Research Article

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

*shared first-authorship
[1]Department of Psychology, Ludwig-Maximilians-Universität München, [2]Facultad de Psicología, Universidad Nacional Autónoma de México, [3]Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, [4]Faculty of Arts & Science, Ambrose University, [5]Department of Psychology and Cognitive Science, University of Trento, [6]Department of Developmental Psychology, University of Göttingen, [7]Research Group Milestones of Early Cognitive Development, Max Planck Institute for Human Cognitive and Brain Sciences, [8]Department of Psychology, The University of British Columbia, [9]School of Psychological Sciences & Center for the Study of Child Development, University of Haifa, [10]Department of Psychology, University of Tennessee, [11]Department of Psychology, University of Oslo, [12]Department of Psychology, James Madison University, [13] Department of English Language and Literature, Chosun University, [14]Neurocognitive

Developmental Psychology, Friedrich-Alexander-Universität Erlangen-Nürnberg, [15]Department of Psychology, University of Gothenburg, [16]School of Psychology and Centre for Brain Research, University of Auckland, [17]Department of Psychology, University of Miami, [18]Department of Psychology, University of Cambridge,

Author Note

† Correspondence concerning this article should be addressed to Lucie Zimmer, Ludwig-Maximilians-Universität, Leopoldstr. 13, 80802 München, Germany. Email: lucie.zimmer@psy.lmu.de

Author Contributions (CRediT report)

Conceptualization: *Leading:* Adrian Steffan. Tobias Schuwerk. *Supporting:* Lucie Zimmer, Laura Franchin, Charlotte Grosse Wiesmann, Louisa Kulke, Julien Mayor, Marek Meristo, Julia Prein, Elizabeth A. Simpson.
Data curation: *Leading:* Adrian Steffan, Daniela Santos Oliveira, Tobias Schuwerk. *Supporting:* Lucie Zimmer, Natalia Arias-Trejo, Rodrigo Dal Ben, Isa Garbisch, Marek Meristo, Seongmin Mun, Elizabeth A. Simpson.
Formal analysis: *Leading:* Lucie Zimmer, Tobias Schuwerk. *Supporting:* Adrian Steffan, Natalia Arias-Trejo, Manuel Bohn, Rodrigo Dal Ben, Isa Garbisch, Naomi Havron, Seongmin Mun.
Funding acquisition: *Leading:* Hannes Rakoczy, Tobias Schuwerk, Tone K. Hermansen, Marek Meristo. *Supporting:* Laura Franchin, Jessica F. Hay, Eon-Suk Ko, Karin Strid.
Investigation: *Leading:* Lucie Zimmer, Laura Franchin, Isa Garbisch, Jessica F. Hay, Tone K. Hermansen, Eon-Suk Ko, Julien Mayor, Julia Prein, Daniela Santos Oliveira. *Supporting:* Natalia Arias-Trejo, Manuel Bohn, Marco A. Flores-Coronado, Charlotte Grosse Wiesmann, J. Kiley Hamlin, Naomi Havron, Louisa Kulke, Marek Meristo, Seongmin Mun, Katrin Rothmaler, Elizabeth A. Simpson, Eleanor S. Smith, Karin Strid, Maleen Thiele, Anna-Lena Tebbe, Francis Yuen, Tobias Schuwerk.
Methodology: *Leading:* Tobias Schuwerk. *Supporting:* Adrian Steffan, Lucie Zimmer, Steven Kalinke, Louisa Kulke, Seongmin Mun, Anna-Lena Tebbe.
Project administration: *Leading:* Lucie Zimmer, Tobias Schuwerk. *Supporting:* Adrian Steffan, J. Kiley Hamlin, Marek Meristo, Hannes Rakoczy, Elizabeth A. Simpson, Anna-Lena Tebbe.
Resources: *Leading:* Adrian Steffan, Tobias Schuwerk. *Supporting:* Lucie Zimmer, Laura Franchin, Charlotte Grosse Wiesmann, Eon-Suk Ko, Louisa Kulke, Seongmin Mun, Katrin Rothmaler, Elizabeth A. Simpson, Eleanor S. Smith, Anna-Lena Tebbe.
Software: *Leading:* Adrian Steffan. *Supporting:* Steven Kalinke, Seongmin Mun.
Supervision: *Leading:* Adrian Steffan, Lucie Zimmer, Jessica F. Hay, Maleen Thiele, Tobias Schuwerk. *Supporting:* Laura Franchin, Charlotte Grosse Wiesmann, J. Kiley Hamlin, Tone K. Hermansen, Eon-Suk Ko, Julien Mayor, Marek Meristo, Julia Prein, Hannes Rakoczy, Elizabeth A. Simpson, Francis Yuen.
Validation: *Leading:* Lucie Zimmer. *Supporting:* Manuel Bohn, Rodrigo Dal Ben, Marek Meristo, David Moreau, Seongmin Mun, Tobias Schuwerk.
Visualization: *Leading:* Lucie Zimmer, Tobias Schuwerk. *Supporting:* Manuel Bohn, Isa Garbisch.
Writing - original draft: *Leading:* Adrian Steffan, Lucie Zimmer, Tobias Schuwerk. *Supporting:* Manuel Bohn, Laura Franchin, Naomi Havron, Krisztina V. Jakobsen, Louisa

