OPEN MIND

Discoveries in Cognitive Science

an open access journal

Lookit (Part 1): A New Online Platform for Developmental Research

Kimberly Scott[1] **and Laura Schulz**[1]

[1]Massachusetts Institute of Technology

Keywords: cognitive development, research methods, Internet, looking time, preferential looking

ABSTRACT

Many important questions about children's early abilities and learning mechanisms remain unanswered not because of their inherent scientific difficulty but because of practical challenges: recruiting an adequate number of children, reaching special populations, or scheduling repeated sessions. Additionally, small participant pools create barriers to replication while differing laboratory environments make it difficult to share protocols with precision, limiting the reproducibility of developmental research. Here we introduce a new platform, "Lookit," that addresses these constraints by allowing families to participate in behavioral studies online via webcam. We show that this platform can be used to test infants (11–18 months), toddlers (24–36 months), and preschoolers (36–60 months) and reliably code looking time, preferential looking, and verbal responses, respectively; empirical results of these studies are presented in Scott, Chu, and Schulz (2017). In contrast to most laboratory-based studies, participants were roughly representative of the American population with regards to income, race, and parental education. We discuss broad technical and methodological aspects of the platform, its strengths and limitations, recommendations for researchers interested in conducting developmental studies online, and issues that remain before online testing can fulfill its promise.

Citation: 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. doi:10.1162/opmi_a_00002

DOI:
http://doi.org/10.1162/opmi_a_00002

Supplemental Materials:
http://dx.doi.org/10.7910/DVN/TMOQMC

Received: 1 March 2016
Accepted: 7 November 2016

Competing Interests: The authors declare that they have no competing interests.

Corresponding Author:
Kimberly Scott
kimscott@mit.edu

The MIT Press

Behavioral research with infants and children stands to illuminate the roots of human cognition. However, many important questions about cognitive development remain unasked and unanswered due to the practical demands of recruiting participants and bringing them into the lab. Such demands limit participation by families from diverse cultural, linguistic, and economic backgrounds; deter scientists from studies involving large sample sizes, narrow age ranges, and repeated measures; and restrict the kinds of questions researchers can answer. It is hard to know, for instance, whether an ability is present in all or only most children, or whether an ability is absent or weakly present. Such small distinctions can have large theoretical and practical implications. Fulfilling the promise of the field depends on scientists' ability to measure the size and stability of effects in diverse populations.

In adult psychology, online testing through Amazon Mechanical Turk (AMT) has begun to lower barriers to research, enabling scientists to quickly collect large datasets from diverse participants (Buhrmester, Kwang, & Gosling, 2011; Paolacci, Chandler, & Ipeirotis, 2010; Rand, 2012; Shapiro, Chandler, & Mueller, 2013). As the technical hurdles involved in online testing dwindle, we are poised to expand the scope of questions developmental science can address as well. Online testing can allow access to more representative populations,

children from particular language groups or affected by specific developmental disorders, and information about children's behavior in the home. Access to larger sample sizes will also allow researchers to estimate effect sizes with greater precision, detect small or graded effects, and generate sufficient data to test computational models. The motivation to bring studies online is bolstered by growing awareness of the importance of direct replication and reproducible results (Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012).

However, before the potential of online developmental research can be fully realized, we need a secure, robust platform that can translate developmental methods to a computer-based home testing environment. Here we present a new online developmental research platform, Lookit. Parents access Lookit through their web browsers, participate at their convenience in self-administered studies with their child, and transmit the data collected by their webcam for analysis. To follow, we address broad ethical, technological, and methodological issues related to online testing, describe the demographics of our online participant population, and offer recommendations for researchers seeking to adapt studies to an online platform. For an empirical report of our user case studies, including raw data and analysis code, please see Scott, Chu, and Schulz (2017). For technical and methodological details regarding the platform itself and video coding procedures, please see the Supplemental Materials (Scott & Schulz, 2017).

ETHICAL ISSUES

Testing children online raises a number of ethical concerns specific to the online environment, including providing fair and noncoercive reimbursement, ensuring the validity of informed consent, and protecting parents' privacy. We addressed these issues through the recruitment and registration procedures described below.

