OPEN MIND
Discoveries in Cognitive Science

an open access journal

Lookit (Part 2): Assessing the Viability of Online Developmental Research, Results From Three Case Studies

Kimberly Scott[1], **Junyi Chu**[1], **and Laura Schulz**[1]

[1] Massachusetts Institute of Technology

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

ABSTRACT

To help address the participant bottleneck in developmental research, we developed a new platform called "Lookit," introduced in an accompanying article (Scott & Schulz, 2016), that allows families to participate in behavioral studies online via webcam. To evaluate the viability of the platform, we administered online versions of three previously published studies involving different age groups, methods, and research questions: an infant (M = 14.0 months, N = 49) study of novel event probabilities using violation of expectation, a study of two-year-olds' (M = 29.2 months, N = 67) syntactic bootstrapping using preferential looking, and a study of preschoolers' (M = 48.6 months, N = 148) sensitivity to the accuracy of informants using verbal responses. Our goal was to evaluate the overall feasibility of moving developmental methods online, including our ability to host the research protocols, securely collect data, and reliably code the dependent measures, and parents' ability to self-administer the studies. Due to procedural differences, these experiments should be regarded as user case studies rather than true replications. Encouragingly however, all studies with all age groups suggested the feasibility of collecting developmental data online and the results of two of three studies were directly comparable to laboratory results.

ASSESSING THE VIABILITY OF ONLINE DEVELOPMENTAL RESEARCH: RESULTS FROM THREE CASE STUDIES

Access to participants is one of the primary bottlenecks in developmental research. Much of the time involved in executing a study is spent, not on science per se, but on participant recruitment, outreach, scheduling, database maintenance, and the cultivation of relationships with partner institutions (preschools, children's museums, hospitals, etc.). This puts pressure on investigators to pursue questions that can be addressed with very few children per condition, motivating elegant designs but also necessarily limiting the questions that researchers can investigate. To address the participant bottleneck, we developed a new web platform, Lookit, that allows researchers to conduct behavioral studies in infants and young children online. Parents access Lookit through their web browsers, self-administer the studies with their child at their convenience, and transmit the data collected by their webcam for analysis. In a companion paper (Scott & Schulz, 2016), we discuss the conceptual motivation behind Lookit and the overall feasibility of the approach. Here we report the results of three user case studies designed to assess the platform's methodological potential.

We conducted three test studies adapted from previously published studies, selected arbitrarily with the criteria that each use video stimuli, focus on a different age group (infants,

Citation: Scott, K., Chu, J., Schulz, L. (2017) Lookit (Part 2): Assessing the viability of online development research, results from three case studies. *Open Mind: Discoveries in Cognitive Science*. Advance Publication. doi:10/11.62/opmi_a_00001

DOI:
http://doi.org/10.1162/opmi_a_00001

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

Received: 21 July 2016
Accepted: 21 July 2016

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

Corresponding Author:
Kimberly Scott
kimscott@mit.edu

The MIT Press

toddlers, and preschoolers), and use a different dependent measure (violation of expectation, preferential looking, and verbal responses). The three studies selected were 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 (36–60 months) based on Pasquini, Corriveau, Koenig, and Harris (2007). All studies were adapted for testing in the online environment and, as such, do not constitute true replications. Although similar results on Lookit and the lab would be grounds for optimism (as in a validity assessment), our goal is neither to better estimate the true effect size in each study (replication) nor to judge whether the results obtained on Lookit are acceptably close to accepted values (formal validation). Rather, these experiments should be regarded as user case studies in the development of a new online platform. Differences from published works may indicate areas where the online methodology needs refinement or further study.

METHODS

Details relevant to the platform as a whole (recruitment, consent, video quality, intercoder agreement, etc.) are discussed in a companion paper (Scott & Schulz, 2016). For additional details of procedures and analysis for each study, and example participant videos, see the Supplemental Materials (Scott, Chu, & Schulz, 2016). Q1

Data Analysis

Data were analyzed using MATLAB 2014b (MathWorks, Inc., Natick, MA) and R version 3.2.1 (R Core Team, 2015); see Supplemental Materials for code. Sample sizes were determined in advance, although the final sample size depended on coding completed after stopping data collection. No conditions were dropped. All dependent measures were collected and participant exclusions are noted.

Exclusion

Participants were excluded if they did not have a valid consent record, if they were not in the age range for the study, and if their video was not usable. Table 1 summarizes the number of children excluded for these reasons in each study. Study-specific exclusions are noted in individual methods sections. One concern was that underrepresented families might be disproportionately likely to encounter technical problems or otherwise be excluded from final analysis. To address this, we performed a logistic regression of whether each participant was included in the final sample (N = 176) or not (N = 345) for all participants with complete demographic data on file, using income, maternal education, multilingual status, Hispanic origin, and White race as predictors. (For details, see Supplemental Materials.) We found

Table 1. Numbers of unique Lookit participants in each study and numbers excluded before study-specific criteria.