Kulke, David Moreau, Katrin Rothmaler, Daniela Santos Oliveira, Elizabeth A. Simpson, Eleanor S. Smith, Maleen Thiele, Anna-Lena Tebbe.
Writing - review & editing: *Leading:* Lucie Zimmer, Tobias Schuwerk. *Supporting:* Natalia Arias-Trejo, Manuel Bohn, Rodrigo Dal Ben, Laura Franchin, Isa Garbisch, Charlotte Grosse Wiesmann, Naomi Havron, Tone K. Hermansen, Krisztina V. Jakobsen, Louisa Kulke, Julien Mayor, David Moreau, Julia Prein, Katrin Rothmaler, Daniela Santos Oliveira, Hannes Rakoczy, Elizabeth A. Simpson, Eleanor S. Smith, Karin Strid, Anna-Lena Tebbe, Francis Yuen.

Acknowledgements

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

Word Count: 9.986

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 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.

Keywords: Eye-tracking; Web-based eye-tracking; Anticipatory looking; Toddlers

Validation of an Open Source, Remote Web-based Eye-tracking Method (WebGazer) for Research in Early Childhood

Eye-tracking technology allows researchers to better understand childrens' interactions with the world. Compared to the manual coding of gaze behaviours, 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 behavioural research (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 and Werchan et al., 2022; for an overview, see, Kominsky et al., 2021; for in-person versus 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., 2020; 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 labor-intensive, requires training, and is difficult for large sample sizes. It may also reduce the replicability and reproducibility of the analysis due to the method's inherent levels of subjectivity. In contrast, automated web-based eye-tracking provides a resource-saving and less subjective alternative. 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 background, 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., 2021); (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 it is 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 30ms, leaving the data noisier.

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 WebGazer.js and jsPsych (de Leeuw, 2015; Papoutsaki et al., 2016). 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, Kampis 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. In line with the results of the ManyBabies2 project, we expect participants to utilize Theory of Mind to anticipate where the chaser will seek the hiding agent. We compared these anticipatory looking behaviors 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 online 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 towards the location that matches the outcome of an agent's action goal (i.e., finding the hiding agent). This would replicate Schuwerk, Kampis 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 usable 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.

Methods

The study was pre-registered. The pre-registration, all materials, data, and the analytic codes are available on Open Science Framework (OSF; https://osf.io/p3f67). The software implementing the experiment can be found on GitHub (https://github.com/adriansteffan/manywebcams-eyetracking/tree/848504f07fa8c25eb3f28444349a4d60151a7895).

Participation Details

In this multi-lab study, participants were recruited by 16 different labs and tested by 11 different labs. 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).

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.

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 https://osf.io/p3f67). Note that labs did not have to contribute 10 included participants. Each number of finally usable datasets was included in the overall sample.