Recruitment, Reimbursement, and Registration

Participants were recruited via AMT and linked to the Lookit site (https://lookit.mit.edu). Participants were paid three to five dollars (depending on the study) for participation to ensure payment of at least the minimum wage nationally, even in cases where parents encountered technical difficulties and contacted the lab. This policy is in accordance with guidelines for researchers regarding fair payment (http://guidelines.wearedynamo.org/). To ensure that parents did not feel any pressure to complete a study, especially if their child was unwilling to continue, parents were paid if they initiated a study, regardless of completion and of any issues in compliance or implementation. Participation required creating a user account and registering at least one child. As in the lab, parents provided their child's date of birth to determine study eligibility. A demographic survey was available for parents to fill out at any point.

Consent and Privacy

At the start of each study, a consent form and the webcam video stream was displayed. The parent was instructed to record a brief verbal statement of consent (see Supplemental Materials for examples), ensuring that parents understood they were being videotaped. Parents were free to end the study at any point. After completing a study, the parent selected a privacy level for the video collected. Across our three test studies, 31% of sessions were marked "private" (video can be viewed only by our research team), 41% "scientific" (video can be shared with other researchers for scientific purposes), and 28% "free" (video can be shared for publicity or educational purposes). Parents also had the option to withdraw their data from the study at this point; this option was chosen by less than one percent of participants and was treated as invalid

consent. A coder checked each consent video before looking at any other video associated with the session. Valid consent was absent for 16% of participants overall (N = 961) due to technical failures in the video or audio transmission, parents not reading the statement, or subsequent withdrawal from the study.

TECHNICAL ISSUES

Video Quality

Multiple short clips were recorded in each study during periods of interest. Video quality varied due to participants' upload speed. Our primary concern for coding looking measures was the effective framerate of the video. Because the putative framerate of the video was unreliable due to details of the streaming procedure, we estimated an effective framerate based on the number of "changed frames" in each clip. A "changed frame" differed from the previous changed frame in at least 20% of pixels; note that this underestimates higher framerates since frames close in time without major movement may actually differ by fewer than 20% of pixels. Videos with an effective framerate under 2 frames per second (fps) were excluded as unusable for looking studies (see SI for examples of video at various effective framerates). The median effective framerate across sessions with any video was 5.6 fps (interquartile range = 2.9–8.6 fps).

Starting webcam video introduced a small (generally less than 1 s) variable delay in the start of the recorded clips. In the preferential looking study, where spoken questions accompanied test videos, onset of audio playback was used to more precisely determine when test periods began. Future versions of this platform will avoid this delay and improve framerates.

Video Usability

Before completing study-specific coding, one coder checked that video from the study was potentially usable. Video was unusable 35% of the time (282 of 805 unique participants with valid consent records). The most common reasons for unusable video were absence of any study videos (44% of records with unusable video), an incomplete set of study videos (20%), and insufficient framerate (15%). Rarely, videos were unusable because a child was present but generally outside the frame (3%) or there was no child present (1%).

METHODOLOGICAL ISSUES

To test the feasibility of online developmental research across a variety of methods and age groups, we conducted three studies: a looking-time study with infants (11–18 months) based on Téglás, Girotto, Gonzalez, and Bonatti (2007), a preferential looking time study with toddlers (24–36 months) based on Yuan and Fisher (2009), and a forced choice study with preschoolers (ages 3 and 4) based on Pasquini, Corriveau, Koenig, and Harris (2007). These allowed us to assess how online testing affected coding and reliability, children's attentiveness, and parental interference. For details on the specific studies, see Scott et al. (2017).

Coding and Reliability of Looking Time Measures

Each session of the looking time study was coded using VCode (Hagedorn, Hailpern, & Karahalios, 2008) by two coders blind to condition. Looking time for each of eight trials per session was computed based on the time from the first look to the screen until the start of the first continuous one-second lookaway, or until the end of the trial if no valid lookaway occurred. Differences of 1 s or greater, and differences in whether a valid lookaway was detected, were flagged and those trials recoded. Agreement between coders was excellent; coders agreed on

whether children were looking at the screen on average 94.6% of the time (N = 63 children; SD = 5.6%). The mean absolute difference in looking time computed by two coders was 0.77 s (SD = 0.94 s).

