Supplemental information

Methods

Data and analysis code: Please see the Open Science Framework repository https://osf.io/9b4qj/.

Sample size determination:
Our goal in planning the test studies was not statistical power but accuracy in parameter estimation. In the absence of objective criteria for acceptable confidence intervals, sample sizes were set at one to two times the original sample sizes, based on the number of children we expected to be able to run during initial development of the Lookit platform. We planned 40 participants in Study 1; the original sample size was 20 (Téglás et al., 2007). We planned 64 participants (16 per condition) in Study 2; Yuan and Fisher (2009) had 8 children per condition in Experiment 1 and 8 children per condition in the same-day condition of Experiment 2 (20 children per condition total in Study 2). We planned 64 three-year-olds and 64 four-year-olds in Study 3, 16 per condition; the original study used 20-25 children in each condition in Experiments 1 and 2, but used a within-subjects design (Pasquini et al., 2007). Actual sample sizes exceed the planned Ns slightly because data collection was ongoing as coding determined inclusion and because conditions were randomly assigned.

Exclusion criteria:
Exclusion criteria for the studies here stem from several sources. First, some participants are necessarily excluded due to incomplete or unusable data, for instance because of invalid consent records, failure to complete the study, or technical problems with video recording. Second, we have tried to follow the original studies' criteria when possible, including age ranges. We did plan and use a larger age range in Study 1 to speed recruitment, and checked after collecting the data that age was not correlated with performance; other age ranges are as in the original studies. Finally, several exact criteria based on the child's and parent's behavior (for instance, how many fussy trials were acceptable) were informed by the distributions of these behaviors, although not by their relationship to main dependent measures, in order to use data as clean as possible without excessive exclusion rates.

Demographics

To address concerns about disproportionate exclusion of data from underrepresented families, we performed a logistic regression of inclusion in the final sample for all participants in the correct age range who had a valid consent video and the following predictors:

1. Total yearly family income, z-scored. Income was categorized based on ranges selected by the parent: 15 (Under $30,000), 30 ($30,000 - $50,000), 50 ($50,000 - $75,000), 75 ($75,000 - $100,000), or 100 (Over $100,000).

2. Maternal education, z-scored, approximated as the parent's education level if the parent completing the survey was female and the parent's spouse's education level otherwise. Education was categorized into four levels: some high school or high school diploma; some college or 2-year degree; 4-year degree, some or attending graduate school; and some graduate or professional school or advanced degree. We used maternal education in addition to income level because it is one of the best single proxies for SES in childhood (Bornstein et al., 2003; Hoffman, 2003).
3. Multilingual status (0 or 1)
4. Hispanic origin reported (0 or 1)
5. Race listed as White (0 or 1).

521 participants had data for all five predictors, of whom 176 were included in studies. We did observe significant correlations between predictors for family income and maternal education ($r = 0.46$, 95% CI [0.39, 0.52], $p < 0.001$), family income and white race ($r = 0.11$, 95% CI [0.02, 0.19], $p = 0.02$), multilingual status and Hispanic origin ($r = 0.36$, 95% CI [0.28, 0.43], $p < 0.001$), and multilingual status and White race ($r = -0.10$, 95% CI [-0.18, -0.01], $p = 0.02$). We found no evidence of any collective effect of demographic details on inclusion in the studies. Overall, the model was not significant (chi square = 1.71, df = 6, $p = 0.89$) and no individual predictors were significant (all p's > 0.2).

It is still possible that parents without valid consent videos differed from those with valid consent videos in regards to these predictors. However, among the in-age-range participants with valid consent videos, rates of demographic survey responses did not differ between included (67%) and excluded (70%) participants.

Study 1

Participants

85 children were excluded due to unusable video; out of 115 sessions with unusable video, 63 had no video, 18 were incomplete due to the parent ending the study early, 6 were missing video due to technical errors on the Lookit server, 20 had video change rates under 2 fps or only audio, 3 did not generally show the child's eyes, and 5 were too dark or blurry to code. One child was excluded due to parental interference on the basis that the parent reported squinting during the study, although this was not obvious on the video.