	Study 1	Study 2	Study 3
Unique participants	269	329	399
Invalid consent	52	51	58
Out of age range	20	13	28
Unusable video	85	125	79
Potentially included	112	140	234

no evidence of any collective effect of demographic details on inclusion in the studies. The model was not significant (chi-square = 1.71, $df = 6$, $p = .89$) and no individual predictors were significant (all p's $> .2$).

STUDY 1: SINGLE-EVENT PROBABILITY, USING LOOKING TIME IN INFANTS

This study was adapted from Téglás et al. (2007), Experiment 1. In the original study, infants were shown videos of "lottery containers" with four bouncing objects inside: three blue and one yellow (counterbalanced). Infants looked longer when the yellow object exited through a narrow opening in the container, demonstrating sensitivity to the probability (25% vs. 75%) of events they had never seen before.

Method

Participants Videos from 112 toddlers (11–18 months) were coded for looking time unless the child was fussy or distracted on more than one test trial ($n = 36$) or the parent's eyes were open on at least one test trial or she peeked on at least two ($n = 13$). Participants were excluded if recording did not start on time ($n = 1$), if brief inability to see the child's eyes affected any looking time measurement by over 4 s ($n = 6$), or, following the original study, if more than two test trials were at ceiling or any test trial had a looking time under 1 s ($n = 7$). Data from 49 toddlers (23 female) are included in the main analysis. Their ages ranged from 11.0 to 18.0 months ($M = 13.9$ months, $SD = 2.2$ months).

Procedure The study started with four familiarization trials. On each familiarization trial, a video appeared showing two blue and two yellow objects bouncing in a container for 14 s. The container was then occluded and parents were instructed to press the space bar to continue the experiment, at which point a single object emerged from the container (blue or yellow, counterbalanced). The occluder was removed, and the video paused for 20 s to measure the infant's looking time to the outcome. The familiarization trials were followed by four test trials in which one of the objects was a different shape and color from the others. Outcomes alternated between probable and improbable, with order counterbalanced. On the last two familiarization trials and all four test trials, parents were instructed to close their eyes from the time they pushed the space bar until audio instructions signaled the end of the trial.

We used the same stimuli as in Téglás et al. (2007). Some changes were necessary to run the study online (see Table 2). Of these, the most important is that videos were not contingent on infant gaze. This would require automated gaze detection, which may be an option in the future (for a recent review, see Ferhat & Vilariño, 2016). We expanded the age range from 12–13 months to 11–18 months only to speed data collection. Additionally, we asked parents to close their eyes rather than wear opaque glasses. To minimize the time during which parents had to close their eyes, we asked them to close their eyes only for each trial outcome, introducing a delay before the start of the trial.

Coding Two coders blind to condition coded each clip for looking time using VCode (Hagedorn, Hailpern, & Karahalios, 2008) and for fussiness, distraction, and parent actions, as described in Scott & Schulz (2016). Looking time for each trial 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. Coders agreed on 95% of frames on average ($N = 63$ children; $SD = 5.6\%$), and the mean absolute difference in looking time between coders was .77 s ($SD = .94$ s).

Table 2. Summary of differences between Téglás et al. (2007) and the Lookit study (Study 1).

	Téglás et al. (2007)	**Lookit**
Parent blinding	Parents wore opaque glasses for the entire study.	Parents were asked to close their eyes during looking time measurements, but could watch the initial videos of objects bouncing in the lottery containers. Coders checked parent compliance.
Familiarization trials	Two	Four (repeated original sequence)
Infant control	Videos of objects bouncing in the container were paused whenever the infant wasn't paying attention; test trials ended after lookaway.	No infant control during videos.
Starting test trials	Test trial started (object exited container) only once infant was looking at the screen (experimenter controlled).	Test trial started after 5-s instructions and once infant was looking at the screen (parent controlled).
Ending test trials	First continuous 2-s lookaway or at 30-s total looking time.	After 20 s; looking time measured from video as time until first 1-s lookaway.
Age range	12–13 months	11–18 months
Exclusion criteria	Looking at ceiling on more than two test trials, looking away in synchrony with object exiting the container, or fussy (20 of 40 participants excluded).	Looking at ceiling on more than two test trials, looking < 1 s on any test trial, or fussy or distracted on more than one test trial (43 of 92 participants excluded, not counting participants excluded for technical problems or parent interference).
Physical setup	Container covering 14 x 14-cm area on a 17-inch monitor; infants 80 cm away from monitor.	Videos displayed full-screen on participants' computer monitors; monitor sizes and distances not measured. Container covers approx. 11 x 11-cm area on a typical 13-inch laptop monitor or 20 x 20-cm area on a 22-inch external monitor.

