DOI: 10.1111/infa.12564

RESEARCH ARTICLE

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

Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood

Adrian Steffan[1] | **Lucie Zimmer**[1] | **Natalia Arias-Trejo**[2] | **Manuel Bohn**[3,4] | **Rodrigo Dal Ben**[5] | **Marco A. Flores-Coronado**[2] | **Laura Franchin**[6] | **Isa Garbisch**[7] | **Charlotte Grosse Wiesmann**[8] | **J. Kiley Hamlin**[9] | **Naomi Havron**[10] | **Jessica F. Hay**[11] | **Tone K. Hermansen**[12] | **Krisztina V. Jakobsen**[13] | **Steven Kalinke**[3] | **Eon-Suk Ko**[14] | **Louisa Kulke**[15] | **Julien Mayor**[12] | **Marek Meristo**[16] | **David Moreau**[17] | **Seongmin Mun**[14] | **Julia Prein**[3] | **Hannes Rakoczy**[7] | **Katrin Rothmaler**[8] | **Daniela Santos Oliveira**[11] | **Elizabeth A. Simpson**[18] | **Sylvain Sirois**[19] | **Eleanor S. Smith**[20] | **Karin Strid**[16] | **Anna-Lena Tebbe**[8] | **Maleen Thiele**[3] | **Francis Yuen**[9] | **Tobias Schuwerk**[1]

Correspondence
Lucie Zimmer, Ludwig-Maximilians-Universität, Leopoldstr. 13, 80802 München, Germany.
Email: lucie.zimmer@psy.lmu.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: DFG RA 2155/7-1, DFG SCHU 3060/2-1; National Institutes of Health, Grant/Award Numbers: NIH 1R01HD083312, NIH 1R15HD099706; National Research Foundation of Korea, Grant/Award Number: NRF-2021R1I1A2051993; National Science Foundation, Grant/Award Number: NSF CAREER 1653737

Abstract

Measuring eye movements remotely via the participant's webcam promises to be an attractive methodological addition to in-person eye-tracking in the lab. However, there is a lack of systematic research comparing remote web-based eye-tracking with in-lab eye-tracking in young children. We report a multi-lab study that compared these two measures in an anticipatory looking task with toddlers using WebGazer.js and jsPsych. Results of our remotely tested sample of 18-27-month-old toddlers (N = 125) revealed that web-based eye-tracking successfully captured

Adrian Steffan and Lucie Zimmer shared first-authorship.

goal-based action predictions, although the proportion of the goal-directed anticipatory looking was lower compared to the in-lab sample ($N = 70$). As expected, attrition rate was substantially higher in the web-based (42%) than the in-lab sample (10%). Excluding trials based on visual inspection of the match of time-locked gaze coordinates and the participant's webcam video overlayed on the stimuli was an important preprocessing step to reduce noise in the data. We discuss the use of this remote web-based method in comparison with other current methodological innovations. Our study demonstrates that remote web-based eye-tracking can be a useful tool for testing toddlers, facilitating recruitment of larger and more diverse samples; a caveat to consider is the larger drop-out rate.

1 | INTRODUCTION

Eye-tracking technology allows researchers to better understand childrens' interactions with the world. Compared to the manual coding of gaze behaviors, eye-tracking can automatically and accurately track gaze patterns on more complex stimuli with higher spatial and temporal resolution (Oakes, 2012; Wass et al., 2013). Best practices for using in-person eye-tracking with young children have been outlined (Oakes, 2012); however, to date, eye-tracking with children has required in-person testing using a commercial eye-tracking system. In adults, remote automated web-based eye-tracking methods have been established in both computational (Valliappan et al., 2020; Xu et al., 2015) and behavioral research (Bogdan et al., 2023; Schneegans et al., 2021; Semmelmann & Weigelt, 2018; Yang & Krajbich, 2021). So far, to our knowledge, none of these systems have been validated in an interactive paradigm for use with young children (for automated gaze coding of already recorded videos, see, Erel et al., 2022; Werchan et al., 2022; for an overview, see, Kominsky et al., 2021; for in-person vs. remote web-based eye-tracking comparison in a looking time paradigm in infants, see, Bánki et al., 2022). Yet, remote automated web-based eye-tracking has become increasingly important in developmental research due to the growing need for testing children at home. During the Covid-19 pandemic, many labs around the world were unable to conduct in-person studies. Remote web-based studies have thus become more popular in recent years (Kominsky et al., 2021; Leshin et al., 2021; Rhodes et al., 2020; Sheskin et al., 2020; Su & Ceci, 2021), with new tools and techniques for moderated versus unmoderated remote studies emerging in developmental psychology (Lo et al., 2021; Oliver & Pike, 2021; Rhodes et al., 2020; Schidelko et al., 2021; Su & Ceci, 2021).

While some of these projects measure children's looking behavior, they still require manual coding from human observers (e.g., Bacon et al., 2021; Bánki et al., 2022; Nelson & Oakes, 2021; Scott & Schulz, 2017). Manual video-coding is still considered, among many researchers, the gold standard. However, it is labor-intensive making it impractical for studies with a large sample size, and requires comprehensive training (Venker & Kover, 2015). To maintain the reliability of manual coding, it is common for coders to participate in lab-wide reliability checks (Yoder et al., 2018) and to report inter-coder agreement for a subset of videos (Fernald et al., 2008), which - though key to coding

reliability and replicability of results - even further exacerbates the problem of significantly greater number of hours spent on manual annotation than on running machine algorithms. In contrast, automated web-based eye-tracking provides a resource-saving alternative. It is more efficient and has –compared to manual coding of gaze direction from video replays– a relatively high temporal and spatial resolution. As a result, automated coding methods are capable of capturing dependent variables that manual coding cannot (e.g., pupil size, or discrete fixations within an AOI), providing, in conjunction with the technology allowing for at-home testing, exciting new areas of exploration (Ozkan, 2018). Additional advantages of conducting eye-tracking studies remotely compared to traditional one-lab in-person studies are that they (1) make it easier to scale up for large samples; (2) enable researchers to reach a more demographically diverse cohort (e.g., linguistic diversity, racial/ethnic/cultural backgrounds, socio-economic status) as remote web-based studies can be performed from around the world, improving generalizability (Byers-Heinlein et al., 2020; Visser et al., 2022; (3) can potentially reduce costs associated with renting lab space, buying expensive equipment, and other expenses associated with in-person studies; (4) are less time-consuming for participants and more comforting as they can do the testing in their natural environment; (5) offer greater flexibility in terms of scheduling and the ability to collect data from participants in different time zones and (6) have the potential to facilitate international collaborations among research groups, as they are more easily reproducible and less subjective.

Despite these clear advantages, the new remote web-based eye-tracking methods are still undergoing development and involve limitations such as poorer image quality and uncontrolled experimental conditions when compared to their in-lab counterparts (i.e., infant positioning, lighting in the room, and presence of distractors; Wass, 2016; Zaadnoordijk et al., 2021). In a traditional lab, the researcher can ensure that participants are following the instructions of the study, whereas in a remote setting, the researcher may not be able to monitor the participant as closely, and the quality of the setup often varies. Additionally, commercial eye-trackers have a higher sampling rate (one sample per two or four milliseconds) compared to the average webcams available to participants taking about one sample each 30 ms, leaving the data more noisy.

Here, we aimed to test the precision of a web-based eye-tracking system that uses the participant's webcam. Our experiment is based on jsPsych and WebGazer.js (de Leeuw, 2015; Papoutsaki et al., 2016). jsPsych is a javascript framework used to create behavioral experiments that run in a web browser. It was used, in this instance, to control the content the participants interacted with during the experiment but cannot collect eye tracking data in isolation. Thus, it was combined with the WebGazer plugin to produce the present paradigm and data collection set-up. WebGazer captures gaze coordinates by predicting the participant's gaze location on the screen from the head and eyes position recorded via webcam, relative to the displayed stimuli. To evaluate whether this web-based eye-tracking method is comparable to lab-based eye-tracking, we aimed to replicate findings of an in-lab paradigm of the ManyBabies2 project, which revealed spontaneous goal-directed action anticipation measured by anticipatory looking using commercial eye-tracking systems (Schuwerk et al., 2022). The paradigm involves two agents, one who moves through an opaque tunnel and hides from the other in one of two locations and a chaser who also enters the tunnel and seeks the agent who is hiding. A goal of the ManyBabies2 project is to replicate the finding that infants and toddlers visually anticipate an agent's action which is based on a false belief (Southgate et al., 2007). Action prediction, measured by anticipatory looking toward the outcome of that action, is a strong indicator of infant's cognitive reasoning that drives these predictions (Falck-Ytter et al., 2006). In the employed anticipatory looking paradigm, before presenting a false belief-based action, simple goal-directed actions are presented to familiarize toddlers with the set-up. Showing that they anticipate a goal-directed action in these trials, something that toddlers at that age are capable of (e.g., Liszkowski et al., 2007; Luo & Baillargeon, 2007), is an

important validity check of this paradigm. We expected participants to anticipate where the chaser will seek the hiding agent. We compared this anticipatory looking behavior recorded in-lab with anticipatory looking behaviors recorded remotely via webcam in 18- to 27-month-old children.

Following the ManyBabies collaborative framework (Frank et al., 2017; Visser et al., 2021), we conducted a cross-sectional web-based eye-tracking experiment with participants recruited and tested across 16 different labs globally. Labs contributed to recruitment, data collection, data analyses, and other related tasks.

The hypotheses of the present study were the following: First, we expected 18- to 27-month-old children in our web-based eye-tracking sample to engage in goal-based action predictions, indicated by above-chance looking toward the location that matches the outcome of an agent's action goal (i.e., finding the hiding agent). This would replicate Schuwerk et al.'s (2022) results obtained using in-lab commercial eye-tracking systems. Second, we then tested whether the eye-tracking method had an effect on the measured proportional looking score, but had no strong directional hypothesis either way. It could have been that due to the reduced accuracy of remote web-based eye-tracking and increased noise of the at-home test setting, the proportional looking score indicating goal-directed action prediction is smaller in remote web-based than in in-lab eye-tracking. Alternatively, the proportional looking score obtained via remote web-based eye-tracking could have been larger, potentially due to beneficial effects of the familiar environment at home, the increased scheduling flexibility to match children's most attentive times, and the lack of an exhausting trip to a lab. It could also have been that the method would have no effect on the proportional looking score—as these two trends might pull in opposite directions. Third, we expected that the proportion of children who contribute useable data would be lower in the remote web-based setting as compared to in-lab eye-tracking.