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. An additional 118 participants were tested but excluded from the analysis for three main reasons: participant-related exclusions (n = 27), technical-related exclusions (n = 50), or exclusions after visual inspection (n = 41). 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. 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). Supplementary Table 1 provides additional details about included/excluded participants per lab.

The lab-based sample consisted of 70 toddlers (39 girls, 31 boys) aged between 18-27 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 participants tested for the pilot study (for the original analysis stricter criteria were applied which led to a final sample of 65 included participants; for details see in Schuwerk, Kampis et al., 2022).

Sample Size

Our sample size rationale was based on two effect sizes: Using the same paradigm with in-lab eye-tracking, Schuwerk, Kampis 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 the 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 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.

Materials and Design

The experimental design was identical to the familiarization phase of the paradigm previously developed for ManyBabies2 (https://manybabies.github.io/MB2/).

Stimuli

General scene setup

We used 3D animations representing a chasing scenario between two agents (chaser and chasee; Figure 1). The scene depicted an open blue-coloured room divided into two sections by a horizontal brown picket fence: an upper section, which was about 1/3 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.

Test trials

All participants viewed four trials (for a detailed description see Schuwerk, Kampis 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, made eye contact with the chaser 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). 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 OSF for the full animation.

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) vs. 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).

Apparatus and Procedure

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 (e.g., Google forms). 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 the computer, 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 towards 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 their browser. During the experiment, the participant's webcam recorded 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. During the initialization of the eye-tracking procedure, the software also controlled for the distance of the participant in relation to the monitor. The distance range accepted by the experiment's software spanned 40 to 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. 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.

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, an animated attention-getter was presented 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. 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, 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 minutes.

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.

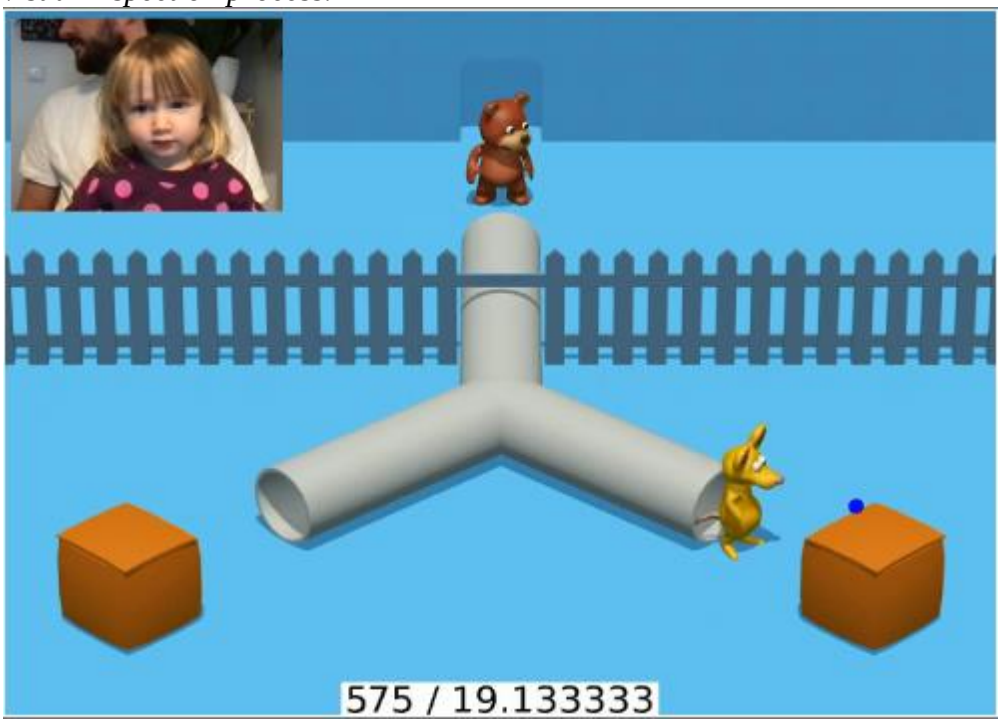
General procedure

We compared the data in the current study to the data collected by Schuwerk, Kampis 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 (Figure 1). 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 2). A gaze point collected with WebGazer has an area of uncertainty of about 100-200 pixels on 1920x1080 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, Kampis 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.