RESULTS AND DISCUSSION

Mean looking times to probable and improbable outcomes for this study and Téglás et al. (2007) are shown in Figure 1. To better assess the effect of outcome probability on infants' looking times, we performed a hierarchical linear regression of the four test trial looking times (grouped by child) against whether the outcome was improbable and trial number, omitting trials where children were distracted or fussy (see Table 3). Improbable outcomes were associated with an increase in looking time of 0.62 s (95% CI [-0.68, 1.92]). This is smaller than the 3.21 s (95% CI [0.36, 6.1]) reported by Téglás et al. (2007), which was replicated twice within the lab with differences of 3.7 s (Téglás, Vul, Girotto, Gonzalez, Tenenbaum, & Bonatti, 2011) and 3.4 s (Téglás, Ibanez-Lillo, Costa, & Bonatti, 2015). However, variances in looking times in each condition are similar (4.0–4.2 s on Lookit compared to 3.6–5.6 s in the original data). We observed no correlations between fraction looking time to improbable outcomes and age, total looking time during training trials, or number of fussy/distracted trials (all $|r| <$ .1 and $p > .5$, Spearman rank order correlation).

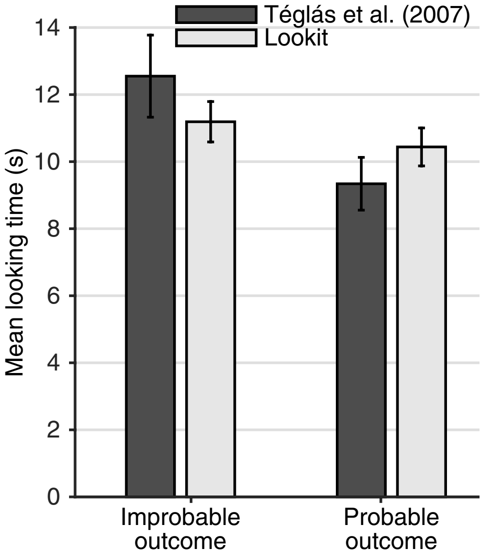


Figure 1. Mean looking times to improbable and probable outcomes for Study 1 (*N* = 49) and Téglás et al. (2007) Experiment 1 (*N* = 20).
Error bars show *SEM*.

This study confirms our ability to collect and reliably code looking times on Lookit, with good intercoder agreement despite varying webcam placement and video quality. The effect we observed on Lookit was smaller but in the same direction as Téglás et al. (2007). Although there was no evidence that our broader age range or fussiness/distraction contributed to the reduced effect, several procedural differences likely contributed. First, parents were asked to close their eyes and press the space bar after the container was occluded and before the outcome was displayed; this introduced a 5-s delay, potentially an untenable memory demand for the infants. Fortunately, this is not due to any fundamental limitation of the online platform; this delay could be avoided by having the parent close her eyes earlier in the trial or throughout. Second, the original experiment paused familiarization videos when the infant looked away and ended test trials upon lookaway. Here both familiarization and test videos were displayed at a fixed rate and neither was contingent on infant gaze. Third, we cannot discount the possibility that the looking time measure was not well-suited to online testing in children's homes, requiring a less distracting or more standardized environment. However, overall mean

Table 3. Coefficients for hierarchical linear regression of test trial looking times, excluding measurements where the child was fussy or distracted (178 measurements, three or four per participant, 49 participants).

	B	*SE B*	*p*	95% CI
Intercept	13.08	.94		[11.23, 14.94]
Improbable (1 or 0)	.62	.66	.35	[-0.68, 1.93]
Order (1, 2, 3, 4)	-.86	.29	.003	[-1.44, -0.29]

Note. *B* = regression coefficient (unstandardized); *SE* = standard error; *p* = p-value for the null hypothesis that *B* = 0; CI = confidence interval.

looking times on Lookit were comparable to those on the same videos in the lab (M = 10.8 s, SD = 3.6 s), suggesting that we did not simply lose the child's attention to more interesting visual environments at home.

STUDY 2: PARTICIPANT ROLES FROM SYNTACTIC STRUCTURE, USING PREFERENTIAL LOOKING IN TWO-YEAR-OLDS

This study was adapted from Yuan and Fisher (2009). The original study demonstrated syntactic bootstrapping in toddlers (27–30 months). A novel transitive or intransitive verb ("blicking") was introduced by videotaped dialogues, absent any potential visual targets for the verb's meaning. Later, children were either asked to "Find blicking" or simply asked "What's happening?" and looking preference for videotaped one- versus two-participant actions was measured. Children who had heard a transitive verb showed a preference for the two-participant actions at test only when asked to find the verb, showing that they had mapped the transitive verbs onto two-participant events.