No children included in the analysis had participated previously; two had previously encountered errors before reaching the first video.

Procedure

All videos were displayed full-screen on the participant's monitor. Parents controlled the start of each phase by clicking, so that if a child was distracted they could wait until she was ready to continue.

Counterbalancing

Test videos always alternated between probable and improbable outcomes. Equal numbers of children saw probable and improbable outcomes first. One of the same four exact sequences of exiting objects was shown in each case. Video sequences were counterbalanced for color and shape of outcomes; some started with a blue object exiting the container and some with a yellow object, and some alternated color of exiting objects (e.g. blue sphere, yellow cube, blue cube, yellow sphere) while some alternated shape (e.g. blue cube, yellow sphere, yellow cube, blue sphere). Table S3 shows numbers of included subjects in each counterbalancing condition.

First test video	First color	Alternate	# subjects	Exact sequence of outcomes
Probable	Blue	Shape	11	A
		Color	4	B
	Yellow	Shape	6	C
		Color	6	D
Improbable	Blue	Shape	9	A
		Color	3	B
	Yellow	Shape	7	C
		Color	3	D

Table S3: Summary of counterbalancing conditions and numbers of children assigned to each.

Coding

Coders recorded looks to and away from the screen, as well as any periods where the child's eyes were out of the frame, to the nearest frame. If the child's eyes were not visible during the entire trial, looking times were computed by treating the periods where eyes were not visible as looking and as not looking and averaged. Children were excluded if the disparity between these two measurements was over 4 s on any test trial. (For instance, a brief period of not being able to see the eyes could make it unclear whether an early lookaway ended the looking time measurement.)

Results

We used a hierarchical linear model to estimate the effect size in order to allow for both decreasing looking time over the course of the study and an effect of the probability of the outcome, while omitting individual trials where a child was fussy or distracted. This approach is both more robust to the potential impacts of exclusion criteria and allows us to keep more of our data while acknowledging that trials where an infant is fussy are unlikely to be meaningful. Gender, outcome object color (blue/yellow), and gender * improbable-outcome and age * improbable-outcome interaction terms were all insignificant ($p > 0.3$) when added to the hierarchical linear regression; there was a marginal main effect of age in months ($B = 0.48$, 95% CI [-0.06, 1.03], $p = 0.08$). Using

the original model without excluding any children for fussiness, distraction, short looking times, or measurements affected by time when the eyes were not visible (and simply omitting those particular trials instead) gives a similar estimate of the improbable-outcome effect size (B = 0.22, 95% CI [-0.90, 1.35], 248 looking times from 84 participants).

The difference between mean looking times to improbable-outcome and probable-outcome trials was 0.75 s (95% CI based on normal distribution, [-0.37 s, 1.87 s]).

Discussion

The exact method of measuring looking time does not appear to explain the decreased effect size: the regression coefficient associated with improbable events remains insignificant when using a 0.5 s lookaway threshold (B = 0.82, 95% CI [-0.47, 2.12]), a 2 s threshold (B = -0.41, 95% CI [-1.95, 1.13]), or total looking time (B = -0.68, 95% CI [-1.89, 0.53]).

Slight variation in the onset of webcam recording relative to the lottery outcome contributed some noise to looking time measurements. Examining trial lengths, which roughly index this delay, we found that the median range of test trial lengths within included subjects was 0.22 s (IQR [0.14, 0.42]). However, this could not on its own explain our failure to observe the much larger looking time difference reported by Téglás et al. (2007).

The delay introduced before looking time measurements was an unfortunate consequence of an attempt to minimize the amount of time parents were required to close their eyes and allow them to view as much of the stimuli as possible. In retrospect, this complicated the procedure unnecessarily given parents' overall compliance, and we would have preferred a longer buffer between when parents were asked to close their eyes and when blinding was required (to allow parents to first see that the video had started correctly).