Figure 1.
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. These overlays were used for the visual inspection process.



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 towards 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). The experiment's software produced raw data for every participant/stimulus combination: For every 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, i.e., 4000 ms preceding the frame in which the chaser exited the tunnel. We then calculated what percentage of their gazes during this critical time frame fall into each category, for every participant/stimulus combination. This relative percentage was necessary, as sampling rates differed between participants. We computed the proportion of looking towards 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, Kampis 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, https://osf.io/b9nrs/), the gaze data were also resampled to 15 Hz; however, the resampled data were not used to run statistical analysis.

Figure 2.
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%).

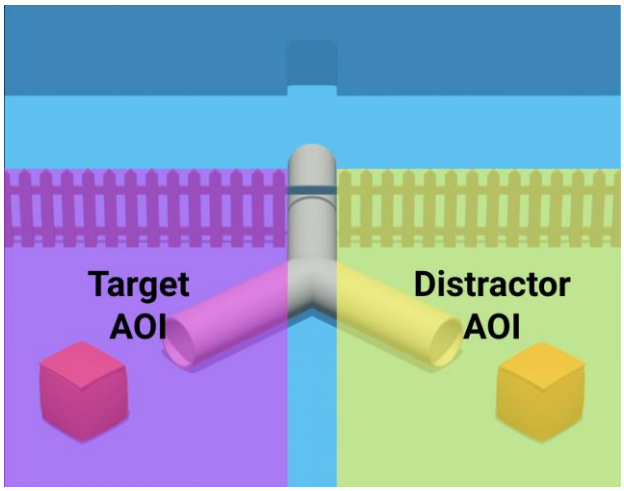
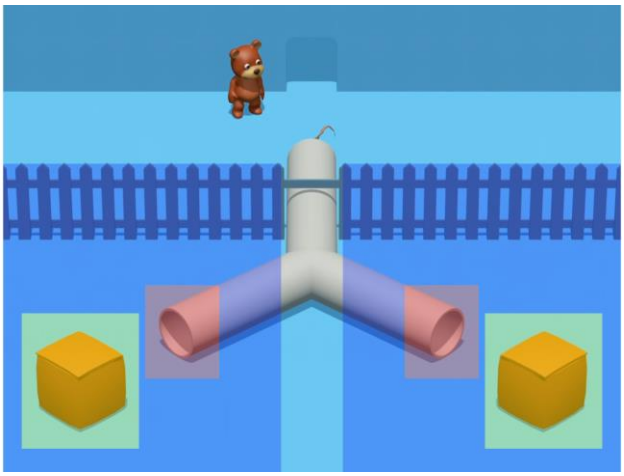


Figure 3.
Illustration of the additional AOIs we used (yellow and red). These replicated the "box" and "exit" AOIs of the in-lab data. Given the accuracy of the web-based eye-tracking, these were only used for exploratory analyses, while our main hypothesis was tested by comparing the data of the larger AOIs. (Dimensions relative to the stimulus video: Left red exit AOI: x: 23% - 35%, y: 20% - 40%; Right red exit AOI: x: 65% - 77%, y: 20% - 40%, Left yellow box AOI: x: 3% - 22%, y: 5% - 34%; Right yellow box AOI: x: 78% - 97%, y: 5% - 34%).



Data Exclusion

Participants were excluded from analyses if technical problems occurred or if participants did not provide at least one usable 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 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), and/or (2) 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: webcam video and recorded gaze coordinates stemmed from two different webcams, 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.

A third of all participants were randomly chosen and coded by a second naive rater to obtain interrater reliability. Cohen's kappa resulted in $\kappa = 0.74$, indicating a substantial inter-rater agreement.