Measuring until the first continuous lookaway of a given duration introduces a thresholding effect in addition to the small amount of noise induced by a reduced framerate. The magnitude of this effect depends on the dynamics of children's looks to and away from the screen. We examined a sample of 1,796 looking times, measured until the first one-second lookaway, from 252 children (M = 13.9 months, SD = 2.6 months) tested in our lab with video recorded at 30 Hz. Reassuringly, in 68% of measurements, the lookaway that ended the measurement was over 1.5 s. We also simulated coding of these videos at framerates ranging from 0.5 to 30 Hz; the median absolute difference between looking times calculated from our minimum required framerate of 2 Hz vs. original video was only .16 s (interquartile range = 0.07–0.29 s; see Figure S1 (Scott, Chu, and Schulz, 2017)).

Coding and Reliability of Preferential Looking Measures

Each session of the preferential looking study was coded using VCode (Hagedorn et al., 2008) by two coders blind to the placement of test videos. Looks to the left and right are generally clear; for examples, see Figure 1. Three calibration trials were included in which an animated attention getter was shown on one side and then the other. During calibration videos, all

Figure 1. Cropped examples of webcam video frames of children looking to the reader's left (left column) and right (right column).

138 participants coded looked on average more to the side with the attention getter. For each of nine preferential looking trials, we computed fractional right/left looking times (the fraction of total looking time spent looking to the right/left). Substantial differences (fractional looking time difference greater than .15, when that difference constituted at least 500 ms) were flagged and those clips recoded. A disagreement score was defined as the average of the coders' absolute disagreement in fractional left looking time and fractional right looking time, as a fraction of trial length. The mean disagreement score across the 138 coded participants was 4.44% (*SD* = 2.00%, range 1.75–13.44%).

Coding Child Attentiveness and Parental Interference

Two natural concerns about online testing are that the home environment might be more distracting than the laboratory or that parents might be more likely to interfere with study protocols. In laboratory-based developmental studies, 14% of infants and children on average are excluded due to fussiness, while only 2% of studies give an operational definition of fussiness (Slaughter & Suddendorf, 2007). Looking times from crying children are unlikely to be meaningful, but subjective exclusion criteria reduce the generalizability of results. Similar issues arise with operationalizing parental interference.

To address these issues, we established criteria for fussiness (defined as crying or attempting to leave the parent's lap), distraction (whether any lookaway was caused by an external event), and various parental actions, including peeking at the video during trials where their eyes should be closed. (See Table S1 and Coding Manual for details.) Exclusion criteria were then based on the number of clips where there was parental interference or where the child was determined to be fussy or distracted. In the two studies using looking measures, two blind coders recorded which actions occurred during individual clips. The first author arbitrated

Table 1. Intercoder reliability for qualitative coding of clips from the looking time study (Study 1, shaded rows, N = 112) and preferential looking study (Study 2, white rows, N = 138) of Scott et al. (2017).

	Agreement	Cohen's κ	Frequency
Fussy (crying or trying to get out of parent's lap)	.85	.55	13.3%
	.99	.42	0.9%
Actively distracted (lookaway caused by external event)	.85	.40	10.4%
	.97	.37	1.8%
Parent talks to child	.98	.16	1.8%
	.95	.70	8.7%
Parent's eyes not visible	.97	.35	4.0%
	.99	.91	7.4%
Parent's eyes open	.98	.67	4.2%
	.97	.70	5.4%
Parent peeks (after start of clip)	.90	.36	9.2%
	.90	.36	11.0%
Parent's eyes open briefly at very start of clip	.77	.49	38.4%
	.93	.68	12.8%

Note. In Study 1, fussiness and distraction were coded for all 8 clips per participant, and parent interaction measures were coded for the last 6 clips; clips were about 20 s long. In Study 2, fussiness and distraction were coded for all 13 clips per participant (ranging from 8 to 50 s), and parent interaction measures for 6 clips (about 8 s each).

disagreements. This constitutes one of the first direct studies of intercoder agreement on these measures. Coders agreed on fussiness and distraction in at least 85% of clips, with Cohen's kappa ranging from .37 to .55 (see Table 1).

