Developmental Psychology

An Asynchronous, Hands-Off Workflow for Looking Time Experiments With Infants

Gal Raz, Sabrina Piccolo, Janine Medrano, Shari Liu, Kirsten Lydic, Catherine Mei, Victoria Nguyen, Tianmin Shu, and Rebecca Saxe

Online First Publication, June 24, 2024. https://dx.doi.org/10.1037/dev0001791

CITATION

Raz, G., Piccolo, S., Medrano, J., Liu, S., Lydic, K., Mei, C., Nguyen, V., Shu, T., & Saxe, R. (2024). An asynchronous, hands-off workflow for looking time experiments with infants.. *Developmental Psychology*. Advance online publication. https://dx.doi.org/10.1037/dev0001791

AMERICAN
PSYCHOLOGICAL
ASSOCIATION

Developmental Psychology

ISSN: 0012-1649

https://doi.org/10.1037/dev0001791

An Asynchronous, Hands-Off Workflow for Looking Time Experiments With Infants

Gal Raz[1], Sabrina Piccolo[2], Janine Medrano[1], Shari Liu[3], Kirsten Lydic[4], Catherine Mei[1], Victoria Nguyen[1], Tianmin Shu[5, 6], and Rebecca Saxe[1, 7]

[1] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[2] Department of Psychology, Northeastern University
[3] Department of Psychological and Brain Sciences, Johns Hopkins University
[4] Annenberg School for Communication, University of Pennsylvania
[5] Department of Cognitive Science, Johns Hopkins University
[6] Department of Computer Science, Johns Hopkins University
[7] McGovern Institute for Brain Research, Massachusetts Institute of Technology

The study of infant gaze has long been a key tool for understanding the developing mind. However, labor-intensive data collection and processing limit the speed at which this understanding can be advanced. Here, we demonstrate an asynchronous workflow for conducting violation-of-expectation (VoE) experiments, which is fully "hands-off" for the experimenter. We first replicate four classic VoE experiments in a synchronous online setting, and show that VoE can generate highly replicable effects through remote testing. We then confirm the accuracy of a state-of-the-art gaze annotation software, iCatcher+ in a new setting. Third, we train parents to control the experiment flow based on the infant's gaze. Combining all three innovations, we then conduct an asynchronous automated infant-contingent VoE experiment. The hands-off workflow successfully replicates a classic VoE effect: infants look longer at inefficient actions than efficient ones. We compare the resulting effect size and statistical power to the same study run in-lab and synchronously via Zoom. The hands-off workflow significantly reduces the marginal cost and time per participant, enabling larger sample sizes. By enhancing the reproducibility and robustness of findings relying on infant looking, this workflow could help support a cumulative science of infant cognition. Tools to implement the workflow are openly available.

Public Significance Statement

Infant looking time experiments have provided critical insights into early cognition, but traditionally very time-consuming and expensive. We run a classical violation-of-expectation experiment through a workflow in which data collection and analysis are automated and compare the results to the same study run in the lab and on Zoom. The automated workflow shows a small reduction in effect size and power, while allowing for significantly larger sample sizes, thereby enabling a more robust developmental science.

Keywords: looking time, violation-of-expectation, asynchronous testing, replication, open tools

Supplemental materials: https://doi.org/10.1037/dev0001791.supp

Melissa M. Kibbe served as action editor.

Gal Raz https://orcid.org/0000-0002-0712-7560

The authors gratefully acknowledge funding from Defense Sciences Office, Defense Advanced Research Projects Agency (Grant 6942927) awarded to Rebecca Saxe and the K. Lisa Yang Integrative Computational Neuroscience Center (Grant 3327570) awarded to Gal Raz. The authors thank Jordan Wilke for help with supervising the automated gaze annotations. The authors thank Melissa Kline-Struhl, Kevin Smith, Tomer Ullman, and Josh Tenenbaum for helpful discussion, and Emily Chen, Mike Frank, Young-eun Kim, and Linette Kunin for feedback on an earlier draft of this article.

Gal Raz played a lead role in conceptualization, formal analysis, investigation, methodology, and visualization and an equal role in data curation, software, writing–original draft, and writing–review and editing. Sabrina Piccolo played a supporting role in data curation, formal analysis, visualization, and writing–original draft and an equal role in investigation. Janine Medrano played a supporting role in formal analysis and investigation and an equal role in data curation. Shari Liu played a supporting role in funding acquisition, supervision, writing–original draft, and writing–review and editing. Kirsten Lydic played a supporting role in writing–review and editing and an equal role in software. Catherine Mei played a supporting role in data curation and formal analysis. Victoria Nguyen played a supporting

continued

To investigate infant cognition, many studies measure how long infants look at events that are designed to surprise them (Aslin, 2007; Spelke, 2022). In these violation-of-expectation (VoE) experiments, infants are first exposed to "familiarization" events that evoke a core principle of the social or physical world (e.g., agents pursue goals, objects are solid). Next, infants are presented with "test" events that either conform to that principle (the expected event) or violate that principle (the unexpected event). When infants look longer at the unexpected event, as long as other features of the displays are well-controlled, researchers conclude that infants understand the underlying physical or social principle (Spelke & Kinzler, 2007). Because such understanding can be probed by measuring looking duration months or years before infants can express abstract principles in words or in other behavior, VoE experiments in many cases offer the earliest evidence of abstract cognition in infant development.

A key challenge for VoE experiments is that data collection and processing are slow and labor intensive. As a result, sample sizes in traditional VoE experiments are small, with typical sample sizes between 16 and 24 infants (Bergmann et al., 2018; Oakes, 2017). The small samples reflect the high marginal cost of each additional infant included in a study. First, infants are traditionally recruited to come to the lab, requiring time for transportation and familiarization with the novel environment. One or two experimenters typically spend up to an hour with each infant and their family, over the course of the visit. The experiments themselves are often administered manually by an experimenter who "stages" the events for the infant and at least one experimenter who records the infant's gaze, to ensure that delivery of the experimental stimuli is contingent on the infant's interest. For example, the experimenter may wait for the infant to attend to the "stage" before beginning an event; and then wait for the infant to lose interest in the stimulus before removing it, to start the next event. Some of the infants withdraw from the study, or do not meet inclusion criteria, increasing the marginal cost of each additional data point that is included (Byers-Heinlein et al., 2022).