In principle, if infants had perfect memory, delays might simply make predictions more dependent on the ratio of blue to yellow objects rather than the objects' positions before occlusion. Téglás et al. (2011) showed that with increasing delays during which the container was occluded, infants' looking time depended more on whether the object exiting was the more common type (e.g. one of three blue cubes versus the single yellow ball) and less on whether an object of that type was closest to the exit before occlusion. At the maximum delay of 2 s, they replicated the basic finding from Téglás et al. (2007), reporting similarly increased looking times when the single object exited the container. The sensitivity infants showed to the delay, however, suggests that playing verbal instructions to the parent during this delay could have disrupted the sophisticated representations infants were maintaining.

Most research that relies on looking time is conducted in dark, quiet lab rooms where there is relatively little else to look at besides the stimuli. Although we assume that the decision-making processes that underlie looking time differences in the lab also underlie

daily learning in noisy environments, effects might be much larger when measured in isolation. Nonetheless, such decision-making processes could exert substantial effects on the child's visual experience (e.g. directing them towards improbable events in the world) when we consider that they are in effect whenever the child's eyes are open.

Study 2

Participants

125 children were excluded due to unusable video. Of the 139 sessions with unusable video, 70 had no video or incomplete video, 37 had a frame change rate under 2 fps or had only audio, 17 were missing or unusable due to technical problems with the Lookit server early in development, and 15 had other video quality problems, e.g. too dark or blurry to see the child clearly. All 138 participants with potentially usable video had practice clips at least 5s long and test clips at least 6s long out of an expected 8s. In otherwise usable cases where parents talked to children during any test trial, a coder blind to condition transcribed the speech and determined whether the speech could have potentially biased the child (e.g., "What are they doing?" or "Look at her!"). Children were not excluded due to neutral parent speech like "Shh honey" or "Keep watching!" In cases where the parent's eyes were not visible during test trials, a coder checked the parent's comments and any previous trials where her eyes were visible; children were only excluded on this basis if there was a problem (the parent reported not always closing her eyes or did not close her eyes on earlier trials). The children excluded due to parent interference include three children excluded due either to speech or the parent's eyes not being visible.

Five included subjects had previously submitted some record of starting the study, but had not progressed at all beyond submitting consent (n = 2), had only gotten through the first training verb (n = 2), or had completed the study three months previous but had not been included due to technical problems (n = 1).

Counterbalancing

Each child was assigned to one of four conditions: transitive (n=16) or intransitive (n=17) experimental group ("Find [verb]!" prompt), or a transitive (n=17) or intransitive (n=17) control group ("What's happening?" prompt). Left/right placement of the one-participant clips was counterbalanced within each condition (7-8 children in each condition saw the one-participant clip on the right), and gender (7-9 girls/condition) and age were balanced across conditions. The specific verbs used (blicking, glorping, meeking, and pimming) were counterbalanced together with action videos. Due to random variation in assignment, 'blicking' and 'glorping' were used with different frequencies across conditions; because there were also potential differences among looking times to the two-participant actions by verb used, we included the specific verb used in further analysis.

Procedure

Dialogue videos were organized in three blocks of four sentences with three second breaks in between blocks. Dialogue videos averaged 48 s long in total (range = 42 – 55 s). Between practice and between test trials the screen was black for three seconds.

All videos were displayed full-screen on the participant's monitor. Parents controlled the start of each phase by clicking, so that if a child was distracted they could wait to continue.

Coding

In one case the first two clips were much longer than expected due to a technical error; video was only analyzed for a trial length equal to the maximum allowable length.

Coders marked the times of looks to the left and right of the screen and away from the screen, and recorded times of looks to and away from the screen during the conversation clip. They also recorded any periods where the child's eyes were not visible. During calibration trials we expected looks to the left during the first third of the clip and to the right during the last third. If coders differed by more than 0.15 on the practice and test trials (when that difference constituted at least 500 ms) or by more than 1 s for the dialogue phase, the difference was flagged and the clip recoded.