Method

Participants We received potentially usable video from 140 toddlers (24–36 months). One child was excluded for previous participation and another due to a technical problem that led to recording starting late. Videos from the remaining 138 participants were coded; children were excluded if the parent pointed, spoke, or had her eyes open during either test trial, or if the parent peeked on both test trials (n = 14). Children were also excluded if their practice scores were under .15 (n = 50), their total looking time during the two test trials was less than 7.5 seconds (n = 1), or they looked at less than 60% of the dialogue (n = 6). Data from the remaining 67 children (32 female) are included in the main analysis. Their ages ranged from 24.2 to 35.8 months (M = 29.2 months, SD = 3.8 months).

Procedure The study started with two practice phases using familiar verbs, one intransitive (clapping) and one transitive (tickling). Each practice phase consisted of the child being asked to find the target verb while a video of the target action was shown on one side of the screen and a distractor video (showing sleeping or feeding, respectively) was simultaneously playing on the other side. This 8-s trial was repeated three times.

Next, children saw a video of two women using one of four novel verbs (blicking, glorping, meeking, or pimming) in three short dialogues comprising four transitive or intransitive sentences each. Finally, children completed a test phase in which a one-participant and a two-participant action were shown simultaneously on opposite sides of the screen. In the experimental condition, children were asked to find the novel verb (e.g., "Find blicking!"). In the control condition they were simply asked, "What's happening?" Parents were instructed to close their eyes during the second practice phase and the test phase. Immediately before the practice and test phases, a 13-s calibration clip was shown in which a colorful spinning ball appeared first on the left and then on the right of the screen.

Differences between this replication and the original experiments are summarized in Table 4. Because delay between the dialogue and test phase was not found to matter in the original studies, we chose a shorter delay, closer to Yuan and Fisher's (2009) Experiment 1, for convenience. We created four stimulus sets (verbs paired with examples of one- and two-participant actions) to reduce the probability that an observed effect (or lack thereof) could be explained by low-level features of the stimuli.

Table 4. Summary of procedural differences between Yuan & Fisher (2009) and Study 2.

	Yuan & Fisher (2009), Exp. 1	**Yuan & Fisher (2009), Exp. 2, Same-Day Condition**	**Lookit**
Dialogues before familiar verbs? ("Mary was clapping")	Yes	No (to avoid superficial learning during training)	No
Familiar verbs	Clapping and tickling		
Delay between end of novel verb dialogue and start of test phase	7 s	100–120 s	13 s (calibration video)
Novel-verb dialogue	8 sentences	12 sentences	12 sentences
Novel verbs and actions used	Blick; one intransitive and one transitive action		Blick, glorp, meek, pimm; four of each type of action paired to verbs
Video events	8 s, on two separate screens		8 s, on same screen
Test trials	2	3	3; first 2 used in analysis to allow stricter inclusion criteria regarding fussiness and parent intervention
Control (What's happening?)	No	Yes	Yes
Age range	26.6 – 30.2 months (*M* = 28.6 months)	26.8 – 30.4 months (*M* = 28.4 months)	24.2 – 35.8 months (*M* = 29.2 months)
Exclusion criteria	Fussiness (6%)	Side bias (3%), distraction (1%), practice trial performance >= 2.5 *SD* below mean (1%), looking time to 2-participant event > = 2.5 *SD* from condition mean (3%)	Parent interference (10%), practice scores below 0.15 (36%), low total looking time to test or dialogues (5%)

Coding Two coders blind to left/right placement of action videos coded each clip using VCode (Hagedorn et al., 2008) and for fussiness, distraction, and parent actions, as described in Scott & Schulz (2016). As in the original study, our dependent measure was preferential looking; although some children pointed at the screen, this was rare, and we expected looking to be a more robust measure. For each trial, we computed the fraction of total looking time spent looking to the right/left (fractional looking time). Mean disagreement between coders on fractional looking time was 4.4% of trial length (*SD* = 2%, *N* = 138 participants).

A practice score was computed for each child as fractional right looking time during the "tickling" practice phase minus fractional right looking time during the "clapping" practice phase. A score of 1 indicates looking only to the target actions. The mean practice score across the 138 coded participants was .24 (*SD* = .22); 36% of children were excluded due to practice scores under .15 (see Figure 2). Children were attentive during the dialogues, looking for an average of 83% (*SD* = 16%) of the duration of the dialogue videos.

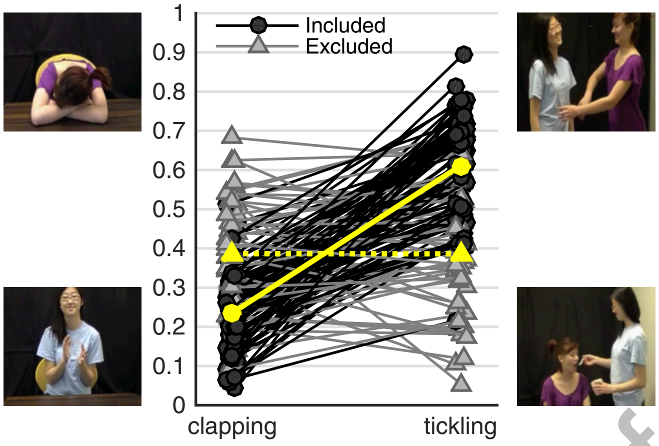


Figure 2. Fraction of time spent looking to the right during the two practice action phases in Study 2.
In the first phase, children were asked to find clapping (on the left), and in the second they were asked to find tickling (on the right). Each of 138 potentially included children's responses is displayed as one line; the x-axis position is jittered slightly for display. Children were only included in the main analysis if the fraction of looking time to the right was at least 0.15 greater during "tickling" than "clapping" clips. Yellow lines show the mean fractions of time spent looking to the right for included (solid line) and excluded (dotted line) participants.