The marginal costs continue to accumulate during data processing. Infant gaze is typically recorded on video during the experiment. Afterward, looking duration to each trial is extracted by trained researchers. The current gold standard for computing looking times (LTs) is manual frame-by-frame annotation through dedicated software like Datavyu (Datavyu Team, 2014). Annotating a video typically takes at least three times as long as the video itself, requiring more than half an hour of work to annotate an experiment that lasted less than 10 minutes. Overall, generating average looking times to two events (one expected, one unexpected), from a single infant participating in one experimental session, can take 2 hr of effort from a trained experimenter.

The slow and labor-intensive pace of VoE experiments poses a challenge for the rigor and replicability of claims about infant cognition. Recent large-scale, multisite collaborations suggest that true effect sizes in studies of infant gaze are smaller than originally reported, and many published effects are likely moderated by tertiary variables (e.g., ManyBabies Consortium, 2020). A robust foundation for the study of infant cognition requires larger and more diverse samples of infants (Amir & McAuliffe, 2020; Frank et al., 2017). Thus, the practical barriers to including larger samples of infants in VoE experiments must be lowered.

Technical innovations are in progress, to ease the workload associated with both data collection and annotation in infant gaze experiments. Since the COVID-19 pandemic, remote synchronous testing, especially via Zoom, has become increasingly common, easing the burden on experimenters and families to schedule and travel to in-person experimental sessions (e.g., Smith-Flores et al., 2022; Zaadnoordijk et al., 2021). Remote testing also makes a larger and more diverse pool of participants available to developmental researchers (Sheskin et al., 2020). In many cases, remote studies via Zoom appear to replicate in-lab studies of children, with comparable effect sizes (Chuey et al., 2021, 2022; though see Lapidow et al., 2021).

Asynchronous testing has developed in parallel as a way to take experimenters out of the loop completely and allow participants to do studies in their own time. For example, LookIt (operated by Children Helping Science) is a platform that hosts experiments, where children and their families can sign in to participate at their own convenience (Scott et al., 2017; Scott & Schulz, 2017). In LookIt studies, the flow of the experiment is automated and does not require an experimenter to be synchronously present for the instructions, consent process, experimental paradigm, or debriefing. Thus, data collection can proceed completely hands-off for many developmental studies.

On the data processing side, modern computer vision tools are being developed for automatic gaze annotation. OpenFace (Baltrusaitis et al., 2018), RT-GENE (Fischer et al., 2018), WebGazer (Papoutsaki, 2015; Papoutsaki et al., 2018; Steffan et al., 2023) and OWLET (Werchan et al., 2023) are modern tools that estimate gaze direction from webcam videos. While these tools have been offered as general solutions to the problem of tracking gaze, automating the measurement of infant looking time in remote settings poses a set of specific challenges, such as detecting the infant among multiple faces in the scene, and correcting for movement, body position and variable camera angles. iCatcher+, the tool used in this article, was designed to address these specific challenges, and has recently been reported to approach human-to-human reliability in annotating the gaze of infants in experimental studies (Erel et al., 2023).

Importantly, it is not currently possible to asynchronously deliver experimental stimuli contingent on an infant's looking behavior. In a classic VoE experiment, for example, both the habituation procedure and the test trials typically use a lookaway criterion: Trials are terminated when the infant continuously looks away from the screen or stage for a specified amount of time, usually 2 s (Horowitz et al., 1972). This procedure serves to tune the amount of exposure to each infant, and ideally to match the depth of encoding of the familiarization events across infants. Although automated gaze annotation can be used offline, it is not yet possible to use automated gaze coding online in the experimental loop, to detect and respond to infants' (loss of) interest in the stimuli.

role in data curation and formal analysis. Tianmin Shu played a supporting role in methodology and software. Rebecca Saxe played a lead role in conceptualization, funding acquisition, and supervision and an equal role in writing–original draft and writing–review and editing.

Correspondence concerning this article should be addressed to Gal Raz, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 43 Vassar Street, Cambridge, MA 02139, United States. Email: galraz@mit.edu

Here, we develop and test a hands-off workflow which leverages and adds to prior advances to address bottlenecks in data collection and processing, by automating these steps for the experimenter. Our project was composed of four steps. First, we attempted to replicate four classic violation-of-expectation (VoE) findings using synchronous online Zoom testing and standard manual annotation. Out of the four paradigms, we chose the one with the most robust VoE effect: infants look longer at inefficient actions than efficient actions (the "action efficiency" paradigm, Gergely et al., 1995; Liu & Spelke, 2017). Second, we replaced manual annotation of looking time with automatic annotation by iCatcher+. Third, we trained parents to make the flow of the experiment contingent on their infant's gaze. Finally, we re-ran the action efficiency paradigm through our workflow. We estimated the effect of automation on effect size and statistical power by comparing the same experiment run in three settings: in-lab, on Zoom, and using the hands-off workflow.

The overall goal was to establish how much we can accelerate the research lifecycle of infant gaze research using a hands-off workflow, balancing cheaper data acquisition against additional noise, and hence costs to statistical power. We demonstrate that we can easily run hundreds of infants through our workflow, with small marginal per infant time investment. We replicate the VoE effect demonstrated by Liu and Spelke (2017), including effect sizes and statistical power that are only slightly lower to those obtained from conducting the same study conducted on Zoom. Tools to implement the workflow are fully available at https://osf.io/ndkt6/?view_only=1984f6599dc44e37ae8c984465a25c0f/ (Raz, 2024).

Developing Tools

To implement our hands-off workflow, we took three preparatory steps. First, we chose an experimental paradigm to test the workflow by replicating a series of classic VoE studies in a remote but synchronous testing setup. Second, we confirmed iCatcher+'s ability to capture human looking time coding out-of-the-box. Third, we trained and validated parents' ability to control the flow of the VoE paradigm in an asynchronous setting.

Transparency and Openness

All anonymized data, preregistration documents, and analysis scripts associated with this article are openly available at https://osf.io/ndkt6/.

Stimulus Selection

To select a paradigm appropriate for validating a hands-off workflow, we attempted to replicate the main results of several VoE paradigms using online video for stimulus presentation and webcams for data collection. Infants have been reported to look longer at violations of intuitive psychological principles, such as when agents seem to act inefficiently toward a goal object (Gergely et al., 1995), or seem to violate their previously shown preferences between goal objects (Woodward, 1998). In the physical domain, infants look longer when objects seem to pass through each other (Baillargeon et al., 1985), or when objects do not fall when pushed over an edge (Needham & Baillargeon, 1993).

