Evaluations 2 and 4 (adults and infants, respectively). We have no concrete explanation for these effects—in our experience, we simply cannot get much data from some participants, and factors such as age, sex, eye color, and ethnicity seem to be irrelevant. We have had greater data yields when using static images with Tobii Studio. For example, we recently observed 54 UCLA undergraduates in a face perception study and excluded only four; two could not be calibrated after several unsuccessful atempts, one provided $<80\%$ data, and the fourth attended to the targets $<50\%$ of the time. The conditions under which participants were observed were identical to those described in the present report. Additionally, we have noted anecdotally that data yields are higher with an older Tobii model (1750) used in our laboratory, and low yields are more characteristic of Tobii Studio.

Experimental flexibility of the Tobii T60XL is greatly enhanced by the use of outside party software and Tobii's own freely available software

development kit, and there are users worldwide who share applications and tips. A partial list of applications can be found at http://appmarket.tobii.com/wiki/index.php/Application_Market_for_Tobii_Eye_Trackers. As noted previously, two popular applications currently in use in many eye tracking laboratories are E-Prime and MATLAB. E-Prime has a user-friendly graphical user interface (GUI) designed for psychology research, and it requires minimal programming knowledge to set up general experiments. Its sole purpose is the design and implementation of experiments. An important advantage of E-Prime over other software is its "packaged" support for integration with Tobii eye trackers. MATLAB, in contrast, is a general-purpose computing tool. It is not GUI based, although some available MATLAB-based programs, such as the Saliency Toolbox (http://www.saliencytoolbox.net), have an interface. For most functions, MATLAB requires writing code, and it can be intimidating for those who have little experience with programming. In contrast to E-Prime, MATLAB is used in other disciplines, and there is an active MATLAB community (http://www.mathworks.com/matlabcentral/). In summary, E-Prime is geared toward the novice programmer and MATLAB toward more experienced programmers. It is more flexible and powerful, but has a steeper learning curve.

CONCLUDING REMARKS

In some ways, our evaluations raise more questions than they answer. What were the sources of the systematic errors in timing of oculomotor latencies in Evaluations 1 and 2? Are there timing errors in other Tobii models? Are there solutions to this problem? Why was there such a disconnect between calibration representations of our participants and the results of the spatial accuracy test? What can be done to ensure the most spatially accurate data possible? What leads to data loss for some participants? We do not know the answers to these questions, but we believe that investigations such as those in the present report are a good place to start.

We would raise two final issues. First, we have not yet formally evaluated other eye trackers for temporal and spatial accuracy, and so direct comparisons in terms of performance parameters that we have described are not yet possible. Our evaluations were conducted with a specific model of eye tracker and a specific operating system, on specific hardware. The extent to which these particular accuracy values that we report will extend to other eye trackers is not known. Second, we hope that this report will motivate the infant research community to think

carefully about the kinds of dependent measures we use as the basis for our science. In our view, part of the excitement about infant eye tracking is the potential to supplant human observers as the principal means of gathering data about infants' visual behaviors, but the field is not well served if the results are not accurate.

ACKNOWLEDGMENTS

This work was supported by NIH Grant R01-HD40432 and a grant from the McDonnell Foundation. The authors are indebted to Michael C. Frank, Kerri L. Johnson, Bryan Nguyen, Zoe Samson and the UCLA Babylab crew, and the infant participants and their parents for their contributions to this research.

REFERENCES

Amso, D., & Johnson, S. P. (2005). Selection and inhibition in infancy: Evidence from the spatial negative priming paradigm. *Cognition*, *95*, B27–B36.

Amso, D., & Johnson, S. P. (2006). Learning by selection: Visual search and object perception in young infants. *Developmental Psychology*, *6*, 1236–1245.

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

Baillargeon, R. (1995). A model of physical reasoning in infancy. In C. Rovee-Collier & L. P. Lipsitt (Eds.), *Advances in infancy research* (Vol. 9, pp. 305–371). Norwood, NJ: Ablex.

Bertenthal, B. I., Longo, M. R., & Kenny, S. (2007). Phenomenal permanence and the development of predictive tracking in infancy. *Child Development*, *78*, 350–363.

Canfield, R. L., Smith, E. G., Brezsnyak, M. P., & Snow, K. L. (1997). Information processing through the first year of life: A longitudinal study using the Visual Expectation Paradigm. *Monographs of the Society for Research in Child Development*, *62* (2, Serial No. 250).