Parents were largely compliant with testing protocols, including requests that they refrain from talking or close their eyes during parts of the studies to avoid inadvertently biasing the child. However, compliance was far from perfect: 9% of parents had their eyes mostly open on at least one of 3–4 test trials, and 26% briefly peeked on at least one test trial (see Table S2). In the forced-choice study with preschoolers, parents interfered in 8% of trials by repeating the two options or answering the question themselves before the child's final answer, generally in cases where the child was reluctant to answer. Practice trials before the test trials mitigated data loss due to parent interference; additional recommendations based on our experience are covered in the Discussion and Recommendations section.

Counterbalancing

Condition assignment and counterbalancing were initially achieved simply by assigning each participant to whichever condition had the fewest sessions already in the database. Because many sessions could not be included in analysis, we later manually updated lists of conditions needed to achieve more even counterbalancing and condition assignment. Condition assignment was still not as balanced as in the lab, so we used analysis techniques robust to this variation. Future versions of the platform will allow researchers to continually update the number of included children per condition as coding proceeds.

DEMOGRAPHICS

We collected demographic information from participants to see whether the promise of expanded participation in developmental research was fulfilled. Families participating on Lookit were more representative of the U.S. population than typical lab samples on several measures. Fifty percent of participants came from families with a yearly income under $50,000 and 73% from families with a yearly income under $75,000 ($N$ = 552 responding out of 759 unique participants in their study's age range and with a valid consent video). Our participants report 21 distinct languages spoken in the home in addition to English; 9% (N = 571 responding)

Table 2. Percentage of Lookit demographic survey respondents (N = 563) and U.S. census respondents (U.S. Census Bureau, 2014b) by race and Hispanic origin.

	Lookit	U.S. Population
Hispanic	9.2	17.4
White	80.0	77.4
Black	7.1	13.2
Asian	1.6	5.4
American Indian, Alaska Native, Native Hawaiian, or Pacific Islander	0.4	1.4
Two or more races	10.9	2.5

Note. Lookit participants were asked a single free-response question about the race of their child whereas census respondents were asked separately about Hispanic origin and race. For direct comparison with the U.S. census, we included only those who additionally specified a race (N = 550) in computing the remaining percentages.

are multilingual. Participants' races were roughly representative of the American population; Table 2 summarizes the racial distribution of participants compared to recent census data.

Even where research samples are racially or economically diverse, participation in research studies is often skewed toward parents with higher education levels. We estimated the expected distribution of educational attainment by weighting American census data based on the distribution of Lookit parents' age ranges and genders. Lookit parents had education levels much more representative of the American population than, for instance, a sample of 96 parents in a recent study conducted by our lab at the Boston Children's Museum (an institution specifically committed to affordability and diversity; see Figure 2). Nonetheless, additional outreach is likely necessary to reach parents who have not completed high school and to obtain a truly nationally representative sample.

DISCUSSION AND RECOMMENDATIONS

Our case studies confirmed the viability of Lookit as a method for remote data collection in developmental psychology in several important respects. Most importantly, the platform worked, for both parents and researchers. Parents were able to log into the system, select studies, administer the experiments themselves, and upload their child's data. Researchers were able to host multiple studies on the site, control timing and counterbalancing of stimuli, assign participants to conditions, limit the age range for each study, receive transmitted video, monitor consent, and code dependent measures, including preferential looking, looking time, and verbal response measures across ages from 11 months through 4 years.