Four experiments were run as Zoom studies, synchronously controlled by experimenters. One of the experiments was previously published as part of Chuey et al. (2021), and the remaining three experiments were preregistered here at https://osf.io/7vu23?view_only=6722cddbafc94bf4a1dac7e027360217. All four experiments began with six infant-controlled familiarization trials demonstrating one psychological or physical principle (action efficiency, goal persistence, object solidity or object support). These trials were followed by four test trials either violating the principle (the "unexpected" trial), or conforming with it ("the expected" trial). These experiments were moderated by an experimenter using pyHab's online version (Kominsky, 2019; Peirce et al., 2019). An overview of the experimental design is shown in Supplemental Figure S1 and described in detail in the Supplemental Materials. Stimuli can be found on the Open Science Framework (OSF) page (https://osf.io/ndkt6/?view_only=1984f6599dc44e37ae8c984465a25c0f/) under "stimuli."

We tested a total of 134 infants (M_{age} = 7.96 months, range: 5–13 months) across the four experiments (N_efficiency = 28, N_goal = 26, N_solidity = 32, N_support = 48). We assessed the presence of a VoE effect using a preregistered linear mixed effects model of the form: log(looking time) ~ test_type [expected versus unexpected] + trial_n [1, 2, 3, 4] + (1|subj), with a one-tailed test on test type. We replicated VoE effects for three out of the four paradigms. Figure 1A shows unstandardized effect sizes in each of the VoE studies: Infants looked longer at inefficient jumps, efficiency; $\beta = 0.28$, $t(66.60) = 2.08$, $p = .021$; Supplemental Table S1, actions incongruent with previous goals, goals; $\beta = 0.20$, $t(56.54) = 1.72$, $p = .045$; Supplemental Table S2, and objects floating in mid-air, support; $\beta = 0.23$, $t(115.38) = 2.18$, $p = .016$; Supplemental Table S3. Infants did not look longer at a fan seeming to rotate through another object, solidity; $\beta = -0.07$, $t(80.35) = -0.55$, $p = .292$; Supplemental Table S4, consistent with results of a similar online study which failed to find an effect of solidity violations on infants' looking (Smith-Flores et al., 2022), despite recent in-lab replications (Perez & Feigenson, 2022). Raw and mean looking times are shown in Supplemental Figure S2A.

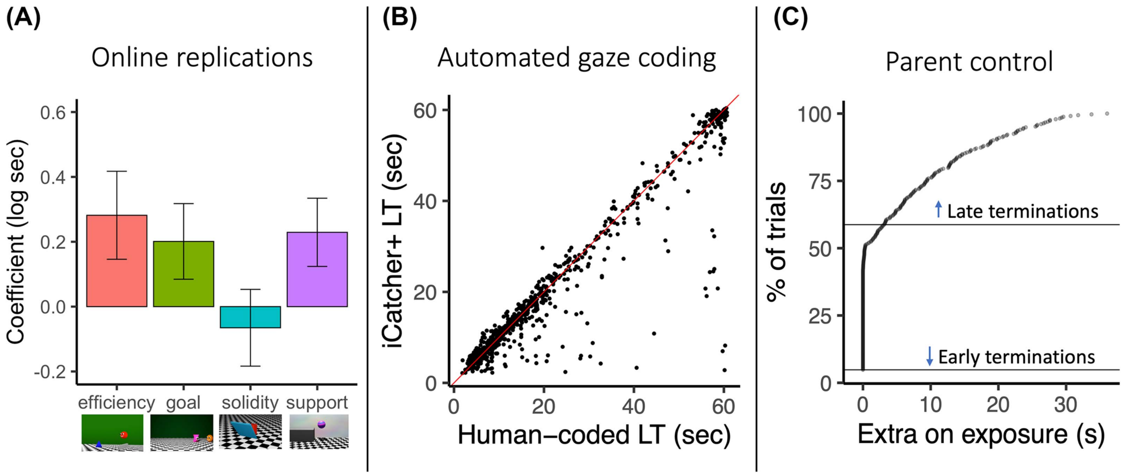
We use the coefficients on "test type" (expected vs. unexpected) as the primary measure of effect size, given that there is no consensus around a standardized effect size measure for within-subject designs like the VOE design (see e.g., Lakens, 2013). Using Cohen's D did not qualitatively change the results (Supplemental Figure S2B). Since the efficiency paradigm showed the largest coefficient on expectedness, we selected this paradigm to use for testing the fully hands-off workflow. Another advantage was the availability of previously published data collected in-lab using the same stimuli (Liu & Spelke, 2017; Experiment 1).

Automatic Annotation

After data collection, data annotation is a second central bottleneck in extracting trial-level looking times. To automate this process, we used a recent model for automated gaze coding, iCatcher+, which is a neural network trained for coarse gaze classification from video (Erel et al., 2023). Importantly, iCatcher+ was pretrained on webcam videos with infants and children, and therefore can be used out-of-the-box without any training. Furthermore, it can run a local machine, therefore complying with most ethics protocols which do not allow cloud processing of sensitive videos.

Figure 1

Validation of Hands-Off Workflow Tools



Note. (A) Three out of four VoE experiments replicated the main effect of longer looking at unexpected events (error bars are standard errors). (B) iCatcher+-coded looking times (Erel et al., 2023) closely reflect human-coded looking times. (C) Percentage of trials where lookaway has not yet been reached, plotted against how much time is left in the trial. Parents controlling the flow of violation-of-expectation experiments leads to minimal premature trial terminations. VoE = violation-of-expectation; LT = looking time. See the online article for the color version of this figure.

To validate the ability of iCatcher+ to match human annotations in this new setup, we replicated analyses from the original article by Erel et al. (2023) on our own data. We pooled the four Zoom data sets above, annotated them using iCatcher+ and merged the annotations with trial timing information to produce trial-wise looking times. We then compared the human-coded and iCatcher+-coded looking times, which is visualized in Figure 1B. We find high intraclass correlations, with an overall intraclass correlations of 0.944, approximating typical agreement between trained human coders. We found this high agreement despite coding looking duration until a threshold lookaway criterion (2 consecutive seconds; Horowitz et al., 1972) for both human and iCatcher+ coded looking times. A lookaway criterion could in principle lead to larger discrepancies than agreement over a fixed period as used in Erel et al. (2023), because disagreement between humans and iCatcher+ about whether a lookaway was reached or not could lead to differences in the period over which looking time was calculated. In sum, iCatcher+ is a robust strategy for coding the duration of infant attention. We provide a detailed manual and scripts for implementing the automatic annotation workflow here on OSF https://osf.io/ndkt6/?view_only=1984f6599dc44e37ae8c984465a25c0f under “automatic_annotation.”