Left and right looking times for each preferential looking trial were computed based on the portion of the clip starting from the onset of audio (or the last 7.5 seconds, in cases where computer audio was not available).

Practice trial performance

17% of practice scores were less than 0, indicating a shift in fractional looking time between the two practice phases opposite the predicted direction. Low practice scores were not simply explained by low total looking times during the practice trials or by greater uncertainty in coding, either of which could increase variance; we observed small and insignificant correlations between practice score and practice looking time ($r = 0.13$, $p = 0.13$, Spearman rank order correlation) and between practice score and the sum of absolute coder disagreements about looking to each side ($r = 0.01$, $p = 0.89$, Spearman rank order correlation).

Results

We first checked for any superficial effects of condition (transitive/intransitive x experimental/control). We did not observe any effects of condition on total looking time (overall M = 13.7s, SD = 1.3s, with condition means ranging from 13.3 to 13.9 s; all $p > 0.1$ in 2x2 ANOVA with interaction term) or on looking time to conversations (overall M = 88%, SD = 11%, with condition means ranging from 87% to 89%; 2x2 ANOVA with interaction term, all $p > 0.5$).

Because of differences between practice performance measures on Lookit and in the

original study, we did not weight the original data based on practice performance. First, it is possible that the reasons for varying practice performance were different in the two studies; on Lookit, for instance, overall attentional states or volume of the stimuli may have played a larger role in children's behavior. Second, Yuan and Fisher (2009) used the amount of time a child looked at the matching stimuli, which decreases with look-away time, whereas we used the fraction of the child's total looking time spent looking at the matching stimuli for each practice verb. The spread of practice performance in the original data was larger, due in part to exclusion of some Lookit participants on the basis of practice performance. Included Lookit participants looked at the correct actions on average 69% (SD = 7%) of their total looking time during practice trials; Yuan and Fisher's (2009) participants looked at the correct actions for on average 54% (SD = 13%) of the practice trial durations. So although weighting of the data is appropriate on Lookit, it may not be in the original data; results of both the weighted and unweighted regressions are shown in Table S4 for comparison.

Nevertheless, we did observe some evidence of similar effects of practice trial performance across studies. Adding practice trial performance, and all interactions with condition, to the unweighted regressions marginally improves model predictions on the Lookit data ($F(4,61)=2.47$, $p = 0.054$) and substantially improves model predictions on the original data ($F(4,74)=6.52$, $p = 0.00015$). In both cases, we observe qualitatively similar effects (same signs and ordered magnitudes of interactions).

		Weighted		Unweighted	
		Experimental: β_{trans}	Control: β_{trans}	Experimental: β_{trans}	Control: β_{trans}
Lookit, Study 2 (N=67)		0.100 (0.054)	0.001 (0.050)	0.084 (0.053)	0.001 (0.051)
Yuan and Fisher (2009) Expt. 2	All data (N=80)	0.073 (0.043)	0.004 (0.047)	0.132 (0.040)	0.005 (0.040)
	Same-day only (N=32)	0.116 (0.075)	-0.091 (0.064)	0.128 (0.066)	-0.032 (0.066)
	Different-day only (N=48)	0.079 (0.049)	0.118 (0.062)	0.134 (0.052)	0.030 (0.052)

Table S4: Effect sizes associated with transitive verbs, from 2x2 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 Lookit additionally included predictors for the various stimulus sets used. The weighted regressions were performed with data weighted according to variance estimated from practice trial performance; weights were 1 / (1 - practice score)^2. On Lookit the practice score was the fractional looking time to the right when the matching action was on the right minus the fractional looking time to the right when the matching action was on the left; for the original data the practice score was the mean fraction of the trial duration spent looking at the matching action. In both cases, 1 – practice score was approximately proportional to the fraction of time spent looking at the mismatching action. Each regression was run with condition coded Experimental = 0, Control = 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 experimental and control conditions respectively. Standard errors shown in parentheses.