The coding results show that preferential looking and looking time measures can be collected from streamed webcam video, without extensive instruction to parents about positioning. Despite varying webcam placement and video resolution, mean disagreement between blind coders on looking time was less than 1 s, with agreement 95% of the time, typical for offline coding in labs (90–95% agreement reported by Baillargeon, Spelke, & Wasserman, 1985; Feigenson, Carey, & Spelke, 2002; Onishi & Baillargeon, 2005; Starkey, Spelke, &

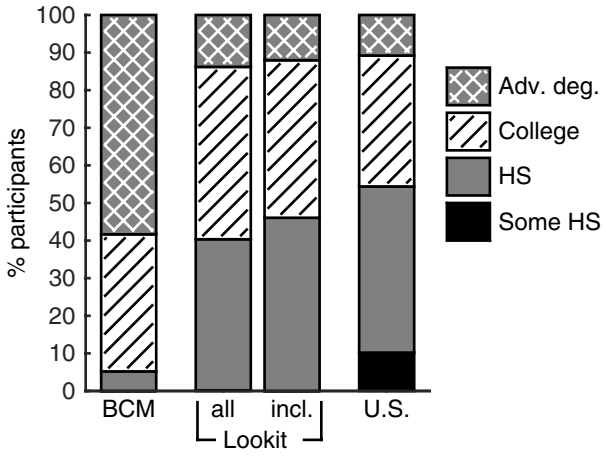

Figure 2. Distribution of the highest level of education completed.
Shown by (i) parents participating in a recently conducted study with their child at the Boston Children's Museum (BCM), $N = 96$; (ii) Parents of Lookit participants ($N = 559$) and included Lookit participants ($N = 184$); and (iii) American adults, based on U.S. census data and weighted by the age and gender distribution of Lookit participants (U.S. Census Bureau, 2014a).

Gelman, 1990; Xu & Garcia, 2008; Xu, Spelke, & Goddard, 2005). Mean disagreement on time spent looking to the left/right of the screen was less than 5% of trial length, also in line with lab estimates, although agreement is typically reported regarding whether the child was looking left, right, or away before summing those intervals (e.g., 91% by Smith & Yu, 2008; 98% by Yuan & Fisher, 2009).

The studies also mitigate concerns about parent-administered testing, although several modifications could improve compliance. The request that posed the biggest challenge was that parents close their eyes. For example, several parents reported peeking due to concerns about what was being shown to their child. If asking parents not to watch what their child is shown, we recommend that researchers provide a clear, simple explanation of why this is necessary and, if possible, allow parents to view stimuli in advance. To address more general difficulties with parent blinding, we recommend including practice trials and asking parents to close their eyes several seconds before blinding is necessary (for instance, so that they can first ensure a video is playing). We also recommend that if possible the parent be asked to face away from the computer with the child looking over their shoulder (to make peeking less tempting and easier to detect). We emphasize, though, that even with minimal instructions, parents generally closed their eyes when asked and failures to comply were readily detectable.

One striking difference between Lookit and traditional lab studies was that our overall yield was quite low at 26%: Of 997 nonrepeat sessions, only 255 were included in the final analysis across these three test studies. Data loss was primarily due to factors unique to the online testing environment (failure to provide informed consent, technical failure of video recording, and parents who left the study early). Because these were apparent early in the coding process, they did not create an undue coding burden. Exclusion rates due to infant behavior were similar to those in the lab, although higher than the stable average of 22% (range 0–87%) for violation-of-expectation paradigms reported by Slaughter and Suddendorf (2007). In the looking time study, 43 of 112 valid video submissions were excluded due to the child's behavior and 20 for parent interference or technical criteria that were not relevant in the original study. The effective exclusion rate was 47%, compared to 50% excluded in Téglás et al. (2007). However, overall exclusion rates due to child and parent behavior were generally higher than in the original studies. In the preferential looking study, we excluded 36% of children due to small differences in looking preferences when asked to find familiar verbs on opposite sides of the screen, in addition to 10% excluded due to parent interference and 5% due to low attention. In contrast, Yuan and Fisher (2009) excluded only 10% due to side bias, distraction, poor practice trial performance, or outlier preference at test, and did not report any parent interference. Finally, in the forced-choice study we excluded 20% of children due to incorrect naming of familiar objects, and 17% due to insufficient valid answers to test questions (see Supplemental Materials of Scott et al., 2017, for details). In contrast, Pasquini et al. (2007) excluded only 6% of children and only due to inaccuracy on questions about familiar object names. However, having conducted studies both online and in the laboratory, we believe the low yield online is unlikely to offset the in-lab costs of outreach, recruitment, and scheduling: it only takes around 2 minutes total to code consent, check video usability, and process an AMT submission, in addition to about 20 minutes per coder to code a complete session (which would be necessary for most in-lab studies using looking measures as well). In contrast, recruiting, scheduling, and testing one child in the lab generally takes around an hour, exclusive of coding. Further technical and user-experience optimization will decrease the rates of invalid consent videos and video failure in online studies.