Parent Control

A key feature of VoE experiments is the infant-contingent experimental design: trials during habituation are terminated by experimenters when infants look away. This procedure serves to tune the amount of exposure to the infants’ interest in the stimuli, to make it more likely that different infants see the test stimuli after similar levels of encoding. Test trials are likewise terminated when infants look away, to maximize their engagement with the subsequent test trials.

In asynchronous setups, however, experimenters are not present to terminate trials, and trial durations are typically fixed. To address this challenge, we trained parents to control the flow of experimental trials by monitoring their child’s looking behavior and terminating familiarization and test trials. To do so, we parents received the following instructions:

> When your child has looked away, start counting slowly, and if your child is still looking away after three seconds, please press the space bar. If your child looks back on the screen before three seconds have passed, you can stop counting, and start again once they look away. Let’s see how that looks in practice!

These instructions were followed by two training videos: (1) a video of a parent and their child facing the camera, and the parent demonstrating how to terminate a trial when their child looked away, and (2) a video shot from behind a child looking at the screen, and looking away after some time. On the second video, participants’ parents were asked to terminate the trial upon a 3-s lookaway. The training videos and associated Lookit generator code are free to use by other researchers and can be found on OSF (https://osf.io/ndkt6/?view_only=1984f6599dc44e37ae8c984465a25c0f/, under “materials”).

The criterion lookaway duration we used in the analysis was 2 s, while parents were instructed to wait 3 s before terminating trials. The 1-s discrepancy between analysis and instructions given to the parents served to reduce the risk of premature trial terminations. The same discrepancy was used for online experimenter-controlled designs too, such as the Zoom studies in Section Stimulus Selection, where we expected that live coding looking during the experiment would be more difficult, because of variability in screen size and lag due to poor connection. The in-lab experiment (Liu & Spelke, 2017) did not have these issues, so experimenters ended trials exactly at 2 s.

We collected data from 35 infants (M_{age} = 8.8 months, range: 7.1–11.2 months) on Lookit who underwent a habituation/test procedure, like the VoE experiments run on Zoom. After training the participant’s parent, each infant saw a series of habituation trials in which an animated creature reached a goal object by overcoming an obstacle.

To assess the quality of parents' trial terminations, we related the point at which the parent stopped each trial to the infant's looking behavior as coded by iCatcher+. There are two types of mistakes parents could make: early terminations and late terminations. Parents could terminate trials prior to their child reaching the criterion lookaway, resulting in an early termination. If a trial is prematurely ended, the looking time during that trial becomes invalid, because one cannot determine how long the infant would have continued to look if the trial had not been terminated. Late terminations are trials in which a criterion lookaway was reached but parents did not terminate the trial. Late terminations are less problematic, because looking time is measured until the lookaway criterion was reached, irrespective of whether the parent terminated the trial late or not. However, late terminations result in overexposure of the infant, thereby increasing the risk of dropout in later trials.

Figure 1C shows the percentage of trials in which lookaway has not yet been reached, plotted against the time left in the trial. When we compared the timing of parents' trial terminations again the ground truth provided by iCatcher+, we found that parents prematurely terminated only 4.8% of trials, 16 of 333. Late terminations led to an average of 5.8 extra seconds of extra exposure per trial, averaged over all trials.

Testing the Hands-Off Workflow

We next combined these tools into an end-to-end hands-off infant behavioral testing workflow. We re-ran the action efficiency experiment on Lookit using the stimuli from Liu and Spelke (2017), trained parents to control the flow of the experiment based on their child's behavior, and automatically coded the data set with iCatcher+ with minimal supervision.

Participants

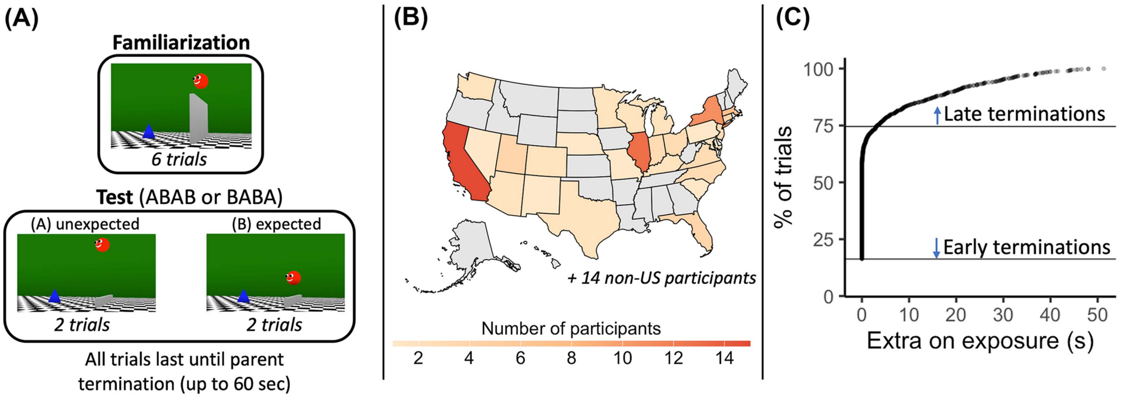
Given the ease of data acquisition and annotation in our hands-off workflow, we collected a larger sample size in this setup: 133 infants aged between 7 and 11.2 months (M = 8.9 months, 67 female). The sample size was a result of our plan to stop data collection in May 2023. Videos from 38 infants were initially excluded due to fussiness, parental interference, poor lighting conditions, frequent movement, unusual camera positions (e.g., on a different screen from the one where stimuli were displayed), or poor iCatcher performance for other (unknown) reasons. For the cases of poor iCatcher performance, we attempted to save the videos by using a new version of iCatcher V0.1.0 (released in July 2023). Seventeen videos were successfully recovered with the new release, leaving 21 infants excluded and unrecoverable. Of the remaining 112 infants, 106 caregivers provided information about the race and ethnicity of the infants: Sixty-four were White, 16 were Asian, two were Hispanic, one was Middle Eastern, one was African American, and 22 participants were multiracial. Of the 112 infants, 97 were from the United States, living in 27 different states, which are depicted in Figure 2B. The remaining infants were from Great Britain (N = 7), Canada (N = 5), Germany (N = 1), and France (N = 1). One infant did not have country information attached. One hundred six parents reported their highest level of education: Sixty-seven parents completed a professional or graduate degree, 29 completed a bachelor's degree, nine completed an associate's degree or community college, and one completed high school. Ninety-nine parents reported their combined annual family income, with 12 families below $50 k, 25 families between $50 and $100 k USD, 36 families between $100 k and $200 k, and 26 families above $200 k.