RESULTS AND DISCUSSION

Based on the original studies conducted by Yuan and Fisher (2009), we expected the proportion of looking time to the two-participant events to be greater for transitive than intransitive verbs in the experimental condition, but not the control condition. Figure 3 shows the fractional two-participant looking time per child. To measure the effect of transitive verbs in each condition, we performed a linear regression of the fractional two-participant looking time on condition (experimental/control), verb type (transitive/intransitive), interaction between condition and verb type, and stimulus set used (dummy-coded). Observations were weighted by $(0.5 - 0.5 * \text{practice\_score})^{-2}$, using the time spent looking at mismatching actions during practice to estimate measurement variance. This reflects the intuition that if a child performed "well" on practice trials then her behavior at test better indicates her understanding of the novel verb. Regression coefficients are shown in Table 5.

For comparison, we performed a similar regression on the data from the 80 children in Experiment 2 of Yuan and Fisher (2009); for details, see the Supplemental Materials. Figure 4 shows the beta coefficients associated with transitive verbs in each condition. We observe effect sizes similar to the original in both conditions on Lookit, with a positive effect in the experimental condition and nearly zero effect in the control condition.

The familiar-verb trials afforded an opportunity to check for effects of family socioeconomic status (SES) on data quality. Among potentially usable sessions with demographic data on file, we did not observe any correlations between practice scores and either SES measure (income: $r = .033$, $p = .74$, $n = 101$; maternal education: $r = .077$, $p = .44$, $n = 103$, Spearman rank order correlations).

This study confirms that we can collect and reliably code preferential looking measures on Lookit, with good intercoder agreement. Like Yuan & Fisher (2009), we observed increased

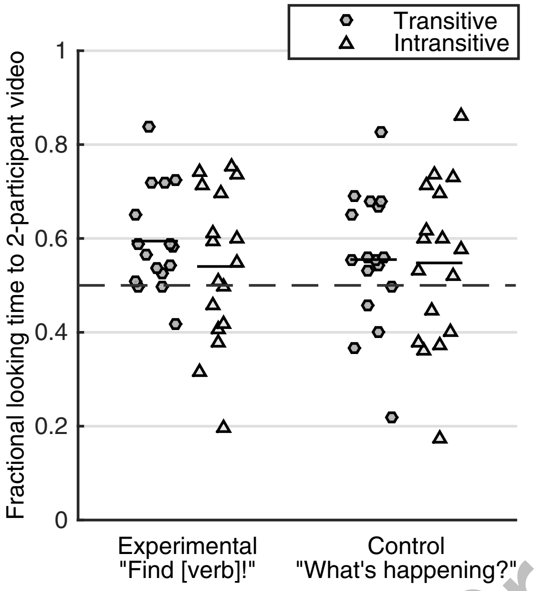


Figure 3. Fraction of total looking time (left and right) across the two test trials spent looking toward the two-participant action by condition.
Each dot represents one child; lines are drawn at the means. The dotted line at 0.5 represents equal looking to one- and two-participant actions.

looking to the two-participant actions only when children were asked to find transitive verbs. However, looking to the correct actions during the practice phases was less reliable than in the original study; additional work is needed to obtain preferential looking responses online that are directly comparable to behavior in the lab and retain more of the data collected. Practice scores were not correlated with SES measures, suggesting that differences were due to the presentation medium rather than the population.

Table 5. Coefficients for practice score weighted linear regression of fractional two-participant looking time on condition, verb type, interaction between condition and verb type, and stimulus set used (*N* = 67).

	B	*SE B*	*p*	95% CI
Intercept	.013	.04	.76	[-.07, .10]
Condition (1 = experimental, 0 = control)	.0010	.05	.98	[-.10, .10]
Verb type (1 = transitive, 0 = intransitive)	.0014	.05	.98	[-.10, .10]
Experimental * transitive	.10	.07	.18	[-.05, .25]
Stimuli set 1 (of 4)	.11	.05	.04	[.01, .21]
Stimuli set 2	-.067	.05	.17	[-.17, .03]
Stimuli set 3	-.002	.05	.97	[-.10, .09]

Note. Fractional two-participant looking time is reduced by 0.5 so that 0 represents equal looking to one- and two-participant actions. The coefficient associated with verb type here corresponds to the effect of verb type within the control condition. *B* = regression coefficient (unstandardized); *SE* = standard error; *p* = p-value for the null hypothesis that *B* = 0; CI = confidence interval.