A successful replication of in-lab results with our remotely tested sample would render remote automated web-based eye-tracking via the participant's webcam an attractive alternative to in-lab eye-tracking for research on cognitive development. Moreover, our open-source tool would provide the community with a free and powerful method for future research.

2 | METHODS

The study was pre-registered on Open Science Framework (OSF).[1] All materials, data, and the analytic codes are also available on OSF.[2] The software implementing the experiment can be found on GitHub.[3]

2.1 | Participation details

In this multi-lab study, participants were recruited by 16 different labs. For feasibility and data protection reasons, only 11 of these 16 labs were involved in testing. The labs were located in Austria ($n = 1$), Canada ($n = 1$), Germany ($n = 5$), Israel ($n = 1$), Italy ($n = 1$), Mexico ($n = 1$), Norway ($n = 1$), United Kingdom ($n = 1$), United States ($n = 2$), South Korea ($n = 1$), and Sweden ($n = 1$). As participants were recruited and tested by several labs, differing recruitment methods were used (e.g., internal database of laboratories, selected kindergartens, online via social media, birth registries from local registration offices). Participants were compensated according to each individual lab policy (e.g., by gifts, cash). The present study was conducted according to guidelines laid down in the Declaration of

[1]permanent link to pre-registration: https://doi.org/10.17605/OSF.IO/SMYA4.
[2]https://osf.io/p3f67/.
[3]https://github.com/adriansteffan/manywebcams-eyetracking/tree/848504f07fa8c25eb3f28444349a4d60151a7895.

Helsinki, with written informed consent obtained from a parent or guardian for each child before any assessment or data collection. All procedures involving human subjects in this study were approved by the respective Institutional Review Boards (IRBs; for a full list see Supplementary Table 1).

2.1.1 | Time-frame

On September 27th, 2021 we sent an email to the ManyBabies mailing list inviting labs to join the project. Three months later, in January 2022, data collection began and ended in August 2022.

2.1.2 | Lab participation criterion

Participation was open to all labs. However, there were some requirements to participate in data collection or recruitment. Labs needed to: 1) provide ethics approval from their local ethics committee by the start of data collection, 2) be able to actively recruit at least 10 participants and/or be able to test them using either their own WebGazer setup or the one provided by LMU Munich, 3) read the ManyWebcams Manual and comply with the ManyBabies code of conduct (for details see OSF). Note that labs did not have to contribute 10 included participants. Each number of finally useable datasets was included in the overall sample.

2.2 | Participants

The final remotely tested sample consisted of 125 participants (67 girls, 58 boys) aged 18–27 months (548–822 days, M_{age} = 21.83 months, SD_{age} = 2.45 months). All toddlers were born full-term (>37 weeks gestation) and had no reported cognitive, visual, or hearing impairments. Since multiple labs around the world collected data, the participants' places of residence were diverse: Germany (n = 52), Norway (n = 11), Italy (n = 10), United States (n = 10), Sweden (n = 9), United Kingdom (n = 8), Canada (n = 6), Austria (n = 5), Israel (n = 5), South Korea (n = 5), and Mexico (n = 4) (see Figure 1). For most of the participants, at least one parent had an educational degree comparable to a bachelor or higher (n = 105). The parent with the higher educational degree spent on average 17.70 years in education. Among the participants, 26% were raised with a second language (n = 32), and 6% with a third language (n = *7)*. Regarding the number of siblings, 54% of participants had no siblings (n = 68), 35% had one sibling (n = 44), 10% had two siblings (n = 12) and 1% had three siblings (n = 1). The majority of participants were going to daycare (n = 87) and spent there 31 h per week on average. An additional 118 participants were tested but excluded from the analysis. There was no indication of any systematic differences between the included and excluded participants except for residence country (for more details see Figure 1 and Supplementary Table 2 and 3). Participants were excluded for three main reasons (see Supplementary Figure 1): participant-related exclusions (n = 27), technical-related exclusions (n = 52), or exclusions after visual inspection (n = 39). Participant-related exclusions were due to a mismatch between participants' age and our predefined age range (n = 9), prematurity (n = 8), reported cognitive (n = 8) or vision (n = 2) impairments. Technical-related exclusions and exclusions after the visual inspection process are described in more detail in the results section.

The lab-based sample consisted of 70 toddlers (39 girls, 31 boys) aged between 18 and 27 months (552–812 days, M_{age} = 22.92 months, SD_{age} = 2.62 months). This sample was collected in seven labs across the world. Note that for the analyses of the current study we were able to use data from 70 toddlers tested for a pilot study of the ManyBabies2 project (for the original analysis stricter criteria were applied which led to a final sample of 65 included toddlers; for further details, including further

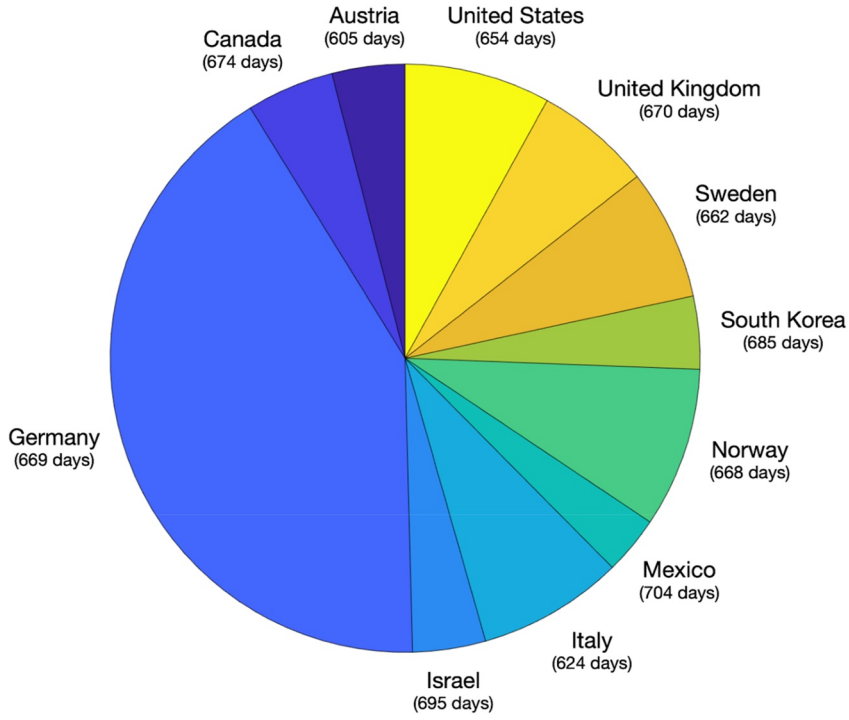

FIGURE 1 Pie chart of the different residence countries of the included participants alongside with their mean age in days in brackets.

information on participating labs, see in Schuwerk et al., 2022). In this pilot study, the appropriateness of the newly developed paradigm was measured. In particular, it was tested if toddlers engage in goal-based action predictions when watching the stimuli.

2.3 | Sample size

Our sample size rationale was based on two effect sizes: Using the same paradigm with in-lab eye-tracking, Schuwerk et al. (2022) observed an effect-size of Cohen's $d = 1.03$ in a sample of 65 toddlers (one sample t test of proportional looking score against chance level). In a pilot study for the current remote web-based version, we tested 40 adults ($M_{age} = 30.10$ years, $SD_{age} = 14.35$ years) and 15 children ($M_{age} = 23.25$ months, $SD_{age} = 10.48$ months). We observed an effect size of Cohen's $d = 0.56$ in a sample of 20 adults who were included in the final analysis, and we did not find a statistically significant effect from the 8 children that were included in the final analysis.

We anticipated two major sources of noise in our data: poorer accuracy of remote web-based eye-tracking as compared to in-lab eye-tracking (Semmelmann & Weigelt, 2018) and more movements artifacts and inattentiveness in toddlers compared to adults (Dalrymple et al., 2018). Based on the observed effect sizes and these considerations, we performed a power analysis for our main hypothesis with the conservative effect size estimate of Cohen's $d = 0.3$. To detect such an effect with a power (1-beta) of 0.95 (using a one sample t test against chance, one-tailed, alpha = 0.05), a minimal sample of 122 toddlers was required. Because in this multi-lab study the exact number of tested

participants could not be determined before the end of data collection, we set $N = 122$ as the minimal sample size of included participants.

2.4 | Materials and design

The experimental design was identical to the familiarization phase of the paradigm previously developed for ManyBabies2.[4]

2.5 | Stimuli

2.5.1 | General scene setup

We used 3D animations representing a chasing scenario between two agents (chaser and chasee; Figure 2). The scene depicted an open, blue-colored room divided into two sections by a horizontal brown picket fence: an upper section, which was about one-third of the height of the room, and a lower section, which was about 2/3 of the height of the room. At the beginning of the scene, two animated agents of the same size were visible in the upper section: a brown bear (chaser) and a yellow mouse (chasee). The agents communicated briefly with pseudo statements. When they moved, one could hear their footsteps. The fence dividing the room was interrupted in the middle by a white inverted Y-shaped tunnel through which the agents could pass from one section to the other. One exit of the tunnel led to the upper section and two identical exits to the lower section of the room, one on the right- and one on the left-hand side. In front of the tunnel exits in the lower section of the room, there were two identical brown boxes with a movable lid, one box in front of each exit.