Statistical Analyses

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 towards 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, Kampis et al. (2022) in a linear mixed effects model using the lme4 package (Bates et al., 2015). This model was set to predict the proportional looking[1] score based on the fixed effect method (web-based vs. lab-based) and a random effect for labs and participants. Significance was calculated using the lmerTest package (Kuznetsova et al., 2017), which applies Satterthwaite's method to estimate degrees of freedom and generate p-values for mixed models. The model specification was:

Proportional looking score ~ method + (1|lab)

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.

[1]Our models assumed that the response variable would be normally distributed. However, given that proportions are bound to be between 0 and 1, a beta distribution would be a more appropriate way to model the response. While this is difficult in the frequentist framework, it can easily be done in a Bayesian one. In the online repository, we, therefore, report a series of Bayesian Multilevel models. Notably, the inferences drawn from these models are the same as for the frequentist models reported in the text.

Exploratory Analysis

To investigate potential effects of age on the proportional looking score, standardized age (z-scores) was added to the model as a fixed effect. Lab was included as random effects. The model specification was:

Proportional looking score ~ method + z_age + (1|lab)

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 and the sampling rate as fixed effects. Lab was included as random effects. The model specification was:

Proportional looking score ~ z_age + sampling rate + (1|lab)

Results

Confirmatory Analysis

Anticipatory looking behavior

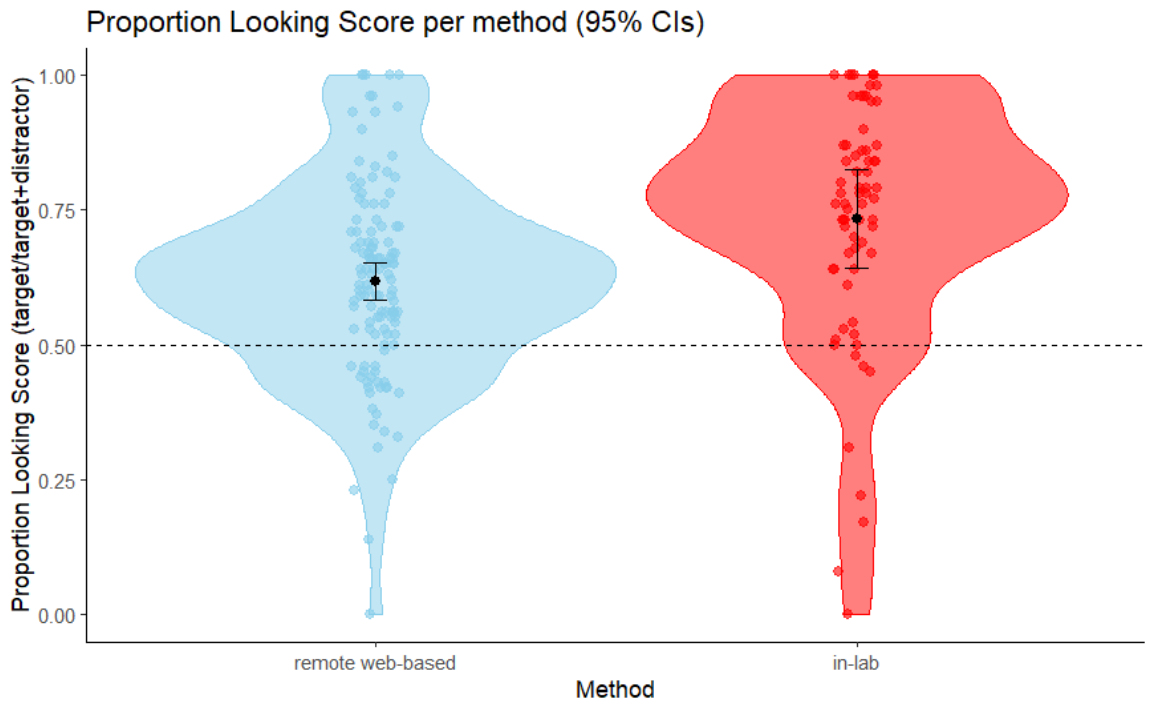
In our web-based sample, the relative looking time towards the location that matched the outcome of the agent's action goal (target AOI; $M = 0.62$, $SD = 0.18$; Figure 4) was significantly different from chance level (0.5), $t(124) = 7.34$, $p < 0.001$, indicating that the participants anticipated the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

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, Kampis et al. (2022) observed a larger effect size of Cohen's $d = 1.03$ (95% confidence interval: 0.50 – 1.56).