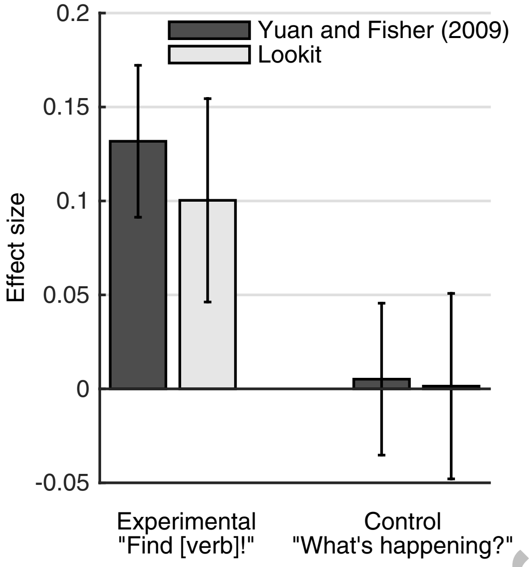


Figure 4. Effect sizes for effect of transitivity on fractional looking time to two-participant actions in experimental and control conditions.
Effect sizes are the coefficients associated with transitive, as compared to intransitive, verbs from 2 x 2 linear regressions of the fraction of looking time each child spent looking at two-participant actions against condition and verb type. The regression of data from Study 2 (*N* = 67) additionally weighted data based on variance estimated from practice trials and included predictors for the various stimulus sets used. Each regression was run with condition coded "Find [verb]!" = 0, "What's happening?" = 1 and vice versa so that the beta coefficients associated with transitive verbs reflected the increase in fractional looking time to transitive verbs in the "Find [verb]!" and "What's happening?" conditions respectively. Error bars show standard errors. Data from Yuan and Fisher (2009) used with permission (*N* = 80).

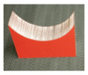
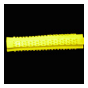
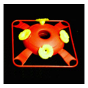
STUDY 3: TRUST IN TESTIMONY, USING VERBAL RESPONSES FROM PRESCHOOLERS

The final user study was adapted from Pasquini et al. (2007), which investigated whether 3- and 4-year-olds monitor and evaluate the previous reliability of informants when deciding which of two informants to trust. Children watched videos in which two informants, one more and one less reliable, labeled four familiar objects and later provided conflicting labels for novel objects. Four-year-olds explicitly identified the more accurate informant and endorsed her novel-object labels in all conditions tested; 3-year-olds' performance was distinguishable from chance when one informant was 100% accurate.

Method

Participants We received potentially usable video from 125 three-year-olds and 109 four-year-olds. Children were excluded for failure to answer at least three of the four familiar-object questions correctly (33 three-year-olds and 14 four-year-olds), failure to choose either of the informants on the first explicit-preference question (9 three-year-olds and 9 four-year-olds), and failure to endorse either informant's label on at least one of the novel-object questions (11 three-year-olds and 10 four-year-olds). (For details of exclusion criteria selection see the Supplemental Materials.) Data from the remaining 72 three-year-olds (45 female; *M* = 3.53 years) and 76 four-year-olds (44 female; *M* = 4.54 years) are included in the main analysis.

Table 6. Object pictures and labels used in Study 3.

Familiar object				
Label 1 (accurate)	spoon	bottle	brush	doll
Label 2 (inaccurate)	duck	apple	plate	cup
Label 3 (inaccurate, used if both informants are inaccurate)	hat	fork	key	tree
Novel object				
Label 1 (given by girl in yellow shirt)	toma	danu	dax	riff
Label 2 (given by girl in red shirt)	gobi	modi	wug	fep

Procedure Following the original study, children completed four familiar-object trials, one initial explicit judgment trial, four novel-object trials, and one final explicit judgment trial. Children were assigned to one of four conditions where informants demonstrated 100% vs. 0%, 100% vs. 25%, 75% vs. 0%, or 75% vs. 25% accuracy, transforming the original within-subjects to a between-subjects design in order to keep the study short (about 10 min). During object trials, children saw a video of two informants taking turns labeling the same object. The informants' answers were repeated and the child was asked what he/she thought the object was called (endorsement measure). Onscreen instructions guided parents to replay the question if needed or to prompt the child without repeating the object labels. The objects and labels used are shown in Table 6. During explicit judgment trials, children were asked, "Who was better at answering these questions, the girl in the yellow shirt or the girl in the red shirt?" Differences between this replication and the original experiments are summarized in Table 7.

See the Supplemental Materials for coding procedures. If the parent gave the answer before the child's final answer, it was treated as an invalid answer. Parents interfered in 8% of trials by repeating the two options or answering the question themselves before the child's final answer.