2.5.2 | Test trials

All participants viewed four trials, with each trial lasting 38s (for a detailed description see Schuwerk et al., 2022). Each trial started with a brief game of tag between two agents, the chaser and the chasee, in which the chasee started either on the left or on the right side. After chasing each other, they stopped, did a high five, and ended up standing side by side in front of the tunnel entrance (left or right position counterbalanced). Both chasee and chaser looked at each other briefly. The chaser continued watching as the chasee headed to the tunnel and entered it. After the chasee disappeared in the tunnel, the chaser moved to the tunnel entrance and remained there until the chasee exited the tunnel (left or right, counterbalanced). During this time, only the sound of footsteps indicated that the chasee was moving through the tunnel. After leaving the tunnel, the chasee turned back, implying eye contact with the chaser, to which the chaser responded by raising their hands, and jumped into the opaque box, which was positioned behind the tunnel exit. The chaser also entered the tunnel and, again, the sound of footsteps indicated their walking through the tunnel (anticipatory period, i.e., 4000 ms). The chaser exited the tunnel on the same side the chasee was hiding. Then, the chaser knocked on the box, the chasee jumped out and, again, the agents did a high five. See the OSF repository for the full animations.

[4]https://manybabies.org/MB2/.

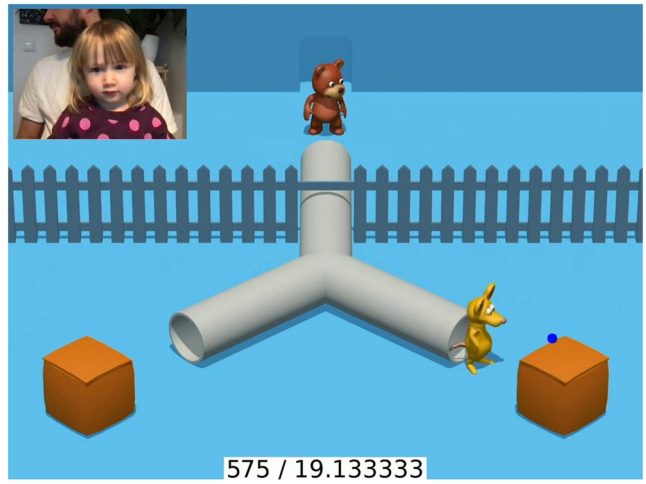

FIGURE 2 A still frame from an overlay of the normalized predictions of gaze location (indicated by the blue dot), the stimuli, and the synchronized webcam video. At the bottom, the current duration (left: frames, right: seconds) is displayed. These overlays were used for the visual inspection process.

2.5.3 | Trial randomization

We used two factors for balancing in the study. First, the location from which the chasee started in the upper section of the room left (L) versus right (R) and second, the box in which the chasee eventually hid (L vs. R). This resulted in four trials: chasee started from the right and ended up in right box (RR); started from the right and ended up in left box (RL); started from the left and ended up in right box (LR); and started from the left and ended up in left box (LL). The order of the four test trials was counterbalanced across participants using two pre-specified pseudo-randomized orders to which they were randomly assigned: LR, LL, RR, RL (Order A); RL, RR, LL, LR (Order B). The trial orders were identical to those used in the in-lab study.

2.6 | Apparatus and procedure

2.6.1 | Testing procedure

Participants met the researcher via a video conference software (e.g., Zoom). Before the test session, the caregiver provided informed written consent via an online survey tool offered by their institution or other third-party software solutions, given the use was covered by their local ethics approval. Subsequently, caregivers completed a demographic questionnaire, which included questions about linguistic and racial/ethnic background, resident country, socio-economic status, caregivers characteristics, and family characteristics. After explaining the general procedure, the researcher offered the caregiver the following instructions. Caregivers were asked to have the child sit in front of a laptop or desktop computer screen with a horizontal screen orientation at a distance of approximately 40 cm. The child could be seated either on their caregiver's lap or in a highchair. Then, the experimenter guided the caregiver to obtain suitable lighting and webcam positioning: If a laptop was used, the caregiver was asked to place it on top of a table and have the child sit in front of it. If a light source (e.g., a window)

caused backlight, the experimenter asked the caregiver to reposition the computer to reach an appropriate angle toward the light source or asked the caregiver to cover it. Caregivers adjusted the angle of the webcam/laptop screen, so that the child's head was centered on the screen, and the caregiver's head was outside of the camera's scope. Alternatively, caregivers were advised to obstruct, close, or move their eyes away from the range of the camera during the experiment, as to not interfere with the eye-tracking procedure. The experimenter then provided the caregiver with a link to access the experimental task and reminded the caregiver to rejoin the video conference after the end of the experiment. Subsequently, the caregiver left the video conference session and accessed the experiment on a browser of their choice (Google Chrome and Firefox were recommended) and started the experiment. During the experiment, the participant's webcam was used to record the child's gaze locations. We also saved the webcam video, which recorded the child's behavior while watching the stimuli. We used a modified version of jsPsych v6.3.1 (de Leeuw, 2015) to control the experimental procedure and stimuli video presentation. To infer the participant's gaze location during the video stimulus presentation, we used WebGazer.js (Papoutsaki et al., 2016). WebGazer is a browser-based eye-tracking library that uses webcam video to infer the participant's gaze locations. It approximates gaze location using a regression model that learns the mapping from pupil positions and eye features to screen coordinates. During the initialization of the eye-tracking procedure, the software also controlled for the distance of the participant in relation to the monitor. To satisfy the headpose requirements enforced by WebGazer, the experiment proceeded only if both eyes were detected within a rectangle (with dimensions equivalent to ⅔ of the webcam feed's height) which was displayed on the screen. Following this requirement, the distance range accepted by the experiment's software spanned 40–130 cm (i.e., 15.7–51.2in). Distances outside of this range caused the program to prompt the participant to move closer or further away from the screen.

At the beginning of the experimental task, a 9-point calibration of the eye-tracking software was displayed, each point appearing for 3 s. During this calibration procedure, a looping animation of a dancing teddy bear was presented as an attention-getter at each calibration point (coordinates in screen percentage [width, height] in order: ([50,50], [50,12], [12,12], [12,50], [12,88], [50,88], [88,88], [88,50], [88,12]) along with an audio cue to attract the participant's attention. This combination of a 9-point calibration procedure and child-friendly attention-getter was used to enhance data accuracy (Zeng et al., 2023). We assessed the quality of the calibration twice: once after the calibration procedure and once after the stimulus display (the second assessment quantified the decrease in eye-tracking quality over time). An attention getter appeared in the middle of the screen for 5 s, and we recorded the average x/y deviations of inferred gaze locations from the center of the screen in pixels during this time. Even though there was no ground truth to compare these values against (making the absolute values difficult to interpret), comparing the average deviations at the two measuring times with each other provides an estimate of the deterioration in eye-tracking quality.

After completion of the experimental task, which lasted approximately 6 min, the experiment software transmitted the data to the experimenter's server for storage and the caregivers returned to the video conference. Caregivers were debriefed on the purpose of the experiment and were given a chance to report any issues faced during the test. The whole experiment lasted approximately 20 min.

2.6.2 | Software setup

The experiment was implemented as a webpage using a modified version of the jsPsych framework v6.3.1 (de Leeuw, 2015). To deliver this page to the participants' machines, we hosted the webpage on an Apache HTTP Server (Version 2.4; Apache Software Foundation, 2012) on a virtual machine running Ubuntu 18.04 LTS (Canonical Ltd, 2018). The participant's browser ran the code controlling

the experiment to present stimuli and record the participant through the webcam. Eye-tracking was performed in real-time on the participant's device. After completing an experiment, the browser sent the data back to the Apache server, where the data was processed and saved using a script written in PHP (Version 8.0; The PHP Group, 2020).

Participating labs had the option of hosting the software on a server of their own using a comparable setup. Alternatively, they could test their participants using the preconfigured server provided by the LMU Munich lab. If they chose to do so, the experiments' software used the ManyKeys library (Steffan & Müller, 2021) to apply end-to-end encryption to the participants' data before transmitting it to the server. This step ensured that only the lab responsible for handling the specific participant's data could access the webcam recordings, enabling different labs to use the same infrastructure for testing while still keeping their participants' data fully private.

2.6.3 | General procedure

We compared the data in the current study to the data collected by Schuwerk et al. (2022). Additionally, data from our pilot study was only used to test our remote web-based eye-tracking paradigm, method feasibility, and sample size rationale, and was not included in the final data analysis.

As WebGazer runs on the participant's device, the achievable sampling rate depends on the participant's hardware capacity. Thus, the sampling rate could not be manipulated but was recorded with our setup for reporting. While we expected a sampling rate of up to 30 Hz for commonly used consumer hardware, our pilot study showed that 15–25 Hz was a more realistic estimate for most devices. Experiments with similar setups reported ranges of 4.50–25.69 Hz (Semmelmann & Weigelt, 2018).

For all videos, we defined two rectangular areas of interest (AOI) around both tunnel exits. We labeled the AOI covering the tunnel exit where the chaser will reappear according to their goal "target AOI" and the other one "distractor AOI". The software tracked whether the child's gaze fell into the left, the right, or neither AOI (Figure 3). According to tests conducted using an adult sample, a gaze

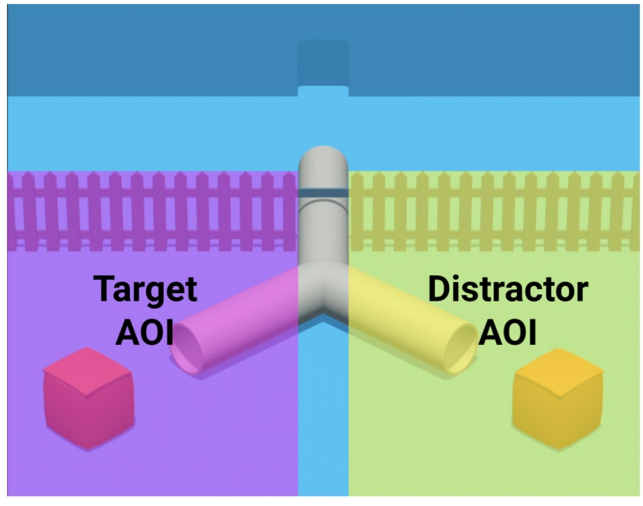

FIGURE 3 Illustration of the scene during the anticipatory period. Colored regions display AOI dimensions we used for our analyses of the web-based eye-tracking data. "Target AOI ″ was the region where the chaser reappeared according to their action goal. "Distractor AOI" was the region covering the other tunnel exit and its surroundings (Dimensions relative to the stimulus video: Left AOI: x: 0%–45%, y: 0%–66%; Right AOI: x: 55%–100%, y: 0%–66%).

point collected with WebGazer has an area of uncertainty of about 100–200 pixels on 1920 × 1080 screens in a practical setting (Papoutsaki et al., 2016). We assumed a similar area of uncertainty for our setup, which is our rationale for choosing AOIs this large (as compared to in-lab data from Schuwerk et al., 2022) for our main analysis. This constituted a necessary trade-off given the technical limitations of our approach. The child's gaze-coordinates, AOI hits, webcam videos, and miscellaneous data (screen size, browser and system information) were submitted to the experimenters' server once the trials concluded.