Procedure

The experiment was conducted via Lookit, with infants participating from home. Caregivers saw a video explaining the structure of the experiment, received instructions on how to position their child during the study (either in a high chair or in their lap), and were asked to pay attention to lighting conditions and keep distractions out of reach.

At the beginning of the experiment, parents received the instructions for parent control as described in 2.3. Like the Zoom and in-lab versions of the action efficiency paradigm (Liu & Spelke, 2017;

Figure 2
Design, Sample Geography and Parent Control



Note. (A) Experimental design of the Lookit VoE study (stimuli from Liu & Spelke, 2017), (B) geographic distribution of U.S. participants, and (C) parent control plot analogous to Figure 1C for Lookit VoE study. VoE = violation-of-expectation. See the online article for the color version of this figure.

Study 1, Chuey et al., 2021; Study 4), the experiment had two parts: familiarization and test. During familiarization, infants saw six trials of an agent jumping over a barrier. At test, infants saw four trials of test trials with the same agent jumping over a shorter barrier. In half the test trials, the agent took an unnecessarily high jump over the barrier (inefficient condition), and in the other half the agent took an appropriately low jump (efficient condition). The stimuli were identical to those shown to infants in the previous studies, and the design is shown Figure 2A. While in-lab stimuli were projected onto a wall in a dark room, Zoom and hands-off workflows had infants sit in high chairs or on their caregiver's lap, while watching the stimuli on a laptop or desktop computer. Age ranges were larger in the remote settings than in the lab (in-lab: 5.5–6.5 months, Zoom: 6–13 months, hands-off: 7–11.2 months).

Occasionally, a trial might repeat due to connectivity issues or parents pausing or restarting the trial. The trials preceding the final repeat were removed from analysis (N = 10, out of 528 test trials), to retain only the trials which were presented without interruptions.

To annotate the data, we first used iCatcher+ to obtain frame-by-frame annotations of the experimental videos. We manually supervised these annotations using the visualization tool of iCatcher+ which overlays the classification label (left, right, away, noface) onto the video. During manual supervision, we aimed to respect the spirit of automation: When we saw that iCatcher+ was making mistakes, we followed a two-step process: We first checked whether mistakes seemed to be caused by interference from other faces in the video (e.g., sometimes the parents' face would being coded). If so, we re-ran iCatcher+ on the same video but cropped out parts of the video that seemed to be the cause of issues, whenever possible. Then, either after cropping, or if cropping would not address the main issue, we spent no more than 5 min on each video for manually correcting iCatcher+ annotations. Videos that would take longer than 5 min to correct were excluded. After supervision, we merged the frame-by-frame annotations with trial boundaries and trial type information to obtain trial-wise looking times.

Figure 2C shows that in this study, parents still performed well when terminating trials. Average extra exposure was 4.6 s per trial, lower than in the study reported in Parent Control section, though early terminations occurred at a somewhat higher rate of 16.2%.

This experiment was designed to estimate the sensitivity of a novel method, not to test a hypothesis, and thus was not preregistered. The data set was collected and analyzed under institutional review board protocol 2003000107 "Social Knowledge in Infants and Children."

Results

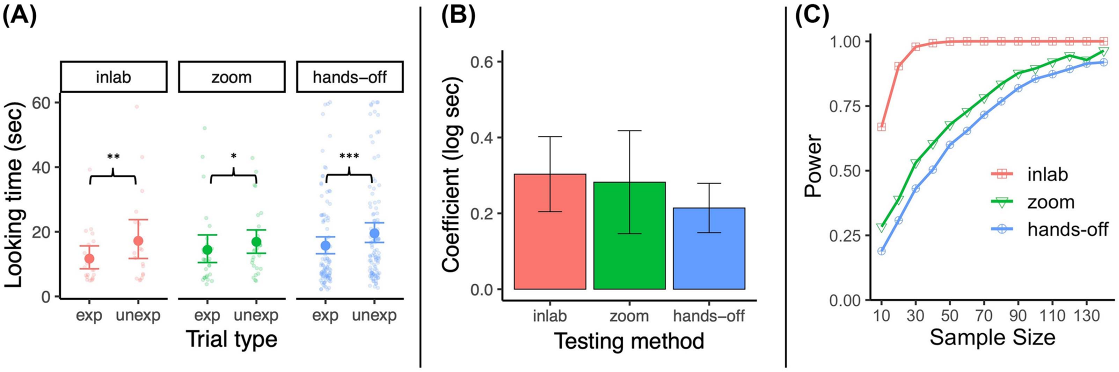
We first tested whether we could replicate the main effect of action efficiency using our hands-off workflow. Running the same analyses that we preregistered for the Zoom version of the study (https://osf.io/7vu23?view_only=6722cddbafc94bf4a1dac7e027360217), we found that infants looked significantly longer at inefficient test trials than efficient test trials, $\beta = 0.21$, $t(291.69) = 3.29$, $p = .001$, therefore replicating the main effect found in the lab and on Zoom (Chuey et al., 2021; Liu & Spelke, 2017). Figure 3A shows the raw and mean looking times in the in-lab, Zoom and hands-off experiments.

We also compared the effect size of the expectedness manipulation, as shown in Figure 3B. We again used the coefficient of the fixed effect of expectedness as the effect size metric. As expected, we found the largest coefficient in the in-lab study (β = 0.30, SE = 0.10), a medium coefficient in zoom study (β = 0.28, SE = 0.14), and the smallest coefficient in the hands-off study (β = 0.21, SE = 0.07; Supplemental Tables S1, S4–S5). The strength of the expectedness manipulation in the hands-off workflow was therefore 71% of the in-lab effect, and 76% of the synchronous Zoom version, in log seconds. To translate these results into seconds, for a case in which looking to the expected test trial was 20 s, the predictions for the unexpected test trial would be 27.09 s for the in-lab study, 26.52 s for the Zoom study and 24.78 s for the hands-off workflow, all else being equal. Using Cohen's D did not qualitatively change the results (Supplemental Figure S3).