Comparison of remote web-based vs. in-lab eye-tracking in toddlers

To test whether the method had an effect on the proportional looking score, we fit a linear mixed model and found a significant main effect of method ($\beta = 0.11$, $t = 3.86$, $p < .001$), reflecting the fact that the proportion of goal-directed anticipatory looking was higher in the in-lab sample (Figure 4).

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).

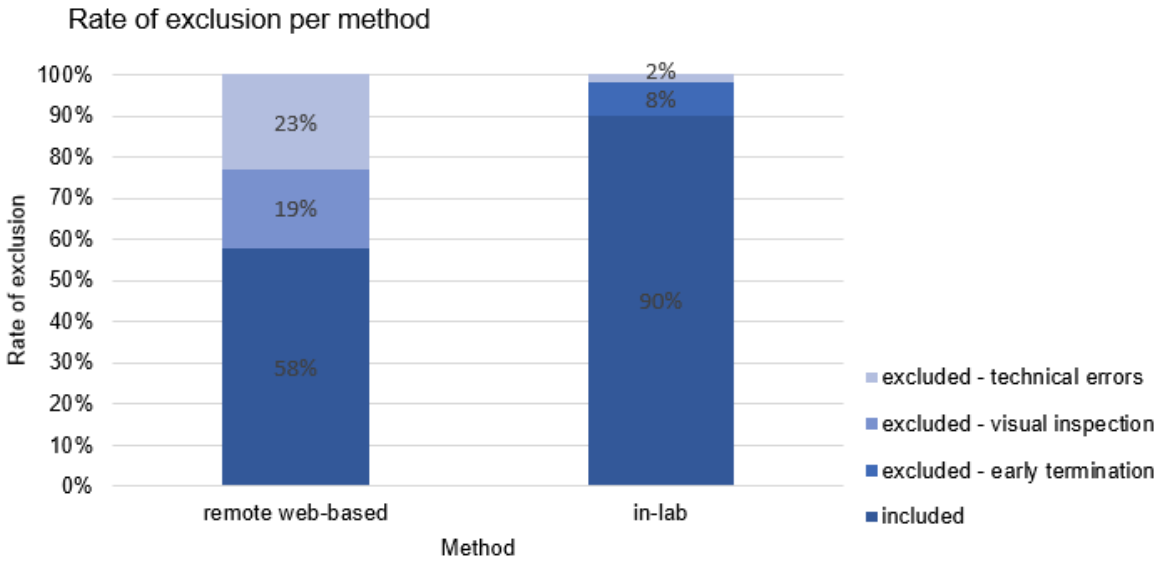


Rate of exclusion

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, 50 toddlers (55% of excluded participants) were excluded due to technical reasons. Technical problems occurred for instance during the stimulus presentation or during data transmission from the participating families to the experimenters

(*n* = 33), a sampling rate below our predefined threshold (*n* = 8), experimenter error (*n* = 2) or technical error without further information (*n* = 7). As a result of the visual inspection process, a total of 41 toddlers were excluded (45% of excluded participants). They were excluded due to a mismatch between gaze coordinates and their head/eyes movement (*n* = 20), interference by caregiver (*n* = 6), inattentiveness of the toddler (*n* = 5), two different active webcams (*n* = 5), suboptimal positioning of the toddler (*n* = 1) and error without further information (*n* = 4). 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, Kampis 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* < .001. See Figure 5 for a comparison of exclusions for in-lab versus web-based methods.

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).



Exploratory Analysis

Change in tracking quality for the web-based sample

We ran calculations for x/y deviations during validation trials for all participants tested by LMU Munich team (n = 71; 56% of the final sample). 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 12.92% (SD = 12.6%) for x coordinates and a mean deviation of 14.49% (SD = 19.40%) for y coordinates. We performed a two-tailed t test for paired samples to compare both validation timepoints. We found no significant difference for either coordinate (X differences: M = 0.314%, SD = 12.603%, $t(70)$ = 0.182, p = .856, delta = .022; Y differences: M = 4.051%, SD = 19.402% , $t(70)$ = 1.418, p = .161, delta = .168). We thus assume that tracking quality did not deteriorate significantly during the trials and did not check for tracking deterioration on the remaining participants.