2.7 | Measures

The experiment consisted of only one trial type in which we manipulated the action sequences of two agents to measure goal-based action predictions via anticipatory looking. We measured the duration of children's gazes toward the target and distractor AOIs between the time the chaser entered the tunnel (first frame the chaser completely disappeared in the tunnel) and the time the chaser exited the tunnel (last frame in which the chaser was entirely inside the tunnel and not yet visible at the tunnel exit). During stimulus playback, the experiment's software sampled gaze predictions as fast as the user's device allowed for, producing the following raw data for every participant/stimulus combination: Per update of the gaze prediction, it included X and Y pixel-coordinates of the estimated gaze location on the screen, which AOIs the gaze fell into (left rectangle, right rectangle, none), and a timestamp specifying how many milliseconds had passed since the stimulus playback started. Using the height and width of the user's browser window, these data were normalized to be relative to the stimulus dimensions. Combining these normalized predictions with the stimulus and webcam video, a replay was created that overlaid the gaze location over the stimulus videos and added the synchronized webcam video in the upper-left corner. These videos were visually inspected to identify trials that had to be excluded (see exclusion criteria below). These trials were omitted from the following pre-processing steps. Participants with a sampling rate below our defined threshold (see Data exclusion) also were excluded. Using information about which AOI is defined as the "target" or "distractor" AOI for a given stimuli version (LR, LL, RR, RL), every captured gaze was classified to fall into one of three categories: "target AOI", "distractor AOI", or "no AOI" (Figure 3). We only included samples with timestamps that fell into the anticipatory period, that is, 4000 ms preceding the frame in which the chaser exited the tunnel. We then calculated for each participant what percentage of gazes fell into each category during this critical time frame aggregated across all trials (for the main hypothesis) and aggregated by trial (for the second hypothesis and the exploratory analyses). This relative percentage was necessary, as sampling rates differed between participants. We computed the proportion of looking toward the target AOI by dividing the number of samples spent looking at the target AOI by the number of samples spent looking at the target plus distractor AOIs (also referred to as total relative looking time; Senju et al., 2009): Proportional looking score = target/(target + distractor).

The score ranged between 0 and 1, whereby a score of 0 meant that the participant had exclusively looked at the distractor, a score of 1 meant that they exclusively looked at the target, and a score of 0.5 meant that they looked for an equally long duration at both AOIs (no preference). By using this proportional score, we were able to compare data across different sampling rates from individual webcams. Further, using this score we could statistically compare the web-based eye-tracking data with in-lab data by Schuwerk et al. (2022), for which we computed the same proportional differential looking score. The resulting data, which now assigned a percentage value to each participant/stimulus/AOI category combination, were used for further statistical analysis. For visualization purposes

(beeswarm plots, available on OSF), the gaze data were also resampled to 15 Hz; however, the resampled data were not used to run statistical analysis.

The collected data points and the processing for the in-lab data by Schuwerk et al. (2022) were comparable, with two differences: First, gaze points were collected using dedicated eye-tracking hardware and resampled to a sampling rate of 40 Hz. Second, the AOIs for the target and distractor were defined to be smaller, as they were not subject to the size increase, we later applied to account for the lower accuracy of webcam-based eye-trackers (see Schuwerk et al., 2022 for more details).

2.8 | Data exclusion

Participants were excluded from analyses if technical problems occurred or if participants did not provide at least one useable trial after the visual inspection. Technical problems included browser freezes that halted the stimulus presentation completely (as reported by the caregiver), crashes due to the hardware being unable to handle real-time eye-tracking, issues with transmitting the data to the experimenters, corrupted data as a result of software failure, and other technical difficulties that can appear in browser-based study setups. As pre-registered, participants providing a sampling rate of 10 Hz or below were also excluded. We chose this cut-off at 1/3rd of the maximum achievable sampling rate of 30 Hz because our pilot data showed that most participants providing sample rates of 10 Hz or lower had very weak hardware, resulting in low refresh rates (around 1–2 Hz). A previous study reported a cut-off at ≤5 Hz (Yang & Krajbich, 2021), but no formal rationale for this cut-off was provided. All webcam video/gaze plot overlays (see Figure 2) were manually checked and individual trials were excluded if: (1) the caregiver interfered with the procedure (e.g., by pointing at stimuli or talking to their toddler), (2) if more than 50% of the gaze data is missing due to inattentiveness of the toddler, and/or (3) the toddler's gaze direction, judged from visual inspection of the webcam video, did not match the recorded gaze coordinates, displayed on the stimulus material as a gaze plot. Reasons for such a mismatch could include: visual properties of the environment (e.g., suboptimal lighting, movements in the background), toddler was looking away, and the gaze coordinates froze at the last location at which the toddler was looking, and/or the toddler attended to the screen, but the gaze coordinates (locations and trajectories) did not match the head and eye movements of the webcam video. Trials were also excluded if a mismatch could not be properly checked due to webcam video and recorded gaze coordinates stemming from two different webcams. WebGazer ensured during initialization that the front-facing webcam was used, but the part of the software responsible for recording the webcam footage for manual checking chose the first available connected webcam, which sometimes resulted in this mismatch in cases when two or more webcams were connected.

A third of all participants were randomly chosen and coded by a second naive rater to obtain interrater reliability (IRR). Cohen's kappa resulted in $\kappa = 0.74$, indicating a substantial inter-rater agreement. Since in our study IRR varies considerably across labs we suggest providing additional guidance for labs demonstrating low IRR in future studies.

2.9 | Statistical analyses

2.9.1 | Confirmatory analysis

All statistical analyses were carried out in R (version 4.1.1, R Core Team, 2021). To test whether participants anticipated goal-directed action outcomes in the web-based method, we measured

above-chance looking toward the location that matched the outcome of the agent's action goal using a one sample *t* test. To test whether the eye-tracking method influenced the measured proportional looking score, we compared web-based eye-tracking data from the current study to lab-based eye-tracking data from the study by Schuwerk et al. (2022) in a generalized linear mixed effects model using the glmmTMB package for R (Brooks et al., 2017). This model was set to predict the proportional looking score based on the fixed effect method (web-based vs. lab-based) and a random effect for labs and participants. We also included trial number (z-transformed) as a control predictor—both as a fixed effect and a random slope within participant. Because proportions are naturally bound to be between 0 and 1, we modeled the data using a beta distribution. The model specification was:

Proportional looking score ~ method + *z*_trial + (1|lab) + (*z*_trial|participant)

A main effect of method would indicate that the way gaze data is sampled in this paradigm has an effect on the proportional looking score, suggesting that this measure of goal-directed anticipatory looking is dependent on the eye-tracking method.

To check whether exclusion rates differed between web-based and in-lab eye-tracking, we computed a Chi-square test on the 2 (web-based vs. in-lab) x 2 (percentage included vs. percentage excluded) contingency table.

2.9.2 | Exploratory analysis

To investigate potential effects of age on the proportional looking score, standardized age and trial (z-scores) were added to the model as fixed effects. Lab was included as a random effect with z_age as a random slope within lab. Participant was included as a random effect with *z*_trial as a slope within participant. The model specification was:

Proportional looking score ~ method + *z*_age + *z*_trial + (*z*_age|lab) + (*z*_trial|participant)

In addition, we analyzed the effect of the recording's sampling rate in the web-based sample on the proportional looking score in an additional model. In this model, we added age, trial and the sampling rate as fixed effects. Lab and participant were included as random effects, with *z*_age and *z*_sampling_rate as random slopes within lab and *z*_trial as a random slope within participant. The model specification was:

Proportional looking score ~ *z*_sampling_rate + *z*_trial + *z*_age + (*z*_age + *z*_sampling_rate|lab) + (*z*_trial|participant)

3 | RESULTS

3.1 | Confirmatory analysis

3.1.1 | Anticipatory looking behavior

In our web-based sample, the relative looking time toward the location that matched the outcome of the agent's action goal (target AOI; $M = 0.62$, $SD = 0.18$; Figure 4) was significantly different from chance level (0.5), $t\ (124) = 7.34$, $p < 0.001$, indicating that the participants anticipated the goal-directed action outcome. In the in-lab sample (Schuwerk et al., 2022), the average proportional looking score was 0.73 ($SD = 0.22$) and participants also showed above-chance looking toward the target AOI, $t\ (69) = 8.80$, $p < 0.001$.

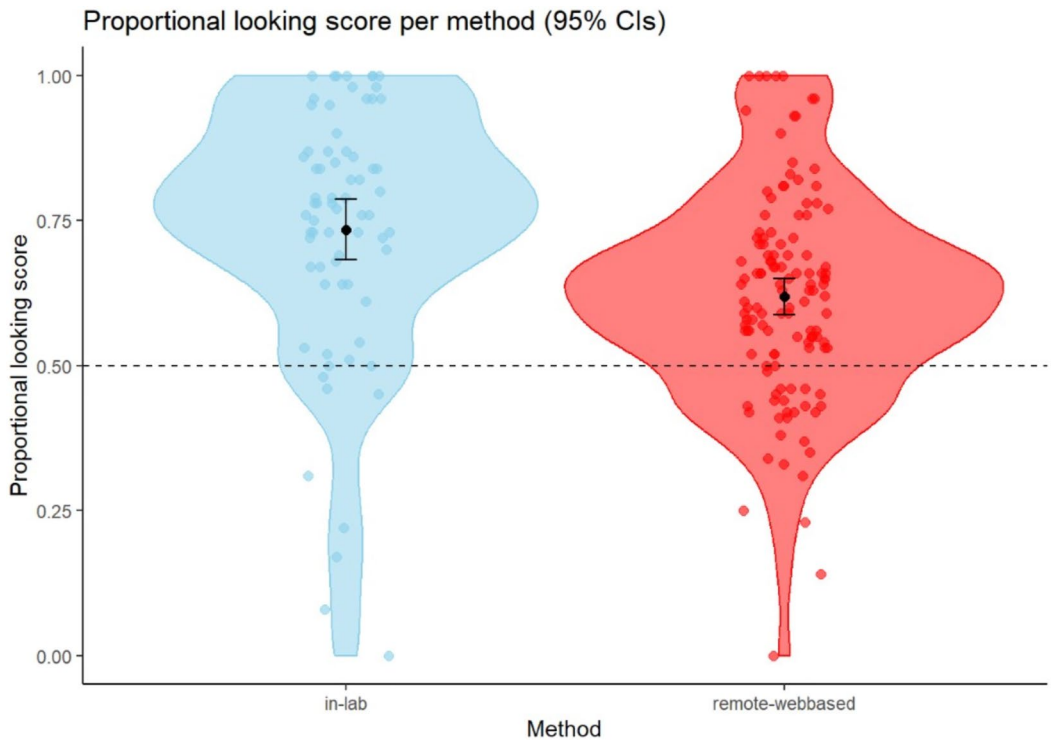

FIGURE 4 Graph depicting the proportional looking score (looking time to target AOI/looking time to target + distractor AOI) (*y* Axis) per method, remote web-based and in-lab eye-tracking (*x* Axis). The error bars represent the 95% confidence intervals (CIs).