Study 3

Materials and procedure

We created our own video and audio stimuli closely following Pasquini et al (2007). The familiar object choices (and incorrect answers) and the novel object labels were identical

to those used in Pasquini et al. (2007) during the 100% vs. 0% condition (Experiment 1) and the 75% vs. 25% condition (Experiment 2). Phrasing of the introduction and of questions posed to informants and children were identical to the original study's.

Before each trial, a picture of the object was shown along with a picture of the two informants, and audio played saying "I bet one of these people can help us find out what it's called." Then a video played showing the two informants sitting together across from a table with the object on it. Each informant responded to voiceover audio asking "Can you tell me what this is called?" After both informants responded, their answers were repeated and the child was asked what the object was really called: for instance, "The girl in the yellow shirt said it's called a spoon, and the girl in the red shirt said it's called a duck. What do you think it's called?" Onscreen instructions guided parents to replay this audio for their children if needed or to gently prompt them without repeating the options. Once the child answered, parents clicked a button to proceed.

In cases where either informant was not always correct or always incorrect, the particular question answered correctly/incorrectly was counterbalanced. In the 75% vs. 25% condition, the more accurate informant missed the same question that the less accurate informant answered correctly. Video of the same two female informants was shown to all children, but the identity of the more accurate informant was counterbalanced. The first informant to be asked to name the object alternated between trials, and the girl in the yellow shirt was always named first in the summary.

Participants

79 participants were excluded due to unusable video; out of 100 records, 53 had no video, 32 were incomplete due to the parent ending the study early, 11 were missing audio, 2 had no child present, 1 had a sibling who interfered with the subject's answers, and 1 withdrew from the study. One child was excluded on the basis of age because database and post-study birthdates disagreed about whether the child was 3 or 4.

Mean ages of children in each of the four conditions ranged from 3.50 to 3.55 years within the three-year-olds and from 4.47 to 4.56 years within the four-year-olds.

No children included in the analysis had participated previously; five had previously encountered errors before beginning the study and one had encountered an error before reaching the second familiar-object video.

Exclusion criteria

In contrast to Pasquini et al. (2007), Lookit participants did not answer all questions by selecting one of the two valid options. We set response-based inclusion criteria to ensure we had both some confidence in children's understanding of the task (based on familiar-object performance) and responses to each type of question, while avoiding removing excessive amounts of data.

In the original study, three-year-olds answered all familiar-object questions correctly 92% of the time and four-year-olds 94%; incorrect responses were allowed only when both informants were incorrect. On Lookit, in contrast, only 51% of three-year-olds and 73% of four-year-olds answered all four familiar-object questions correctly. This was primarily due to non-responses (only 64% and 79%, respectively, gave either the correct or incorrect label every time). We therefore required only three familiar-object questions to be answered correctly, which included most three-year-olds (74%) and four-year-olds (87%).

Similarly, Lookit participants did not always endorse either label during the novel-object phase. Only 16% of three-year-olds and 34% of four-year-olds endorsed one of the two labels in all four trials. We required only that they endorse one of the two labels on at least one trial, which included most children without substantial age bias (82% of three-year-olds and 86% of four-year-olds).

Finally, we required that participants give a valid answer during the first explicit-preference trial, and did not use responses to the second preference trial. We decided not to require both responses in order to avoid unnecessary data loss (73% of three-year-olds and 81% of four-year-olds answered both, compared to 80-83% of three-year-olds and 87% of four-year-olds who answered each individual question). However, we expected differences in performance on the first and second questions due to the intervening and possibly confusing novel-object phase. Additionally, parent interference on the first explicit-preference trial could affect all future responses (novel-object and second explicit-preference trials). Therefore we required a valid response only on the first explicit-preference trial and used this response in our analysis.

Coding

For each of the ten trials (four familiar-object, four novel-object, and two explicit-preference trials) a coder recorded a transcript of child and parent speech, an ordered list of specific actions that took place, and a final answer to be used in analysis. The final answer was determined by dividing the transcript into blocks of child speech separated by parent speech or action. Actions recorded included the parent prompting the child, repeating the question, repeating the two options, giving positive or negative feedback, or answering the question, as well as all child responses during the trial. To allow children to correct themselves but minimize the impact of parent interaction, we used the child's last valid response (one of the two options) within the earliest block in which there was a valid response.