Cohen, L. B. (2002). Extraordinary claims require extraordinary controls. *Developmental Science*, *5*, 211–212.

Collewijn, H., & Tamminga, E. P. (1984). Human smooth and saccadic eye movements during voluntary pursuit of different target motions on different backgrounds. *Journal of Physiology*, *351*, 215–250.

Diggle, P. J., Liang, K. Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford: Oxford University Press.

Falck-Ytter, T., Gredebäck, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience*, *9*, 878–879.

Fantz, R. L. (1961). The origin of form perception. *Scientific American*, *204*, 66–72.

Fischer, B., & Weber, H. (1993). Express saccades and visual attention. *Behavioral and Brain Sciences*, *16*, 553–610.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. New York: John Wiley and Sons.

Franchak, J. M., Kretch, K. S., Soska, K. C., & Adolph, K. E. (in press). Head-mounted eye-tracking: A new method to describe the visual ecology of infants. *Child Development*.

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

Gredebäck, G., Johnson, S. P., & von Hofsten, C. (2010). Eye tracking in infancy research. *Developmental Neuropsychology*, *35*, 1–19.

Gredebäck, G., & von Hofsten, C. (2004). Infants' evolving representation of moving objects between 6 and 12 months of age. *Infancy*, *6*, 165–184.

Hainline, L. (1981). An automated eye movement recording system for use with human infants. *Behavior Research Methods & Instrumentation*, *13*, 20–24.

Haith, M. M. (1980). *Rules that babies look by*. Hillsdale. NJ: Erlbaum.

von Hofsten, C., Kochukhova, O., & Rosander, K. (2007). Predictive tracking over occlusions by 4-month-old infants. *Developmental Science*, *10*, 625–640.

Hunnius, S., & Geuze, R. H. (2004). Developmental changes in visual scanning of dynamic faces and abstract stimuli in infants: A longitudinal study. *Infancy*, *6*, 231–255.

Hunter, M. A., & Ames, E. W. (1988). A multi factor model of infant preferences for novel and familiar stimuli. In L. P. Lipsitt (Ed.), *Advances in child development and behavior* (pp. 69–95). New York: Academic Publishers.

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

Johnson, S. P., Davidow, J., Hall-Haro, C., & Frank, M. C. (2008). Development of perceptual completion originates in information acquisition. *Developmental Psychology*, *44*, 1214–1224.

Johnson, S. P., & Johnson, K. L. (2000). Early perception-action coupling: Eye movements and the development of object perception. *Infant Behavior & Development*, *23*, 461–483.

Johnson, S. P., & Shuwairi, S. M. (2009). Learning and memory facilitate predictive tracking in 4-month-olds. *Journal of Experimental Child Psychology*, *102*, 122–130.

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*, 185–201.

Kato, M., de Wit, T. C., Stasiewicz, D., & von Hofsten, C. (2008). Sensitivity to second-order motion in 10-month-olds. *Vision Research*, *48*, 1187–1195.

Kirkham, N. Z., Slemmer, J. A., Richardson, D. C., & Johnson, S. P. (2007). Location, location, location: Development of spatiotemporal sequence learning in infancy. *Child Development*, *78*, 1559–1571.

Kochukhova, O., & Gredebäck, G. (2007). Learning about occlusion: Initial assumptions and rapid adjustments. *Cognition*, *105*, 26–46.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis for discrete and continuous outcomes using generalized linear models. *Biometrika*, *84*, 13–22.

McMurray, R., & Aslin, R. N. (2004). Anticipatory eye movements reveal infants' auditory and visual categories. *Infancy*, *6*, 203–229.

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLEARN: A two-process theory of infant face recognition. *Psychological Review*, *98*, 164–181.

Quinn, P. C., Doran, M., Reiss, J. E., & Hoffman, J. E. (2009). Time course of visual attention in infant categorization of cats versus dogs: Evidence for a head bias as revealed through eye tracking. *Child Development*, *80*, 151–161.

Richmond, J., & Nelson, C. A. (2009). Relational memory during infancy: Evidence from eye tracking. *Developmental Science*, *12*, 549–556.

Turati, C., Valenza, E., Leo, I., & Simion, F. (2005). Three-month-olds' visual preference for faces and its underlying visual processing mechanisms. *Journal of Experimental Child Psychology*, *90*, 255–273.