In our web-based sample, we observed an effect-size of Cohen's $d = 0.66$ (95% confidence interval: 0.29–1.02) in the one sample directed t test contrasting the proportional looking score against chance level. Schuwerk et al. (2022) observed an effect size of Cohen's $d = 1.03$ (95% confidence interval: 0.50–1.56).

3.1.2 | Comparison of remote web-based versus in-lab eye-tracking in toddlers

To test whether the method had an effect on the proportional looking score, we fit a generalized linear mixed model and found a significant main effect of method ($\beta = 0.52$, $z = 4.46$, $p < 0.001$), reflecting the fact that the proportion of goal-directed anticipatory looking was higher in the in-lab sample (Figure 4).

3.1.3 | Rate of exclusion

In our web-based sample, 125 out of 216 tested participants (58%), that matched our predefined eligibility requirements, were included in the final sample. Thus, 91 participants (42%) were excluded. From these, 52 toddlers (57% of excluded participants) were excluded due to technical reasons. Caregivers had the chance to report any technical issues after completing the experimental task when they returned to the video conference meeting with the experimenter for their debriefing. Techni-

cal problems mainly occurred during the stimulus presentation or during data transmission from the participating families to the experimenters ($n = 35$, 67% of technical errors). Other technical reasons for exclusions were a sampling rate below our predefined threshold ($n = 8$, 15% of technical errors), experimenter error ($n = 2$, 4% of technical errors) or technical error without further information ($n = 7$, 13% of technical errors). As a result of the visual inspection process, a total of 39 toddlers were excluded (43% of excluded participants). They were excluded due to a mismatch between gaze coordinates and their head/eyes movement ($n = 20$, 51% of visual exclusions), interference by caregiver ($n = 4$, 10% of visual exclusions), inattentiveness of the toddler ($n = 4$, 10% of visual exclusions), two different active webcams ($n = 6$, 15% of visual exclusions), suboptimal positioning of the toddler ($n = 1$, 3% of visual exclusions) and error without further information ($n = 4$, 10% of visual exclusions). In contrast, in the in-lab sample, 70 out of 78 tested participants were included, which results in an exclusion rate of 10%. Reasons for exclusion were early termination of the experiment ($n = 6$) and technical problems with data collection ($n = 2$; Schuwerk et al., 2022). We compared web-based and in-lab exclusion rates and found a statistically significant difference, χ^2 (1, $n = 294$) = 24.65, $p < 0.001$. See Figure 5 for a comparison of exclusions for in-lab versus web-based methods.

3.2 | Exploratory analysis

3.2.1 | Change in tracking quality for the web-based sample

We ran calculations for x/y deviations during validation trials for all included participants ($n = 125$). To adjust for different screen resolutions, all values are reported as percentages relative to the screens' width and height. Across all validation trials, we found a mean deviation of 10.51% ($SD = 10.18\%$) for x coordinates and a mean deviation of 11.83% ($SD = 12.59\%$) for y coordinates. We performed a two-tailed t test for paired samples to compare both validation time points. We found no significant difference for either coordinate (X differences: M = 0.196%, $SD = 12.539\%$, t (124) = 0.175, $p = 0.861$, delta = 0.016; Y differences: $M = 1.991\%$, $SD = 16.898\%$, t (124) = 1.317, $p = 0.190$, delta = 0.118). We thus assume that tracking quality did not deteriorate significantly during the trials (for more details

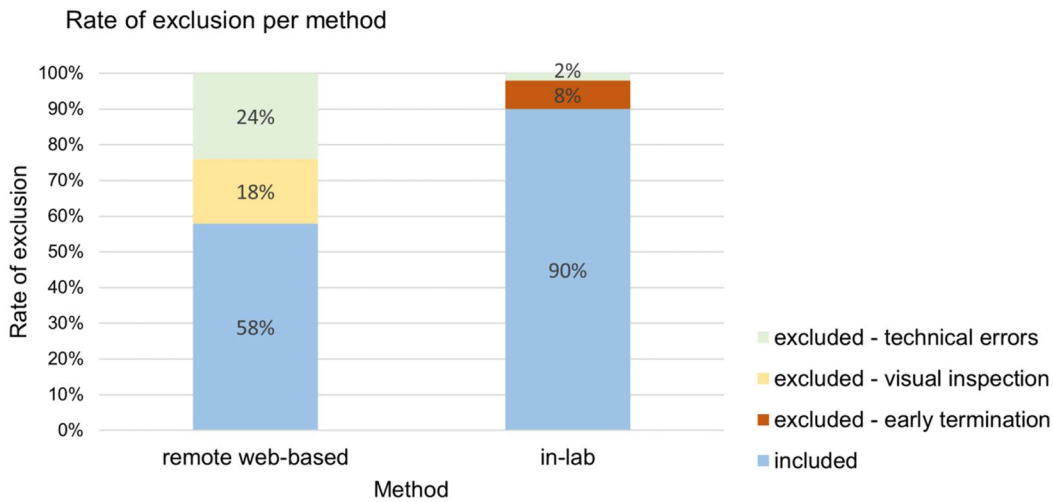

FIGURE 5 The graph depicts the rate of exclusion and reasons for exclusion (y-Axis) per method, remote web-based and in-lab eye-tracking (x-Axis).

about the participants' technical specifications see Supplementary Figure 2 and for visualizations of the first and second validation see Supplementary Figure 3).

3.2.2 | Age analysis

Using the previously described generalized linear mixed effects model, we did not find a statistically significant effect of age on the proportional looking score ($\beta = -0.05$, $z = -0.88$, $p = 0.379$), meaning that in our sample the toddler's age had no influence on anticipatory looking in the web-based task.

3.2.3 | Sampling rate analysis

We observed sampling rates between 10.42 and 40.10 Hz, resulting in a mean sampling rate of 22 Hz ($SD = 7.3$ Hz) in our web-based sample after exclusions. We did not find a statistically significant effect of the sampling rate on the proportional looking score ($\beta = -0.005$, $z = 0.08$, $p = 0.554$), meaning that the sampling rate had no effect on anticipatory looking in our remote sample.

4 | DISCUSSION

In the present study, we validated an open-source, remote web-based eye-tracking method for young children by replicating an anticipatory looking paradigm designed for commercial in-lab eye-trackers (Schuwerk et al., 2022). We measured anticipatory looking behavior via participants' webcams and compared our findings with results of an in-lab study. Although the eye-tracking performance in our remote web-based sample was lower and attrition rate was higher than in the in-lab sample, we successfully replicated in-lab findings, which demonstrates that remote web-based eye-tracking in toddlers is feasible. The fact that we were able to replicate the effect of goal-based anticipatory looking in multiple labs with multiple experimenters, introducing substantial variability in the data collection procedure, strengthens this conclusion. By testing children remotely and collaboratively, we were able to access participants from diverse parts of the world (Asia, Europe, North America and South America) and thus contributed an important first step in reaching more diversity in developmental research, especially in terms of a diverse cultural background.

4.1 | Measuring goal-based action prediction using remote web-based eye-tracking

We found that 18- to 27-month-olds' goal-based action predictions—reflected in above-chance looking toward the location that matches the outcome of an agent's action goal—occurred in our remotely tested sample, replicating results obtained with in-lab commercial eye-tracking systems (Schuwerk et al., 2022). This finding shows that web-based eye tracking can be used successfully to assess children's goal-based action predictions and is in line with previous studies reporting that moderated web-based test sessions with children are comparable to in-lab sessions (Chuey et al., 2021, 2022; Prein et al., 2022; Schidelko et al., 2021). Also, in line with previous remote studies in children, we found no statistically significant age effect (Chuey et al., 2022), suggesting that our web-based eye-tracking method may capture anticipatory looking behavior equally well among 18- to 27-month-olds.

4.2 | Comparing performance of web-based versus in-lab eye-tracking

We found that the eye-tracking method influenced the measured proportional looking score: the in-lab sample's mean proportional looking score toward the target location was higher than the web-based sample's score. This suggests that there may be limitations to remote web-based eye-tracking. Two main limitations of the web-based eye-tracking we used here are lower sampling rate and lower accuracy as compared to when using commercial eye-tracking systems in the lab. In the in-lab data we used for a comparison, the eye-trackers had sampling rates ranging from 60 to 500 Hz. Further, pupil-corneal reflection eye-tracking has a much higher accuracy in measuring x/y-coordinates of gaze points than the regression model WebGazer uses based on webcam videos. Although we took both these limitations into account and adjusted the AOIs in our web-based sample, we unsurprisingly still were not able to track the gaze behavior as fine-grained as in the lab. We assume that lower sampling rate and accuracy in the web-based sample led to noisier data which drove the proportional looking score toward chance-level. Replicating main findings from the lab given such added noise is thus a robust demonstration of the method.