Age analysis

Using the previously described mixed effects model, we did not find a statistically significant effect of age on the proportional looking score (β = -0.01, t = -0.98, p = 0.330), meaning that in our sample the toddler's age had no influence on anticipatory looking in the web-based task.

Sampling rate analysis

We observed sampling rates between 10.42Hz and 40.10Hz, resulting in a mean sampling rate of 22Hz (SD = 7.3Hz) in our web-based sample after exclusions. We did not find a statistically significant effect of the sampling rate on the proportional looking score (β = 0.001, t = 0.47, p = .638), meaning that the sampling rate had no effect on anticipatory looking in our remote sample.

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, Kampis 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. By testing children remotely and collaboratively, we were able to access participants from all over 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.

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 towards 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, Kampis et al., 2022). This finding shows that web-based eye tracking can be used successfully to assess children's goal-based action predictions. This finding 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; Chuey et al., 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.

Comparing Performance of Web-Based vs. 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 towards 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 60Hz to 500Hz. 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 towards chance-level.

Comparing Data Quality of Web-Based and In-Lab Eye-Tracking

We found support for our hypothesis that the proportion of children who contributed usable 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 accidentally 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. 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., 2021), 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., 2022).

Limitations

While this study examined the replicability of an in-lab paradigm, we did not explicitly measure the accuracy of WebGazer for infants. Using an in-lab eye-tracker concurrently while running a WebGazer experiment could provide us with ground truth 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.

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, Kampis et al. (2022) under these conditions. Still, paradigms that use the full width of the screen (33% increase in presentation space) would be even less bothered by the accuracy drop from using WebGazer as opposed to in-lab eye-tracking.

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 OS, or internet

connection strength. 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. Nonetheless, 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.

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 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.

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. Additionally, free, open-source approaches such as WebGazer have several advantages over these commercial platforms. First, the transparency of open-source code is desirable in a research context, as it

allows other researchers to verify the validity of the research and promotes openness and accessibility in research, 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.

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, OWLET only allows a fixed four-point calibration. 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 utilises the improved accuracy of OWLET with the convenience and flexibility that WebGazer provides.

Deep Learning

While WebGazer and OWLET use traditional computer vision algorithms, applying 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 consist 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 gazes 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.

Conclusion

Web-based eye-tracking can be used to capture toddlers' goal-based action anticipation. Thus, in-lab findings can be replicated using remote web-based testing. 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, they come with substantially higher costs and physical 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 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 a reduced spatial resolution can be tolerated, web-based webcam eye-tracking using WebGazer is a promising method.

References

Apache Software Foundation (2012). Apache HTTP Server (Version 2.4) [Computer Software]. https://httpd.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*, 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*, 733933. https://doi.org/10.3389/fpsyg.2021.733933

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

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

Canonical Ltd. (2018). Ubuntu (Version 18.04 LTS) [Computer Software]. https://ubuntu.com/

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*, 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

Dalrymple, K. A., Manner, M. D., Harmelink, K. A., Teska, E. P., & Elison, J. T. (2018). An examination of recording accuracy and precision from eye tracking data from toddlerhood to adulthood. *Frontiers in Psychology*, *9*, 803. https://doi.org/10.3389/fpsyg.2018.00803

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

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). "LabVanced: a unified JavaScript framework for online studies [Conference paper]," in International Conference on Computational Social Science IC2S2S, Cologne, (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., Kasneci, E., Kingstone, A., Knox, P. C., Kok, E. M., Lee, H., Lee, J. Y., Leppänen, J. M., Macknik, S., Majaranta, P., Martinez-Conde, S., Nuthmann, A., Nyström, M., Orquin, J. L., Otero-Millan, J., Park, S. Y., Popelka, S., Proudlock, F., Renkewitz, F., Roorda, A., Schulte-Mecklenbeck, M., Sharif, B., Shic, F., Shovman, M., Thomas, M. G., Venrooij, W., Zemblys, R., & Hessels, R. S. (2022). Eye tracking: empirical foundations for a minimal reporting guideline. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01762-8