Next, we analyzed the statistical power to detect a VoE effect in the different methods, using a bootstrapped power analysis. Simulating data sets of different sizes via bootstrapping, we compute

Figure 3

Raw Data, Effect Sizes and Statistical Power



Note. Testing methods compared in terms of (A) looking times to expected versus unexpected test trials by testing method (* p < .05. ** p < .01. *** p < .001), (B) coefficients on trial type in mixed effects model (error bars show standard errors) and (C) power to detect a significant main effect of test type (p < .05, one-tailed) as a function of sample size, obtained through a bootstrap simulation. See the online article for the color version of this figure.

the proportion of significant main effects of test type ($p < .05$; one-tailed) for each sample size. Figure 3C shows that while the ordering remains similar to the effect size comparison (in-lab > Zoom > hands-off), the in-lab study shows strong separation from the other two methods, crossing 80% power after 20 participants, whereas the Zoom and hands-off workflow do so after 80 and 90 participants, respectively.

In exploratory analyses, we further investigated the effect of testing method on other variables of interest. We compared the trajectories of looking times to the familiarization stimuli. While the in-lab procedure used a habituation paradigm (i.e., familiarizing until looking time drops to half), the other testing methods used a fixed number of six familiarization trials. Figure 4A shows looking times during familiarization across the testing methods. Initial interest was higher in-lab than in the other two testing methods, as shown by an analysis of variance, $F(2, 152) = 3.55$, $p = .031$, and post hoc tests showing that in-lab looking times ($M = 60.0$, $SD = 0.0$) to the first familiarization trial were higher than on Zoom ($M = 45.8$, $SD = 16.8$, $p = .027$), but not significantly higher than in the automated study ($M = 50.6$, $SD = 15.7$, $p = .066$). The Zoom and hands-off studies did not significantly differ from each other in initial interest ($p = .561$).

We also compared the frequency with which infants were excluded in the different testing methods, for any reason (fussiness, parental interference, experimenter errors or other technical issues). Figure 4B visualizes the proportion of exclusions by testing method. In the in-lab study, 26% of infants (seven of 27) were excluded, in the Zoom study only 7% of infants (two of 29) were excluded, and in the hands-off study 16% of infants (21 of 133) were excluded. A Fisher's exact test suggested that these proportions are not significantly different ($p = .168$). The proportion of exclusions due to fussiness specifically (in-lab: 11%, Zoom: 7%, hands-off: 6%) were also not significantly different ($p = .54$).

Finally, we investigated the effects of infant age and sex in the combined data set. We found no significant main effects of age and sex on looking times, or interactions with either testing method or test type (Supplemental Tables S7–S12). Thus, infant age (within the tested range of 6–13 months) and sex did not affect infant looking time in this paradigm.

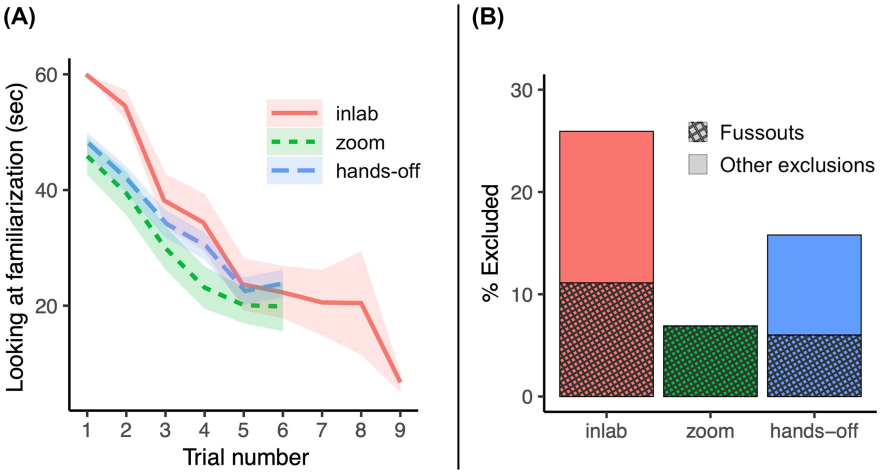
Discussion

In this article, we aimed to implement and evaluate a fully hands-off yet infant-contingent VoE experiment workflow. Critically, we replicated the robust main effect of expectedness on infant looking time, in the hands-off workflow: Infants look longer when an agent takes an inefficient (vs. an efficient) path to their goal (Gergely et al., 1995; Liu & Spelke, 2017). Comparing our results from the same VoE experiment run in-lab and via Zoom, we find that the hands-off workflow achieves only somewhat smaller effect sizes than alternative ways of testing infants online. Statistical power to detect a VoE effect was very similar to that on Zoom.

In our initial set of Zoom studies, we replicated three VoE main effects, increasing evidence for the robustness of these previously reported results (Baillargeon et al., 1985; Gergely et al., 1995; Woodward, 1998). A fourth paradigm, testing infants' looking time to solidity violations, did not replicate here. A previous study by Smith-Flores et al. (2022) conducted two studies testing support and solidity violation on Zoom and, like us, found longer looking toward support violations but not solidity violations. Multiple published, in-lab studies do find that infants look longer toward violations of solidity (Baillargeon et al., 1985; Perez & Feigenson, 2022; Spelke et al., 1992). It is possible that infants' beliefs about solidity are weaker for video displays than for real 3D objects, suggesting limitations on the aspects of infant cognition that can be measured in online studies with 2D stimuli.

Figure 4

Familiarization and Exclusions



Note. Other methodological comparisons: (A) trajectories of familiarization looking times (in-lab study employed a full infant-controlled habituation procedure, whereas Zoom and hands-off had a fixed number of six familiarization trials) and (B) percentage of infants who fussed out, or were excluded for other reasons (experimenter error, parental interference, technical errors, etc.). See the online article for the color version of this figure.

To move to an asynchronous setting, we trained parents to implement infant-contingent trial durations. We saw few early trial terminations, and some late terminations. Infants were not more likely to lose patience and withdraw from the experiment, and the total rate of exclusions was not higher in the automated pipeline. We conclude that for the VoE paradigm, it is possible to replace experimenter control with parent control.

Our findings promise a significant advance in the automation of infant data collection and annotation. The implementation of this hands-off workflow allows a drastic reduction in the marginal time and cost associated with data from each infant. Once an experiment is established, the process of data acquisition and annotation becomes markedly faster, allowing researchers to conduct studies with larger sample sizes. Larger sample sizes are essential for ensuring the reproducibility and robustness of research findings. Particularly in developmental studies, where behavioral readouts are often noisy and indirect and power is therefore low, small sample sizes inflate both false positive and false negative results (Button et al., 2013). Power analysis of the existing literature suggests that the sample size required for 80% power in VoE studies ($N = 80$–100) is substantially larger than current standard study samples ($N = 20$–30; Kunin et al., 2023). This meta-analytic estimate converges with our estimate for the Zoom and hands-off VoE study.