4.3 | Comparing data quality of web-based and in-lab eye-tracking

We found support for our hypothesis that the proportion of children who contributed useable data was lower in web-based as compared to in-lab eye-tracking; this is likely largely due to poorer data quality and/or technical challenges with the remote web-based approach. Because the participating families were responsible for allowing data transmission to our servers, the dropout due to transmitting failures were particularly high. For instance, if the caregiver accidently closed the experiment's browser window after completing the last trial but before the process of data transmission was finished, the data transmission to our servers stopped. In future studies, this source of data loss could be minimized by explicitly instructing participants to keep the browser open for a longer time during the instruction, as well as using clearly visible warning displays during the data transfer. Programming a regular backup of the data during the experiment can be useful if this does not interrupt the experimental flow. Our high attrition rate in the web-based sample is in line with results of previous web-based eye-tracking studies with infants using a commercial eye-tracking platform (52% in Bánki et al., 2022), but also with adults using automated gaze coding (62% in Yang & Krajbich, 2021; 66% in Semmelmann & Weigelt, 2018). Interestingly, attrition rates in child and adult samples seem to converge when testing remotely, despite the fact that higher attrition rates are usually observed in young children compared to adults in in-lab studies using commercial eye-tracking systems (Holmqvist et al., 2023).

4.4 | Limitations

While this study examined the replicability of an in-lab paradigm, we did not explicitly measure the accuracy of WebGazer for toddlers. Using an in-lab eye-tracker concurrently while running a WebGazer experiment could provide us with a proper benchmark to compare against the inferred gaze coordinates. These data points would allow us to create accuracy measures that are directly comparable to the measures reported by Papoutsaki et al. (2016), thus providing a better idea of how the noise levels differ between infant and adult data for webcam eye-tracking.

Even though WebGazer estimates x/y gaze coordinates, the reliability of these measurements is greatly reduced by the noisy nature of the prediction. Therefore, WebGazer could suffer in designs

with a larger number of AOIs, limiting the kind of studies it can be deployed in. As our present design only included two AOIs, we cannot make claims about the performance of WebGazer in these more complex scenarios. Further experiments with a larger number of AOIs need to be conducted to make definitive statements about the general usefulness of WebGazer.

To make the data of this study comparable to the in-lab sample, we used the same 4:3 aspect ratio for the stimulus material. As most computer screens today have a widescreen aspect ratio of 16:9, the stimulus material did not fill the screen's full width but left borders on both sides of the video. We replicated the findings of Schuwerk et al. (2022) under these conditions. Still, paradigms that use the full width of the screen (33% increase in presentation space) should be even less affected by the accuracy drop from using WebGazer as opposed to in-lab eye-tracking, as the horizontal eye movements would occur more clearly when a larger area is used to display the stimuli.

The range of head-to-screen distances accepted by WegGazer was large, spanning 40–130 cm (i.e., 15.7–51.2in). Calibration procedures like the one employed by WebGazer help to normalize gaze location estimations across different distances, but it is possible that children sitting particularly close to or far away from the screen could have exhibited worse tracking performance. As the head-to-screen distance was not recorded, the current study cannot determine if the highly variable distances affected the tracking quality. Additionally, this variability in distances also prevented us from calculating visual angles for our validation, which limits comparisons to other eye-tracking tools.

We decided to place two calibration quality checks in the experiment - one directly after calibration and one after the stimulus presentation. The number and placement were chosen to not overly disturb the stimulus presentation and to stay as close to the in-lab experiment as possible. However, these checks on tracking quality only compare the quality at two discrete time points. If, say, tracking quality worsened during stimulus presentation but improved to normal levels toward the end, our validation data would not report a change, even though the resulting data may be affected negatively.

Remote testing comes with an inherently higher exclusion rate than in-lab data as additional sources of errors are introduced. While software improvements could aid in lowering the attrition rate, there are many variables to control for when testing on participants' devices, such as available hardware, software characteristics like operating system, Internet connection strength, or available webcams. In addition, we recommend systematically collecting parents' feedback on technical difficulties they experienced to gain more information about potential reasons for data loss. Thus, at this point, remote testing is unlikely to reach levels comparable to in-lab studies.

Our remote sample was more diverse and global than samples from most in-person developmental studies (Singh et al., 2021), but it was still primarily a WEIRD sample (Western, Educated, Industrialized, Rich, Democratic). Thus, it is far from representing a multifaceted set of different linguistic, cultural, ethnic or socio-economic backgrounds. For example, the fact that possessing or having access to a computer is a precondition to participation already excludes large parts of the world's population. Participants outside of Western Europe or North America comprised only 11.2% of our final data, and interestingly, exclusion rates were higher in that sample (53.33% vs. 21.62%). Were we to obtain a more geographically diverse sample, we could have tested the effect of background culture on different aspects of our analysis. For example, cultural context can affect children's visual perception of scenes (e.g., Nisbett & Miyamoto, 2005). Additionally, children from different cultures might differ in their performance on different tasks. For example, Callaghan et al. (2011) found that tasks that involved pretense or graphic symbols showed cultural differences. Canadian children developed such skills sooner than Indian and Peruvian children. In ManyBabies4, which tests infants' social evaluation development, there is an attempt to tackle such cross-cultural comparisons as a spin-off project which examines cultural values and behaviors and their relation to children's social evaluation development (Wang et al., 2023).

The method used here has potential to enable research outside privileged research environments: first, by providing researchers with a low-cost eye-tracking solution, and second, by the possibility to reach participants in their homes, leveraging burdens to participate such as geographical distance to the lab or lack of time or resources to get there.

Another factor that could have influenced our remote sample is the Covid-19 pandemic. The testing took place over a period (January-August 2022) of stay-at-home restrictions. The toddlers' experience with electronic devices, the time spent with these devices, and their screen time were most likely increased during that period, and higher than the in-lab sample. This difference between the samples could have partially affected the results. However, it is important to keep in mind that at least in the last two decades, and probably before, a rapid increase in the number of electronic toys, and toys linked to electronic media, marketed for infants and toddlers (e.g., Levin & Rosenquest, 2001), and the development of the interactive mobile media technology in general (e.g., Courage et al., 2021), has led to a remarkable and widespread increase of toddlers' experience with electronic devices and their time spent with these devices, even before the Covid-19 pandemic.

4.5 | Current method in the larger context of recently emerging technical approaches

Recently, online experiment platforms such as Lookit (Scott & Schulz, 2017) have enabled remote testing of infants and toddlers using webcam video. While these platforms make it easier for labs to collect data online, they currently require manual coding of video frames to derive dependent variables. This data coding method is time-consuming when dealing with large datasets and introduces objectivity issues, so employing automated methods is desirable. However, we implemented visual inspection only as a preprocessing step to efficiently reduce noise in the data. Unlike manual coding, it involves a holistic assessment (e.g., a general impression whether the look is completely off or not by reviewing the time-locked gaze coordinates and the participant's webcam video overlayed on the stimuli), making the process more efficient, which typically does not take longer than the duration of the trial itself. As a next step, implementing an automated pre-selection process to identify trials that potentially need to be excluded would be ideal to minimize the number of videos that require visual inspection. Also, the IRR should be more consistent across labs in future studies that also implement our visual inspection process. In our study, we observed an IRR ranging from 0.5 to 1, which demonstrates a high variability across labs and results in a substantial agreement. Given that there were some labs with very high IRR, we assume that the low IRR stems from simple misunderstandings. Thus, we suggest a thorough review of the coding process with labs demonstrating low IRR.

Currently, there are several commercial online webcam-based eye-tracking platforms (e.g., Finger et al., 2017; GazeRecorder, 2010; Lewandowska, 2019). Bánki et al. (2022) used LabVanced (Finger et al., 2017) for remote eye-tracking studies with infants, but in general, these platforms have yet to be widely validated for infant research. While these platforms allow for researchers to quickly set up experiments with little programming knowledge, free, open-source approaches such as WebGazer provide considerable advantages of their own. First, the transparency of open-source code is desirable in a research context, as it allows other researchers to verify the validity of the analysis and promotes openness and accessibility, which can help democratize the scientific process and make research more inclusive. Furthermore, due to the code being available and modifiable, scientists can change the software to fit specific research needs, like making the calibration procedure more infant-friendly. This can save time and resources, as researchers can build on existing code and incorporate it into their own work, rather than starting from scratch. Lastly, the low cost of the method enables labs with fewer resources to use eye-tracking, an important factor for promoting research outside of privileged research infrastructures.

The potential of webcam-based eye-tracking is further amplified through recent applications in both educational and clinical settings (Hutt & D'Mello, 2022; Wong et al., 2023).

4.5.1 | Post hoc gaze inference

WebGazer performs real-time gaze location prediction on the participant's device, which has at least two downsides. The achievable sampling rate depends on the participant's hardware capacity and thus varies among participants. Also, real-time gaze inference requires frequent updates, limiting the complexity of the predictive models. Using more sophisticated methods or computationally expensive deep learning models to capture the face's geometry, locate the pupil, and infer gaze locations is not currently feasible in a real-time setting (Erel et al., 2022; Valliappan et al., 2020).

An alternative approach is to capture webcam footage online but run the calculations to determine gaze locations after the experiment concluded. Doing so would lift the restrictions on inference speed, and the computation of gaze location would not need to be performed on the participants' hardware.

Werchan et al. (2022) recently presented OWLET, an infant-focused webcam eye-tracking system that follows this approach, performing gaze data processing post hoc. OWLET may outperform WebGazer on some dimensions. For instance, the best-performing inference models of WebGazer achieve an average error of 4.17° in an adult sample with a controlled calibration (Papoutsaki et al., 2016). OWLET reported mean absolute x/y calibration deviations of 3.36°/2.67° across infants with a simpler, infant-friendly calibration.

While our study validated WebGazer exclusively on PCs, OWLET can also infer gaze location from video captured on tablet computers and mobile devices. In a study testing the robustness of OWLET, the authors found higher socioeconomic and racial/ethnic diversity in their sample using mobile devices compared to laptops (Werchan et al., 2022). The ability to run eye-tracking studies on these devices would, therefore, be desirable for projects aiming to diversify samples, such as the ones under the ManyBabies framework (Frank et al., 2017; Visser et al., 2021).