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, 702710. https://doi.org/10.3389/fpsyg.2021.702710

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

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

Lewandowska, B. (2019). RealEye Eye-tracking system Technology Whitepaper. Retrieved December 19, 2022, from https://support.realeye.io/realeye-accuracy/

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

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*, 3866. https://doi.org/10.3389/fpsyg.2021.733218

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*, 168–183. https://doi.org/10.1080/15295192.2019.1694838

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI), 3839–3845.

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. Vienna, Austria. 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*(4573). 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, J. K., Haun, D., Hepach, R., Hunnius, S., Hyde, D. C., Kármán, P., Kosakowski, H. L., Kovács, Á. M., Krämer, A., Kulke, L., Lee, C., Lew-Williams, C., Liszkowski, U., Mahowald, K., Mascaro, O., Meyer, M., Moreau, D., Perner, J., Poulin-Dubois, D., Powell, L. J., Prein, J., Priewasser, B., Proft, M., Raz, G., Reschke, P., Ross, J., Rothmaler, K., Saxe, R., Schneider, D., Southgate, V., Surian, L., Tebbe, A.-L., Träuble, B., Tsui, A. S. M., Wertz, A. E., Woodward, A., Yuen, F., Yuile, A. R., Zellner, L., Zimmer, L., Frank, M.C., & 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 (*shared co-first authorship).

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

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]. https://www.php.net/

Van Rossum, G., & Drake, F. L. (2009). Python 3 reference manual. CreateSpace.

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. https://doi.org/10.1038/s41467-020-18360-5

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., Simpson, E. A., Singh, L., Soderstrom, M., Sullivan, J., van den Heuvel, M., Westermann, G., Yamada, Y., Zaadnoordijk, L., & Zettersten, M. (2022). Improving the generalizability of infant psychological research: The ManyBabies model. *Behavioral and Brain Sciences, 45,* e35. https://doi.org/10.1017/S0140525X21000455

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

Wass, S. V. (2016). "The use of eye-tracking with infants and children," in Practical Research with Children 1st ed, eds J. Prior and J. Van Herwegen (Milton Park: Routledge), 24–45. https://doi.org/10.4324/9781315676067

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*. https://doi.org/10.3758/s13428-022-01962-w

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

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

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*, 703234. https://doi.org/10.3389/fpsyg.2021.703234

Supplementary Materials

Deviations from pre-registration

We intended to perform the eye-tracking quality check before and after the trials by calculating the percentage of gazes falling within a 200-pixel radius of an attention getter. (In doing so, 100% would have been a near perfect calibration for the radius, while lower percentages represent a non-optimal calibration quality.) However, the fixed radius made comparisons among differing screen resolutions difficult to interpret, so we instead looked at the deviation from the attention getter, transformed to a percentage of the screen's resolution.

Supplementary Table 1.
Overview of the number of participants included and excluded per lab.

Lab	*N* included	*N* excluded participant-related exclusion	*N* excluded technical-related exclusion	*N* excluded after visual inspection
BLT_Trento	10	2	1	4
CBL_Gwangju	5	3	2	2
GAUG_Göttingen	12	5	-	9
INCH_Gothenburg	9	-	1	-
LMU_Munich*	48	7	21	10
MPI_EVA_Leipzig	5	-	3	2
UBC_Vancouver	6	-	7	7
UH_Haifa	5	3	2	-
UIO_Oslo	11	-	8	1
UNAM_CdMéxico	4	7	1	5
UTK_Knoxville	10	-	4	1

* Note that participants were tested by LMU_Munich, but recruited by PaL_Cambridge, FAU_Erlangen, MPI_HCBS_Leipzig, PLUS_Salzburg, UM_CoralGables, and LMU_Munich.