Slight variants of each of the novel labels were allowed (e.g. "tuma" for "toma"), but any that were ambiguous between the two options (e.g. "goma" when "toma" and "gobi" were options) were coded as "combination" responses rather than endorsements of either informant.

Results

Note that because the original study was conducted within-subjects (each subject saw two to three pairs of informants, rather than one pair), differences between the Lookit replication and the original study are likely to be correlated between conditions.

We used only the initial explicit-judgment question rather than including the final question because some children did not answer both, and we expected there could be differences between their performance before and after the novel-object questions.

The original study drew conclusions about age differences based primarily on whether performance was significantly above chance in each condition; we instead performed a logistic regression of the 148 responses to the explicit judgment question against age, whether the child was prompted, condition (dummy-coded, with the 100% vs. 0% condition as reference) and age x condition interactions. The results are summarized in Table S5. We observed a positive but insignificant effect of age ($z = 0.38$, $p = 0.70$), a strong negative effect of prompting ($z = -2.62$, $p = 0.0088$), and negative effects of being in the more difficult conditions, although only one difference from the 100% vs. 0% condition was individually significant.

	Odds ratio e^B	95% CI e^B	z	P(>|z|)
Intercept	11.25	[4.20, 36.45]		
Age (in years, mean-centered)	1.36	[0.28, 7.49]	0.38	0.70
Prompted on this question (1 or 0)	0.32	[0.13, 0.74]	-2.62	0.0088 (**)
Condition: 100% vs. 25%	0.51	[0.15, 1.65]	-1.10	0.27
Condition: 75% vs. 0%	0.27	[0.080, 0.85]	-2.18	0.029 (*)
Condition: 75% vs. 25%	0.42	[0.13, 1.32]	-1.45	0.15
Age * (100% vs. 25%)	0.67	[0.076, 5.53]	-0.37	0.71
Age * (75% vs. 0%)	1.34	[0.16, 10.99]	0.28	0.78
Age * (75% vs. 25%)	0.33	[0.041, 2.35]	-1.09	0.28

Table S5: Logistic regression results for 148 children's explicit judgment question performance (correct = 1, incorrect = 0) in Study 3. For instance, being prompted on a given question is associated with being only 32% as likely to answer it correctly.

To assess the relationship between performance on the explicit judgment and endorsement questions, we replicated the analysis of covariance (ANCOVA) in Pasquini et al. (2007). The dependent variable was the difference between more-accurate informant endorsements and less-accurate informant endorsements, to take into account variation in how many questions children answered. Age group (3-, 4-year-olds) and accuracy condition were both between-subjects variables in our design, and explicit judgment performance was entered as a covariate. As in Pasquini et al. (2007), this ANCOVA revealed a main effect of explicit judgment performance ($F(1,139) = 3.94$, $p = 0.049$), although the effect was smaller ($\eta^2 = 0.028$, 90% CI [0.000, 0.082] vs. $\eta^2 = 0.12$ in the original data). We did not observe a significant effect of age group ($F(1,139) = 0.03$), although the original study found an effect of about the same size as that of explicit judgment performance. As in the original paper, we did not observe significant effects of condition ($F(3,139) = 0.37$) or of the interaction between age group and condition ($F(3,139) = 0.33$).