On the other hand, our setup is more flexible and easier to use than the OWLET. Whereas WebGazer can be configured to allow any calibration scheme and webcam format, OWLET by default only allows a fixed four-point calibration and only processes 16:9 webcam videos with a framerate of 30 frames per second or higher. Moreover, WebGazer can be plugged into any online experiment set up with jsPsych to produce inferred gaze coordinates without additional post hoc processing through dedicated software. This advantage is important for big team science collaborations like ManyBabies, for example, by reducing the need for additional software installations for all participating labs. Furthermore, given that WebGazer provides real-time tracking, and assuming enough computational power, only WebGazer could be adapted to create infant-controlled experiments.

In sum, when choosing a web-based eye-tracking solution, researchers must consider these tradeoffs based on their resources and paradigm. With further work on streamlining the process, a system can be built that utilizes the improved accuracy of OWLET with the convenience and flexibility that WebGazer provides.

4.5.2 | Deep learning

While WebGazer and OWLET use traditional computer vision algorithms to extract facial information and map them to screen coordinates based on regressions and polynomial functions respectively, applying end-to-end deep learning algorithms trained on large datasets shows great potential for

webcam eye-tracking. Valliappan et al. (2020) used deep-learning models to achieve gaze-tracking accuracy for adults comparable to specialized eye-tracking software using only a smartphone's front camera. Unfortunately, the software they developed is not openly available and needs to be reimplemented to be used in experiments. Furthermore, their training data exclusively consisted of adults, so the generalizability to infant footage remains unknown. Nonetheless, their results show the potential of webcam-based eye-tracking through deep learning algorithms.

iCatcher+ also uses deep learning algorithms to classify gaze points into either left, right, or away (Erel et al., 2022). The model was trained on a hand-labeled dataset of infant webcam footage. iCatcher+ reaches gaze coding accuracy comparable to that of human coders, making it a viable choice for paradigms with binary dependent variables. Until deep learning solutions for x/y coordinate inference from webcam footage are created, online studies that require more fine-grained paradigms have to rely on tools like OWLET or WebGazer.

5 | CONCLUSION

Web-based eye-tracking can be used to capture toddlers' goal-based action anticipation. Thus, in-lab findings can be replicated using remote webcam-based testing, which provides children and their caregivers with a more comfortable participation experience in their natural environment. In developmental research, eye-tracking is commonly performed using in-lab pupil-corneal reflection eye-tracking. While this specialized hardware enables high gaze tracking accuracy that software-only solutions cannot match, it comes with substantially higher costs and physical and/or social boundaries that are hard to overcome. Collecting eye-tracking data remotely using common computers and WebGazer substantially reduces the cost of running experiments, makes testing young participants less time-consuming and more flexible, while providing the opportunity to test demographically diverse, large international samples under comparable conditions. For experiments in which the benefits of remote testing are substantial, such as with children, and a reduced spatial resolution can be tolerated, web-based webcam eye-tracking using WebGazer is a promising method.

AFFILIATIONS

[1]Department of Psychology, Ludwig-Maximilians-Universität München, München, Germany

[2]Facultad de Psicología, Universidad Nacional Autónoma de México, Ciudad de México, México

[3]Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[4]Institute of Psychology, Leuphana University Lüneburg, Lüneburg, Germany

[5]Faculty of Arts & Science, Ambrose University, Calgary, Alberta, Canada

[6]Department of Psychology and Cognitive Science, University of Trento, Trento, Italy

[7]Department of Developmental Psychology, University of Göttingen, Göttingen, Germany

[8]Research Group Milestones of Early Cognitive Development, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[9]Department of Psychology, The University of British Columbia, Vancouver, British Columbia, Canada

[10]School of Psychological Sciences & Center for the Study of Child Development, University of Haifa, Haifa, Israel

[11]Department of Psychology, University of Tennessee, Knoxville, Tennessee, USA

[12]Department of Psychology, University of Oslo, Oslo, Norway

[13]Department of Psychology, James Madison University, Harrisonburg, Virginia, USA

[14]Department of English Language and Literature, Chosun University, Gwangju, Korea

[15]Developmental Psychology with Educational Psychology, University of Bremen, Bremen, Germany

[16]Department of Psychology, University of Gothenburg, Göteborg, Sweden

[17]School of Psychology and Centre for Brain Research, University of Auckland, Auckland, New Zealand[18]Department of Psychology, University of Miami, Coral Gables, Florida, USA

[19]Department of Psychology, Université du Québec à Trois-Rivières, Trois-Rivières, Québec, Canada

[20]Department of Psychology, University of Cambridge, Cambridge, UK

ACKNOWLEDGMENTS

Jessica F. Hay was supported by NIH 1R01HD083312 and NIH 1R15HD099706; Eon-Suk Ko was supported by NRF-2021R1I1A2051993; Hannes Rakoczy was supported by DFG RA 2155/7-1; Tobias Schuwerk was supported by DFG SCHU 3060/2-1; Elizabeth A. Simpson was supported by NSF CAREER 1653737.

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest with regard to the funding source for this study.

ORCID

Lucie Zimmer https://orcid.org/0000-0002-5766-2991
Naomi Havron https://orcid.org/0000-0001-6429-1546
Eon-Suk Ko https://orcid.org/0000-0003-3963-4492
Julien Mayor https://orcid.org/0000-0001-9827-5421
Marek Meristo https://orcid.org/0000-0001-6792-3123
Elizabeth A. Simpson https://orcid.org/0000-0003-2715-2533
Maleen Thiele https://orcid.org/0000-0002-1695-1850

REFERENCES

Apache Software Foundation. (2012). Apache HTTP server. Version 2.4) (Computer Software). https://apache.org/

Bacon, D., Weaver, H., & Saffran, J. (2021). A framework for online experimenter-moderated looking-time studies assessing infants' linguistic knowledge. *Frontiers in Psychology*, *12*. Article 703839. https://doi.org/10.3389/fpsyg.2021.703839

Bánki, A., de Eccher, M., Falschlehner, L., Hoehl, S., & Markova, G. (2022). Comparing online webcam-and laboratory-based eye-tracking for the assessment of infants' audio-visual synchrony perception. *Frontiers in Psychology*, *12*. Article 733933. https://doi.org/10.3389/fpsyg.2021.733933

Bogdan, P. C., Dolcos, S., Buetti, S., Lleras, A., & Dolcos, F. (2023). Investigating the suitability of online eye tracking for psychological research: Evidence from comparisons with in-person data using emotion–attention interaction tasks. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02143-z

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated Generalized Linear Mixed Modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., Kominsky, J. F., Kosie, J. E., Lew-Williams, C., Liu, L., Mastroberardino, M., Singh, L., Waddell, C. P. G., Zettersten, M., & Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 349–363. https://doi.org/10.1037/cap0000216

Callaghan, T., Moll, H., Rakoczy, H., Warneken, F., Liszkowski, U., Behne, T., & Tomasello, M. (2011). Early social cognition in three cultural contexts: III. Individual studies. *Monographs of the Society for Research in Child Development*, *76*(2), 34–104. https://doi.org/10.1111/j.1540-5834.2011.00606.x

Canonical Ltd. (2018). Ubuntu. Version 18.04 LTS) (Computer Software).

Chuey, A., Asaba, M., Bridgers, S., Carrillo, B., Dietz, G., Garcia, T., Leonard, J. A., Liu, S., Merrick, M., Radwan, S., Stegall, J., Velez, N., Woo, B., Wu, Y., Zhou, X. J., Frank, M. C., & Gweon, H. (2021). Moderated online data-collection for developmental research: Methods and replications. *Frontiers in Psychology*, *12*. Article 734398. https://doi.org/10.3389/fpsyg.2021.734398

Chuey, A., Boyce, V., Cao, A., & Frank, M. C. (2022). Conducting developmental research online vs. in-person: A meta-analysis. *PsyArXiv*. https://doi.org/10.31234/osf.io/qc6fw

Courage, M. L., Frizzell, L. M., Walsh, C. S., & Smith, M. (2021). Toddlers using tablets: They engage, play, and learn. *Frontiers in Psychology*, *12*. Article 564479. https://doi.org/10.3389/fpsyg.2021.564479

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

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Erel, Y., Shannon, K. A., Chu, J., Scott, K. M., Kline Struhl, M., Cao, P., Tan, X., Hart, P., Raz, G., Piccolo, S., Mei, C., Potter, C., Jaffe-Dax, S., Lew-Williams, C., Tenenbaum, J., Fairchild, K., Bermano, A., & Liu, S. (2022). iCatcher+: Robust and automated annotation of infant's and young children's gaze direction from videos collected in laboratory, field, and online studies. *PsyArXiv*. https://doi.org/10.31234/osf.io/up97k

Falck-Ytter, T., Gredebäck, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience*, *9*(7), 878–879. https://doi.org/10.1038/nn1729

Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In I. A. Sekerina, E. Fernandez, & H. Clahsen (Eds.), *Developmental Psycholinguistics: On-line methods in children's language processing* (pp. 97–135). John Benjamins.

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). *LabVanced: A unified JavaScript framework for online studies [conference paper]*. International Conference on Computational Social Science IC2S2S. (Germany).

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. https://doi.org/10.1111/infa.12182

GazeRecorder (2010). GazeRecorder (Computer Software) https://gazerecorder.com/

Holmqvist, K., Örbom, S. L., Hooge, I. T. C., Niehorster, D. C., Alexander, R. G., Andersson, R., Benjamins, J. S., Blignaut, P., Brouwer, A.-M., Chuang, L. L., Dalrymple, K. A., Drieghe, D., Dunn, M. J., Ettinger, U., Fiedler, S., Foulsham, T., van der Geest, J. N., Hansen, D. W., Hutton, S. B., …, & Hessels, R. S. (2023). Eye tracking: Empirical foundations for a minimal reporting guideline. *Behavior Research Methods*, *55*(1), 364–416. https://doi.org/10.3758/s13428-021-01762-8

Hutt, S., & D'Mello, S. K. (2022). Evaluating calibration-free webcam-based eye tracking for gaze-based user modeling. In *Proceedings of the 2022 international conference on multimodal interaction* (pp. 224–235).