A recent approach to increase sample sizes in studies of infants has been the advent of multilab collaborative studies (Frank et al., 2017; ManyBabies Consortium, 2020), in which several laboratories run the same design and distribute the work associated with data collection and annotation across groups. Here, we offer a workflow in which single laboratories can conduct appropriately powered studies, avoiding the cost of coordination across laboratories, while reaping the benefits of more robust and reproducible results. The reduced geographical limitation inherent to asynchronous testing also offers potential for a diverse study population, similar to multisite studies. The main study here tested infants from 27 American states and four additional countries (Figure 2B).

The automation of infant looking time experiments could facilitate testing computational models of infant cognition. Models have been used to generate predictions for VoE experiments: A metric of expectedness can be derived from a computational model and then correlated with infants' looking time (Liu et al., 2017; Téglás et al., 2011). However, to disambiguate between multiple candidate models whose predictions differ only in fine-grained patterns of the data, small samples are insufficient. Tools for well-powered infant studies could thus enable more reliable testing grounds for these models.

Our work also has implications for the long-standing debates about VoE results themselves. Some scholars have argued that many VoE results do not replicate, or that the entire method is suspect (Blumberg & Adolph, 2023; Paulus, 2022). Here, we show that one previously published VoE result (Gergely et al., 1995) replicated three times: in the lab (Liu & Spelke, 2017), on Zoom, and using our hands-off workflow. These results show that (a) this specific paradigm is highly robust and replicable, and more generally and (b) the method of VoE in infants can generate highly replicable effects. Hands-off workflows will facilitate future replications, since the exact stimuli and experimental protocol can be transferred to other researchers for replication and extension, increasing the potential for a cumulative science of infant cognition.

The tools we introduce have some limitations. The hands-off workflow introduces various new sources of errors. Training parents to assist in delivering the experiment relies on compliance. How parents terminate trials introduces variability that a standardized experimenter-controlled design does not. Automated gaze coding via iCatcher+ has known failure modes in poor lighting conditions or unconventional poses (Erel et al., 2023). In the current experiment, 21 out of 135 videos could not be coded using iCatcher+ and so were dropped from the data set (though some of these videos may have been excluded even if they were manually coded, e.g., due to fussiness, which often co-occurs with poor automatic annotations). While our workflow enables collection of larger samples, these larger samples may also be necessary to overcome new sources of errors.

Furthermore, a degree of manual work is still necessary: Although gaze coding is automated, experimenters must still manually supervise the annotations and screen the videos for distractions and events that would invalidate trials, as well as major failures by iCatcher+. A trained experimenter can supervise iCatcher+ annotations in less than 5 min per infant. This is a significant improvement from manual annotation, which can take 4–5× the duration of the experimental video (in our case, about 45 min). Still, similarly to a hands-off workflow, supervision requires training experimenters on exclusion and annotation criteria.

The workflow was validated specifically for one VoE experiment, and we do not test the importance of specific features of this context. For example, there is an ongoing debate about the importance of controlling exposure in the familiarization period of VoE experiments and/or of fully habituating infants (Kucharský et al., 2022; Zaharieva et al., 2021). In the automated paradigm, all infants saw six trials, whereas in-lab infants watched up to nine trials until habituated (looking time reduced by half). Some meta-analytic evidence suggests that the familiarization procedure does not influence the size of observed VoE effects (Kunin et al., 2023). However, future studies could test whether the effect size in the hands-off workflow is more similar to the in-lab study when infants are allowed to habituate.

Furthermore, we do not measure the effectiveness of our approach for ages outside our 7–11 months range (though iCatcher+ was originally validated on a sample between 4 and 14 months, see Erel et al., 2023), or other experimental designs that measure gaze, such as preferential looking paradigms. Similarly, we do not directly address more complex child-contingent designs, such as those requiring verbal responses based on the child's behavior. In principle, there is potential for training parents to undertake more complex experimenter roles which could further expand the applicability of our method.

Given these limitations, using automation may not be appropriate for all studies of infant cognition. For each study, researchers should evaluate the tradeoff between benefits, like improved reproducibility and smaller marginal costs per infant, versus costs, like the smaller effect size and limitations on the paradigm and stimuli. The hands-off workflow also requires substantial technical investment, both for creating paradigms that can be delivered automatically and for using iCatcher+ for annotation. The total time investment to run a first study in the hands-off workflow is unlikely to be lower than a traditional paradigm; only the marginal investments per infant will be lower, especially initially.

Another concern is that the ease of data collection and annotation could lead to the exhaustion of the asynchronous study pool. Manual data collection and annotation put natural limits on how many participants are required for each study, but automating these processes could raise the demand for participants. If a hands-off workflow is adopted by many researchers, this move should be supplemented with efforts to recruit new participants to online data collection platforms like Lookit. In doing so, it will also be important to maintain and advance the representativeness of the Lookit pool (Nielsen et al., 2017; Scott et al., 2017; Scott & Schulz, 2017).

In summary, we present an advance in the automation of infant data collection and annotation in VoE experiments. We tested a hands-off infant-contingent violation-of-expectation (VoE) workflow, combining classic VoE experiments with new tools like iCatcher+ for automated annotation, parent control for infant-contingency and Lookit for asynchronous testing. The results demonstrate a robust main effect of expectedness on infant looking time, achieving effect sizes close to traditional in-lab and Zoom-based methods. Automation greatly reduces the marginal time and cost per infant of data collection, enabling larger sample sizes, which are crucial for the reproducibility and robustness of developmental studies.