RESULTS AND DISCUSSION

We analyzed two measures of children's performance: their answers to the initial explicit-judgment question and the number of endorsements of each informant's labels during the novel-object phase. To compare these endorsements to the original results, we calculated the fraction correct (number of endorsements of the more accurate informant divided by total number of endorsements). Performance on each measure in the present study and Pasquini et al. (2007) is shown in Figure 5. Children were less likely to endorse either label on Lookit than in the original study (where all children were required to answer all questions); only 21% of 3-year-olds and 45% of 4-year-olds gave valid answers to all four endorsement questions. Overall, we did observe modestly lower performance on Lookit on both explicit judgment and endorsement questions; weighting each condition and age group equally, Lookit scores were 7.2 percentage points lower on explicit preference questions and 10 percentage points lower on endorsement questions. However, the overall patterns observed were similar: the eight explicit judgment and endorsement performances, per condition and age group, were highly

Table 7. Summary of procedural differences between Pasquini et al. (2007) and Study 3.

	Pasquini et al., 2007, Exp. 1	**Pasquini et al., 2007, Exp. 2**	**Lookit**
Accuracy conditions	100% vs. 0% 100% vs. 25% 75% vs. 0%	75% vs. 0% 75% vs. 25%	100% vs. 0% 100% vs. 25% 75% vs. 0%
Design	Within-subjects; different informants and objects used for each condition		Between-subjects; same informants and objects used for each condition. (Familiar objects and labels, and novel- object labels, were the ones used by Pasquini et al., 2007, for the 100% vs. 0% condition of Exp. 1 and the 75% vs. 0% condition of Exp. 2.)
False-belief task	Yes (no relationship with other measures found)	No	
Dependent measures	Endorsement, explicit judgment, and ask ("which person would you like to ask?"). Children were also asked what each object was called before the informants answered. No main effects of question type were found on ask vs. endorsement questions with explicit judgment as a covariate.		Endorsement and explicit judgment; to keep the experiment short we omitted the "ask" measure and did not ask children what objects were called before trials.
Explicit judgment question	One of these people was not very good at answering these questions. Which person was not very good at answering these questions?	1. [For each informant] Was the girl with the ___ shirt good at answering the questions or was she not very good at answering the questions? 2. Who was better at answering the questions: the girl in the ___ shirt or the girl in the ___ shirt?	Who was better at answering these questions, the girl in the yellow shirt or the girl in the red shirt?
Exclusion criteria	Incorrect response to any familiar-object question, unless both informants were incorrect (6% overall)		Failure to answer at least 3 of 4 familiar-object questions correctly (20%), failure to choose an informant first explicit-preference question or an informant's label on at least one endorsement question (17%)

correlated across the two studies (explicit preference: $r = .78$, $p = .024$; endorsement: $r = .88$, $p = .004$).

To check for effects of SES on performance, we conducted a logistic regression of explicit judgment responses from the 105 included subjects with demographic information on file, using a composite SES score (mean of z-scored maternal education level and z-scored family income) in addition to age in years, condition, and age by condition interactions. The effect