Socioeconomic status and performance

	SES				Age (years)			
	B	95% CI B	t(158)	p(>|t|)	B	95% CI B	t(158)	p(>|t|)
Correct response to familiar-object questions (0-4)	0.12	[-0.06, 0.31]	1.31	0.19	0.41	[0.12, 0.70]	2.76	0.006 **
Number of endorsement responses to novel-object questions (0-4)	-0.24	[-0.49, 0.004]	-1.94	0.054	0.45	[0.07, 0.84]	2.34	0.021 *
Number of trials with parental interference (0-5)	-0.10	[-0.28, 0.070]	-1.17	0.24	-0.16	[-0.42, 0.11]	-1.15	0.25
Number of trials with parent prompting (0-5)	-0.25	[-0.54, 0.04]	-1.68	0.10	-0.88	[-1.33, -.43]	-3.85	0.0002 ***

Table S6: Regression coefficients and statistics for two selected predictors, SES and age, in regressions of each of four dependent variables. Each regression also used accuracy condition (dummy-coded) as a predictor. Values of B for can be interpreted as the increase in the score (range given in the first column) associated with being one year older or with one-standard-deviation increases in maternal education level and family income level.

We additionally checked for potential effects of socioeconomic status (SES) on performance in this study that could affect inclusion. We used the average of z-scored maternal education level and z-scored family income, each binned as in *Demographics*, as a simple measure of SES, to maximize our power to detect an overall effect. We included in the analysis the 164 children aged 3-5 with valid consent, potentially usable video and both maternal education level and family income on file. We performed regressions against age (in years), accuracy condition (dummy-coded), and SES of the following dependent variables: number of correct responses (0-4) during familiar-object trials, number of endorsement responses (0-4) during novel object trials, number of test trials (0-5) where parents interfered before the child's final answer by answering the question or repeating the two options, and number of test trials (0-5) where parents prompted the child (including by repeating the options) before his/her final answer. Regression coefficients for SES and age are shown in Table S6. Although, as expected, older children answered more familiar object questions correctly, answered more endorsement questions, and required less prompting, no associations between SES and dependent measures were significant.

Although parent prompting is associated with less accurate answers on a per-trial basis, the weak effect of SES on parent prompting does not lead to any association between SES and correct responses; when adding composite SES as a predictor to the model shown in Table S5, restricted to the 105 children who have demographic data on file, the odds ratio $e^{B} = 1.00$ [0.56, 1.83]. In a study design where such differences in parent behavior could produce a confound, it would be important to standardize the protocol further by reducing the parent's involvement and/or providing more training.

References

Bornstein, M. H., Hahn, C., Suwalsky, J.T.D., & Haynes, O. M. (2003). Socioeconomic status, parenting, and child development: The Hollingshead Four-Factor Index of Social Status and the socionomic index of occupations. In M. H. Bornstein & R. H. Bradley (Eds.), *Socioeconomic status, parenting, and child devleopment* (pp. 29–82). Mahwah, NJ: Lawrence Erlbaum Associates.

Hoffman, L. W. (2003). Methodological issues in studies of SES, parenting, and child development. In M. H. Bornstein & R. H. Bradley (Eds.), *Socioeconomic status, parenting, and child development* (pp. 125–143). Mahwah, NJ: Lawrence Erlbaum Associates.

Video legends

Note: All videos are from sessions where parents chose a privacy level of "free" and confirmed that publicity and educational use of the video are acceptable.

SI Video 1: An example session from Study 1. The bottom panel shows the participant video collected; the top panel shows the stimuli this child was viewing at the time (reconstructed based on the condition, not recorded directly). Clips are labeled in the top left corner; the first four clips are practice trials and the last four clips are test trials.

SI Video 2: An example session from Study 2. The bottom panel shows the participant video collected; the top panel shows the stimuli this child was viewing at the time (reconstructed based on the condition, not recorded directly). Clips are labeled in the top left corner. Clips 1, 5, and 10 are calibration trials; 2, 3, 4, 6, 7, and 8 are practice trials; 9 is the dialogue introducing the new verb; and 11, 12, and 13 are test trials.

SI Video 3: An example session from Study 3. The bottom panel shows the participant video collected; the top panel shows the stimuli this child was viewing at the time (reconstructed based on the condition, not recorded directly). Clips are labeled in the top left corner. Clips 1–4 are familiar-object questions, 6–9 are novel-object questions, and 5 and 10 are explicit preference questions.