References

Amir, D., & McAuliffe, K. (2020). Cross-cultural, developmental psychology: Integrating approaches and key insights. *Evolution and Human Behavior*, *41*(5), 430–444. https://doi.org/10.1016/j.evolhumbehav.2020.06.006

Aslin, R. N. (2007). What's in a look? *Developmental Science*, *10*(1), 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208. https://doi.org/10.1016/0010-0277(85)90008-3

Baltrusaitis, T., Zadeh, A., Lim, Y. C., & Morency, L. P. (2018, May). *Openface 2.0: Facial behavior analysis toolkit* [Conference session]. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). https://doi.org/10.1109/FG.2018.00019

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development*, *89*(6), 1996–2009. https://doi.org/10.1111/cdev.13079

Blumberg, M. S., & Adolph, K. E. (2023). Protracted development of motor cortex constrains rich interpretations of infant cognition. *Trends in Cognitive Sciences*, *27*(3), 233–245. https://doi.org/10.1016/j.tics.2022.12.014

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2022). Six solutions for more reliable infant research. *Infant and Child Development*, *31*(5), e2296. https://doi.org/10.1002/icd.2296

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

Datavyu Team. (2014). *Datavyu: A video coding tool*. New York University. Databrary Project. https://datavyu.org

Erel, Y., Shannon, K. A., Chu, J., Scott, K., Struhl, M. K., 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. (2023). iCatcher+: Robust and automated annotation of infants' and young children's gaze behavior from videos collected in laboratory, field, and online studies. *Advances in Methods and Practices in Psychological Science*, *6*(2). https://doi.org/10.1177/25152459221147250

Fischer, T., Chang, H. J., & Demiris, Y. (2018). *Rt-Gene: Real-Time eye gaze estimation in natural environments* [Conference session]. Proceedings of the European Conference on Computer Vision (ECCV). https://doi.org/10.1007/978-3-030-01249-6_21

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

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193. https://doi.org/10.1016/0010-0277(95)00661-H

Horowitz, F. D., Paden, L., Bhana, K., & Self, P. (1972). An infant-control procedure for studying infant visual fixations. *Developmental Psychology*, *7*(1), Article 90. https://doi.org/10.1037/h0032855

Kominsky, J. F. (2019). PyHab: Open-source real time infant gaze coding and stimulus presentation software. *Infant Behavior and Development*, *54*, 114–119. https://doi.org/10.1016/j.infbeh.2018.11.006

Kucharský, Š., Zaharieva, M., Raijmakers, M., & Visser, I. (2022). Habituation, part II. Rethinking the habituation paradigm. *Infant and Child Development*, *33*(1), Article 2383. https://doi.org/10.1002/icd.2383

Kunin, L., Piccolo, S., Saxe, R., & Liu, S. (2023). *Conceptual and perceptual novelty reflect distinct motives of infant looking: Meta-analytic evidence*. PsyArxiV. https://doi.org/10.31234/osf.io/kx76y

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, Article 863. https://doi.org/10.3389/fpsyg.2013.00863

Lapidow, E., Tandon, T., Goddu, M., & Walker, C. M. (2021). A tale of three platforms: Investigating preschoolers' second-order inferences using in-person, Zoom, and Lookit methodologies. *Frontiers in Psychology*, *12*, Article 731404. https://doi.org/10.3389/fpsyg.2021.731404

Liu, S., & Spelke, E. S. (2017). Six-month-old infants expect agents to minimize the cost of their actions. *Cognition*, *160*, 35–42. https://doi.org/10.1016/j.cognition.2016.12.007

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041. https://doi.org/10.1126/science.aag2132

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science*, *3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Needham, A., & Baillargeon, R. (1993). Intuitions about support in 4.5-month-old infants. *Cognition*, *47*(2), 121–148. https://doi.org/10.1016/0010-0277(93)90002-D

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy*, *22*(4), 436–469. https://doi.org/10.1111/infa.12186

Papoutsaki, A. (2015). *Scalable webcam eye tracking by learning from user interactions* [Conference session]. Proceedings of the 33rd Annual ACM

Conference Extended Abstracts on Human Factors in Computing Systems.

Papoutsaki, A., Gokaslan, A., Tompkin, J., He, Y., & Huang, J. (2018). *The eye of the typer: A benchmark and analysis of gaze behavior during typing* [Conference session]. 2018 ACM Symposium on Eye Tracking Research & Applications. https://doi.org/10.1145/3204493.3204552

Paulus, M. (2022). Should infant psychology rely on the violation-of-expectation method? Not anymore. *Infant and Child Development*, *31*(1), Article e2306. https://doi.org/10.1002/icd.2306

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Perez, J., & Feigenson, L. (2022). Violations of expectation trigger infants to search for explanations. *Cognition*, *218*, Article 104942. https://doi.org/10.1016/j.cognition.2021.104942

Raz, G. (2024). *An asynchronous, hands-off workflow for looking time experiments with infants*. https://osf.io/ndkt6/

Scott, K., Chu, J., & Schulz, L. (2017). Lookit (Part 2): Assessing the viability of online developmental research, results from three case studies. *Open Mind: Discoveries in Cognitive Science*, *1*(1), 15–29. https://doi.org/10.1162/OPMI_a_00001

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

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

Smith-Flores, A. S., Perez, J., Zhang, M. H., & Feigenson, L. (2022). Online measures of looking and learning in infancy. *Infancy*, *27*(1), 4–24. https://doi.org/10.1111/infa.12435

Spelke, E. S. (2022). *What babies know: Core knowledge and composition* (Vol. 1). Oxford University Press. https://doi.org/10.1093/oso/9780190618247.001.0001

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605–632. https://doi.org/10.1037/0033-295X.99.4.605

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Steffan, A., Zimmer, L., Arias-Trejo, N., Bohn, M., Dal Ben, R., Flores-Coronado, M. A., Franchin, L., Garbisch, I., Grosse Wiesmann, C., Hamlin, K., Havron, N., Hay, J., Hermansen, T. K., Jakobsen, K., 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*. PsyArXiv. https://doi.org/10.31234/osf.io/7924h

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science*, *332*(6033), 1054–1059. https://doi.org/10.1126/science.1196404

Werchan, D. M., Thomason, M. E., & Brito, N. H. (2023). OWLET: An automated, open-source method for infant gaze tracking using smartphone and webcam recordings. *Behavior Research Methods*, *55*, 3149–3163. https://doi.org/10.3758/s13428-022-01962-w

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

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

Zaharieva, M., Kucharský, Š., Colonnesi, C., Gu, T., Jo, S., Luttenbacher, I., Mannsdörfer, L., Matetovici, M., Mickute, U., Raijmakers, M., Staaks, J., Torma, Z., & Visser, I. (2021). *Design choices in the infant Habituation Paradigm: A pre-registered crowd-sourced systematic review and meta-analysis*. PsyArxiV. https://doi.org/10.31234/osf.io/bdtx9

Received January 17, 2024
Revision received April 22, 2024
Accepted April 22, 2024 ■