Moving protocols from the lab to the web browser will require continued methodological refinement. Encouragingly, we found looking times similar to those reported in the lab in the looking time study and excellent attention (over 80% looking) to the dialogues in the preferential looking study. However, despite children's overall attentiveness, we recommend that researchers choose methods as robust as possible to minor interruptions. For this reason, we suggest using preferential looking rather than looking time paradigms if possible. In preferential looking, a distracted lookaway (e.g., to the family dog) simply decreases the measurement period; in a looking time paradigm it ruins an entire measurement. Preferential looking based on audio prompts may need to be optimized for online presentation to induce more reliable responses, as we observed wide variation in 2-year-olds' looking to familiar verbs. We suggest that labs seeking to use verbal response measures design studies to be robust to nonresponses (e.g., by using many short questions and pooling responses); provide guidance for parents in prompting their children to respond, including example videos or practice trials; and encourage engagement by allowing the online interface to respond contingently (e.g., by repeating back an answer the child chose, as selected by the parent). As we move studies online, it will also be crucial to clearly define behavioral criteria for exclusion of trials or participants in order to harness Lookit's full potential for replicable results.

Another striking difference between Lookit and traditional lab studies was the diversity of incomes, parental education levels, races, and language backgrounds represented among the participants. Inclusion criteria, both technical and behavioral, did not disproportionately affect lower socioeconomic status (SES) families in any of the studies; nor was performance linked to SES where there were clear normative choices (see Scott et al., 2017, for details). In the context of the current studies, we believe the absence of any relationship between performance and SES is encouraging with respect to the accessibility of the platform. However, the diversity of the participant pool suggests that for those studies where SES is a critical variable, online testing may be an appropriate interface for assessing its impact.

Nonetheless, there are limitations to online testing. Any dependence on the child's behavior must be implemented via the parent, for instance, by allowing the parent to pause a study or repeat a question; infant-contingent displays are not yet possible. The experimenter cannot directly engage in joint attention or pedagogical cueing or adjust fluidly for momentary distractions. For studies where synchronous attention or direct pedagogical engagement is critical, or its effects are the question of interest (e.g., Csibra & Gergely, 2009; Yu & Smith, 2012), or where visual angle subtended by stimuli must be tightly controlled, Lookit will not be an appropriate interface.

Even within the scope of methods adaptable to the online environment, several issues must be addressed before online testing achieves its full potential. First, Lookit is currently a prototype and does not yet include a "plug and play" interface. Programming expertise beyond what is expected in most graduate developmental programs was required to implement studies on the platform. In collaboration with the Center for Open Science, we are working toward an easy-to-use experimenter interface. Second, AMT is not especially designed for parents, so recruitment is not as efficient as it might be. Lookit may become a more effective tool as it connects with sites that directly target parents. At that stage, however, maintaining parents' interest will require a steady supply of novel content from developmental labs. Thus, expanding interest in the site from both researchers and parents should be mutually reinforcing.

We look forward to creative uses of this method. Although it will not be appropriate for every study, online research can expand access to both more representative populations

and rare populations, and make it easier to conduct large-scale longitudinal studies, detect small and graded effects, generate data sufficient for testing computational models, and assess individual differences and developmental change. We hope this tool will also be used to replicate classic effects and make the replication of new results easier. Finally, in connecting families and scientists, Lookit offers exciting new opportunities for education and outreach. As a venue for "citizen science" as well as scientific research, our goal for Lookit is to expand the scope of both the questions we ask and the people we reach.