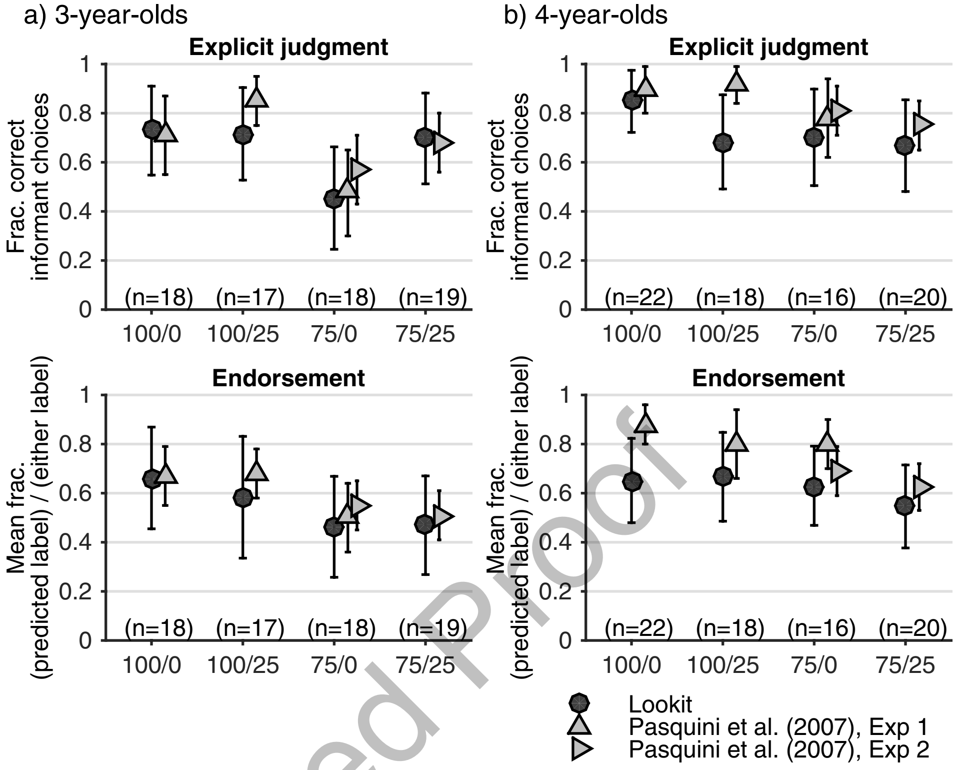


Figure 5. Mean performance on explicit judgment questions (top row) and endorsement questions (bottom row), by age group.
Means and 95% confidence intervals are plotted for Study 3 and for Experiments 1 and 2 of Pasquini et al. (2007). The study was conducted between-subjects on Lookit and within-subjects in both Pasquini et al. experiments. Explicit judgment performance is 0 or 1 for Lookit participants and 0, 0.5, or 1 for Pasquini et al. participants. A child's endorsement question performance is the number of times (over the four trials) that she chose the label of the more accurate informant divided by the number of times she endorsed either label.

of SES on the probability of giving a correct response was small and nonsignificant ($e^B = 1.00$, 95% CI [0.56, 1.83], $z = .02$, $p = .98$).

In this study we confirmed the viability of Lookit for collecting verbal responses from preschoolers. Despite increased variation expected due to the between-subjects design and slightly reduced performance overall, we observed very similar patterns of performance based on age group and condition compared to the original study.

GENERAL DISCUSSION

Collectively, our user case studies confirm the feasibility and suggest the promise of conducting developmental research online. Parents of children ranging from infants to preschoolers were able to access the platform and self-administer the study protocols in their homes at their convenience. Researchers were able to securely collect and reliably code looking time, preferential looking, and verbal response measures. We did not observe any relationships between SES and children's performance. More critically (since such relationships may well emerge or be the topic of investigation in other studies), SES differences did not adversely affect parents' ability to interact with the platform: there was no effect of SES on exclusion rates. This suggests that online testing can fulfill the goals of expanding access and lowering barriers to participation in developmental research.

The current project was designed to investigate the possibility of collecting looking time, preferential looking, and verbal response measures online; the results of the user case studies suggest that in these respects, the project was successful. However, in adapting the studies to the online environment, the studies fell short of true replications. Assessing the degree to which various designs and results are directly reproducible online, and whether sample diversity moderates effect size, remains an important direction for future research, and will be critical to understanding the relationship between online testing and laboratory-based protocols.

Although we cannot yet conclude that measures collected on Lookit are directly comparable with those collected in the lab, the similarity of the results of Studies 2 and 3 to published results is very encouraging. In Study 1, we observed effects in the same direction as the lab-based study, but a smaller effect size than initially reported; further research must determine to what extent this was due to the protocol differences we introduced or to difficulties adapting infant looking time measures to the online environment. As noted in the accompanying conceptual paper (Scott & Schulz, 2016), online testing is not appropriate for every study, and may be more appropriate for some designs than others (i.e., preferential looking rather than looking time). The initial empirical results, however, provide grounds for optimism about the potential of Lookit to extend the scope, transparency, and reproducibility of developmental research.

ACKNOWLEDGMENTS

We thank all of the families who participated in this research. Thanks to the Boston Children's Museum where we conducted early piloting of computer-based studies. We thank 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 Kathleen Corriveau, Cynthia Fisher, and Erő Téglás for original stimuli and helpful comments; 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

KMS developed the methodology, designed the studies, and collected data with the advice of LES. Data analysis and interpretation was performed by KMS with contributions from JC. KMS and LES prepared the manuscript.

REFERENCES

Ferhat, O., & Vilariño, F. (2016). Low cost eye tracking: The current panorama. *Computational Intelligence and Neuroscience, 2016*(3), 1–14. doi.org/10.1155/2016/8680541

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

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. doi/10.1037/0012-1649.43.5.1216

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Scott, K. M. & Schulz, L. E. (2016). Lookit: A new online platform for developmental research. *Open Mind*.

Q2 Scott, K. M., Chu, J., & Schulz, L.E. (2016). Replication data for: Assessing the viability of online developmental research: Results from three case studies. dx.doi.org/10.7910/DVN/TMOQMC, Harvard Dataverse, V1.

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

Téglás, E., Ibanez-Lillo, A., Costa, A., & Bonatti, L. L. (2015). Numerical representations and intuitions of probabilities at 12 months. *Developmental Science*, *2*, 183–193. doi.org/10.1111/desc.12196

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. Retrieved from http://10.1126/science.1196404

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. dx.doi.org/10.1111/j.1467-9280.2009.02341.x