Kominsky, J. F., Begus, K., Bass, I., Colantonio, J., Leonard, J. A., Mackey, A. P., & Bonawitz, E. (2021). Organizing the methodological toolbox: Lessons learned from implementing developmental methods online. *Frontiers in Psychology*, *12*. Article 702710. https://doi.org/10.3389/fpsyg.2021.702710

Leshin, R., Leslie, S.-J., & Rhodes, M. (2021). Does it matter how we speak about social kinds? A large, pre-registered, online experimental study of how language shapes the development of essentialist beliefs. *Child Development*, *92*(4), e531–e547. https://doi.org/10.1111/cdev.13527

Levin, D. E., & Rosenquest, B. (2001). The increasing role of electronic toys in the lives of infants and toddlers: Should we be concerned? *Contemporary Issues in Early Childhood*, *2*(2), 242–247. https://doi.org/10.2304/ciec.2001.2.2.9

Lewandowska, B. (2019). RealEye eye-tracking system technology whitepaper. Retrieved https://support.realeye.io/realeye-accuracy/ 19 December 2022.

Liszkowski, U., Carpenter, M., & Tomasello, M. (2007). Pointing out new news, old news, and absent referents at 12 months of age. *Developmental Science*, *10*(2), F1–F7. https://doi.org/10.1111/j.1467-7687.2006.00552.x

Lo, C. H., Mani, N., Kartushina, N., Mayor, J., & Hermes, J. (2021). e-Babylab: an open-source browser-based tool for unmoderated online developmental studies. *PsyArXiv*. https://doi.org/10.31234/osf.io/u73sy

Luo, Y., & Baillargeon, R. (2007). Do 12.5-month-old infants consider what objects others can see when interpreting their actions? *Cognition*, *105*(3), 489–512. https://doi.org/10.1016/j.cognition.2006.10.007

Nelson, C. M., & Oakes, L. M. (2021). "May I grab your attention?": An investigation into infants' visual preferences for handled objects using Lookit as an online platform for data collection. *Frontiers in Psychology*, *12*. Article 3866. https://doi.org/10.3389/fpsyg.2021.733218

Nisbett, R. E., & Miyamoto, Y. (2005). The influence of culture: Holistic versus analytic perception. *Trends in Cognitive Sciences*, *9*(10), 467–473. https://doi.org/10.1016/j.tics.2005.08.004

Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy*, *17*(1), 1–8. https://doi.org/10.1111/j.1532-7078.2011.00101.x

Oliver, B. R., & Pike, A. (2021). Introducing a novel online observation of parenting behavior: Reliability and validation. *Parenting*, *21*(2), 168–183. https://doi.org/10.1080/15295192.2019.1694838

Ozkan, A. (2018). Using eye-tracking methods in infant memory research. *The Journal of Neurobehavioral Sciences*, *5*, 62–66.

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the 25th international joint conference on artificial intelligence* (pp. 3839–3845). IJCAI).

Prein, J. C., Bohn, M., Kalinke, S., & Haun, D. B. M. (2022). Tango: A reliable, open-source, browser-based task to assess individual differences in gaze understanding in 3 to 5-year-old children and adults. *PsyArXiv*. https://doi.org/10.31234/osf.io/vghw8

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rhodes, M., Rizzo, M. T., Foster-Hanson, E., Moty, K., Leshin, R. A., Wang, M., Benitez, J., & Ocampo, J. D. (2020). Advancing developmental science via unmoderated remote research with children. *Journal of Cognition and Development*, *21*(4), 477–493. https://doi.org/10.1080/15248372.2020.1797751

Schidelko, L. P., Schünemann, B., Rakoczy, H., & Proft, M. (2021). Online testing yields the same results as lab testing: A validation study with the false belief task. *Frontiers in Psychology*, *12*. Article 703238. https://doi.org/10.3389/fpsyg.2021.703238

Schneegans, T., Bachman, M. D., Huettel, S. A., & Heekeren, H. (2021). Exploring the potential of online webcam-based eye tracking in decision-making research and influence factors on data quality. *PsyArXiv*. https://doi.org/10.31234/osf.io/zm3us

Schuwerk, T., Kampis, D., Baillargeon, R., Biro, S., Bohn, M., Byers-Heinlein, K., Dörrenberg, S., Fisher, C., Franchin, L., Fulcher, T., Garbisch, I., Geraci, A., Grosse Wiesmann, C., Hamlin, K., Haun, D. B. M., Hepach, R., Hunnius, S., Hyde, D. C., Karman, P., …, & rakoczy, h. (2022). Action anticipation based on an agent's epistemic state in toddlers and adults. *PsyArXiv*. https://doi.org/10.31234/osf.io/x4jbm

Scott, K., & Schulz, L. (2017). Lookit (Part 1): A new online platform for developmental research. *Open Mind*, *1*(1), 4–14. https://doi.org/10.1162/OPMI_a_00002

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: An absence of spontaneous theory of mind in Asperger syndrome. *Science*, *325*(5942), 883–885. https://doi.org/10.1126/science.1176170

Sheskin, M., Scott, K., Mills, C. M., Bergelson, E., Bonawitz, E., Spelke, E. S., Fei-Fei, L., Keil, F. C., Gweon, H., Tenenbaum, J. B., Jara-Ettinger, J., Adolph, K. E., Rhodes, M., Frank, M. C., Mehr, S. A., & Schulz, L. (2020). Online developmental science to foster innovation, access, and impact. *Trends in Cognitive Sciences*, *24*(9), 675–678. https://doi.org/10.1016/j.tics.2020.06.004

Singh, L., Cristia, A., Karasik, L. B., Rajendra, S. J., & Oakes, L. (2021). Diversity and Representation in Infant Research: Barriers and bridges towards a globalized science of infant development. *PsyArXiv*. https://doi.org/10.31234/osf.io/hgukc

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. *Psychological Science*, *18*(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Steffan, A., & Müller, T. (2021). ManyKeys. Version 1.0. Computer Software) https://github.com/adriansteffan/manykeys/tree/bed46cdaf3cb8a578c6277eff669b0abb36c3a26

Su, I. A., & Ceci, S. (2021). Zoom Developmentalists": *Home-based videoconferencing developmental research during COVID-19*. *PsyArXiv*. https://doi.org/10.31234/osf.io/nvdy6

The PHP Group. (2020). PHP. Version 8.0) (Computer Software).

Valliappan, N., Dai, N., Steinberg, E., He, J., Rogers, K., Ramachandran, V., Xu, P., Shojaeizadeh, M., Guo, L., Kohlhoff, K., & Navalpakkam, V. (2020). Accelerating eye movement research via accurate and affordable smartphone eye tracking. *Nature Communications*, *11*(1), 4553. Article 4553. https://doi.org/10.1038/s41467-020-18360-5

Venker, C. E., & Kover, S. T. (2015). An open conversation on using eye-gaze methods in studies of neurodevelopmental disorders. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1719–1732. https://doi.org/10.1044/2015_JSLHR-L-14-0304

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., Franchin, L., Frank, M. C., Geraci, A., Hamlin, J. K., Kaldy, Z., Kulke, L., Laverty, C., Lew-Williams, C., Mateu, V., Mayor, J., Moreau, D., Nomikou, I., Schuwerk, T., … Zettersten, M. (2022). Improving the generalizability of infant psychological research: The ManyBabies model. *Behavioral and Brain Sciences*, *45*, e35. Article e35. https://doi.org/10.1017/S0140525X21000455

Wang, Y., Alpcetin, B., Zhu, J., Buyurucu, G., Sancar, B. H., Kaya, M. E., Dresel, M., Exner, A., Hamlin, J. K., Havron, N., Henderson, A., Martin, A., Partridge, T. T., Schuwerk, T., Shainy, M. R., Su, Y., Tsang, C. K. A., Uzefovsky, F., Wong, T. T.-Y., … Lucca, K. (2023). *Individual differences in infants' social evaluations across cultures: A spin-off project of many babies*. CEO Conference on Cognitive Development. [Poster Presentation]. The 13th Budapest https://osf.io/jp532

Wass, S. V. (2016). The use of eye-tracking with infants and children. In J. Prior & J. Van Herwegen (Eds.), *Practical research with children* (1st ed., pp. 24–45). Routledge. https://doi.org/10.4324/9781315676067

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

Werchan, D. M., Thomason, M. E., & Brito, N. H. (2022). Owlet: An automated, open-source method for infant gaze tracking using smartphone and webcam recordings. *Behavior Research Methods*, *55*(6), 3149–3163. https://doi.org/10.3758/s13428-022-01962-w

Wong, A. Y., Bryck, R. L., Baker, R. S., Hutt, S., & Mills, C. (2023). Using a webcam based eye-tracker to understand students' thought patterns and reading behaviors in neurodivergent classrooms. In *LAK23: 13th international learning analytics and knowledge conference* (pp. 453–463).

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., & Xiao, J. (2015). TurkerGaze: Crowdsourcing saliency with webcam based eye tracking. arXiv http://arxiv.org/abs/1504.06755

Yang, X., & Krajbich, I. (2021). Webcam-based online eye-tracking for behavioral research. *Judgment and Decision Making*, *16*(6), 1485–1505. https://doi.org/10.1017/S1930297500008512

Yoder, P. J., Lloyd, B. P., & Symons, F. J. (2018). *Observational measurement of behavior* (2nd ed.). Paul H. Brookes.

Zaadnoordijk, L., Buckler, H., Cusack, R., Tsuji, S., & Bergmann, C. (2021). A global perspective on testing infants online: Introducing ManyBabies-AtHome. *Frontiers in Psychology*, *12*. Article 703234. https://doi.org/10.3389/fpsyg.2021.703234

Zeng, G., Simpson, E. A., & Paukner, A. (2023). Maximizing valid eye-tracking data in human and macaque infants by optimizing calibration and adjusting areas of interest. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-02056-3

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Steffan, A., Zimmer, L., Arias-Trejo, N., Bohn, M., Dal Ben, R., Flores-Coronado, M. A., Franchin, L., Garbisch, I., Grosse Wiesmann, C., Hamlin, J. K., Havron, N., Hay, J. F., Hermansen, T. K., Jakobsen, K. V., Kalinke, S., Ko, E.-S., Kulke, L., Mayor, J., Meristo, M., … Schuwerk, T. (2023). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood. *Infancy*, 1–25. https://doi.org/10.1111/infa.12564