ACKNOWLEDGMENTS

We thank all of the families who participated in this research and the Boston Children's Museum where we conducted early piloting of computer-based studies. Thanks to Joseph Alvarez, Daniela Carrasco, Jean Chow, DingRan (Annie) Dai, Hope Fuller-Becker, Kathryn Hanling, Nia Jin, Rianna Shah, Shirin Shivaei, Tracy Sorto, Yuzhou (Vivienne) Wang, and Jean Yu for help with data collection and coding. Special thanks to Paul Muentener for demographic data from a study at the Children's Museum; lab managers Rachel Magid and Samantha Floyd for logistical support; and Elizabeth Spelke, Joshua Tenenbaum, and Rebecca Saxe for helpful discussions. This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1429216, NSF Graduate Research Fellowship under Grant No. 1122374, and by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

AUTHOR CONTRIBUTIONS

KS developed the methodology and designed the studies with the advice of LS. Data collection, analysis, and interpretation were performed by KS. KS and LS prepared the manuscript.

REFERENCES

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

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. doi.org/10.1177/1745691610393980

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. doi.org/10.1016/j.tics.2009.01.005

Feigenson, L., Carey, S., & Spelke, E. (2002). Infants' discrimination of number vs. continuous extent. *Cognitive Psychology*, *44*, 33–66. doi.org/10.1006/cogp.2001.0760

Hagedorn, J., Hailpern, J., & Karahalios, K. (2008). VCode and VData: Illustrating a new framework for supporting the video annotation workflow. *Proceedings of the Workshop on Advanced Visual Interfaces AVI*, *2008*, 317–321. doi.acm.org/10.1145/1385569.1385622

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*(5719), 255–258. doi.org/10.1126/science.1107621

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 943. doi.org/10.1126/science.aac4716

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. doi.org/10.1177/1745691612465253

Pasquini, E. S., Corriveau, K. H., Koenig, M., & Harris, P. L. (2007). Preschoolers monitor the relative accuracy of informants. *Developmental Psychobiology*, *43*(5), 1216–1226.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology*, *299*, 172–179. doi.org/10.1016/j.jtbi.2011.03.004

Scott, K. M., Chu, J., & Schulz, L. E. (2017). *Validation of online collection of looking time, preferential looking, and verbal response measures*. Manuscript submitted for publication.

Scott, K. M., & Schulz, L. E. (2017). Replication data for: Lookit: A new online platform for developmental research. doi.org/10.7910/DVN/TMOQMC, Harvard Dataverse, V1.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220. doi.org/10.1077/2167702612469015

Slaughter, V., & Suddendorf, T. (2007). Participant loss due to "fussiness" in infant visual paradigms: A review of the last 20 years. *Infant Behavior and Development*, *30*, 505–514. doi.org/10.1016/j.infbeh.2006.12.006

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. doi.org/10.1016/j.cognition.2007.06.010

Starkey, P., Spelke, E. S., & Gelman, R. (1990). Numerical abstraction by human infants. *Cognition*, *36*, 97–127. doi.org/10.1016/0010-0277(90)90001-Z

Téglás, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(48), 19156–19159. doi.org/10.1073/pnas.0700271104

U.S. Census Bureau. (2014a). Educational attainment of the population 18 years and over, by age, sex, race, and Hispanic origin: 2014. Retrieved from www.census.gov/hhes/socdemo/education/data/cps/2014/Table1-01.xlsx

U.S. Census Bureau. (2014b). QuickFacts. Retrieved from www.census.gov/quickfacts/table/PST045215/00

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(13), 5012-5015. doi.org/10.1073/pnas.0704450105

Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, *8*(1), 88–101. doi.org/10.1111/j.1467-7687.2005.00395.x

Yu, C., & Smith, L. B. (2012). Embodied attention and word learning by toddlers. *Cognition*, *125*(2), 244-262. doi.org/10.1016/j.cognition.2012.06.016

Yuan, S., & Fisher, C. (2009). "'Really? She blicked the baby?"' Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science*, *20*(5), 619–626. doi.org/10.1111/j.1467-9280.2009.02341.x