nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-0913-y

Check for updates

Response latencies and eye gaze provide insight on how toddlers gather evidence under uncertainty

Sarah Leckey[1,2]✉, Diana Selmeczy[1,2], Alireza Kazemi[1,2], Elliott G. Johnson[1,3], Emily Hembacher[1,2] and Simona Ghetti[1,2]✉

Toddlers exhibit behaviours that suggest judicious responses to states of uncertainty (for example, turning to adults for help), but little is known about the informational basis of these behaviours. Across two experiments, of which experiment 2 was a pre-registered replication, 160 toddlers (aged 25 to 32 months) identified a target from two partially occluded similar (for example, elephant versus bear) or dissimilar (for example, elephant versus broccoli) images. Accuracy was lower for the similar trials than for the dissimilar trials. By fitting drift–diffusion models to response times, we found that toddlers accumulated evidence more slowly but required less evidence for similar trials compared with dissimilar trials. By analysing eye movements, we found that toddlers took longer to settle on the selected image during inaccurate trials and switched their gaze between response options more frequently during inaccurate trials and accurately identified similar items. Exploratory analyses revealed that the evidence-accumulation parameter correlated positively with the use of uncertainty language. Overall, these findings inform theories on the emergence of evidence accumulation under uncertainty.

The ability to recognize one's uncertainty is critical for learning. When uncertain, we gather disambiguating information from knowledgeable sources[1,2], seek information to fill gaps in understanding[3,4] or opt out of responding to avoid mistakes[5,6]. The examination of the ability to experience uncertainty can illuminate the mechanisms that underlie metacognition (for example, monitoring ongoing mental processes[7]), including insight into function and boundary conditions[8]. However, little is known about the emergence of this ability.

Recent innovative studies[9–11] have suggested that infants exhibit behaviours that may track their states of knowledge or ignorance. In one study, infants aged 20 months solicited parents' help more frequently when they had not seen or didn't remember the location of a toy[12]. In another study, 18-month-old infants were more persistent in searching for a hidden object at the correct location after shorter delays compared with after longer delays[13]. Overall, infants and toddlers seemingly respond to uncertain situations with hesitation or information seeking behaviours. However, these studies have not examined how the available evidence is assessed before rendering an overt decision (for example, pointing or turning to a parent). This assessment arguably provides the informational basis for decisions such as asking for help in infants and toddlers. Failure to account for this assessment represents a substantial gap given that evidence accumulation is central to theories of adult metacognition and decision-making[14].

It has been proposed that implicit error signals support bids for help in infants and overt uncertainty judgements in children and adults[8,15]. Our guiding hypothesis is that these signals manifest in toddlers' behaviours during the time between stimulus onset and response selection. Specifically, we argue that response latencies[16], as well as looking times and gaze switches, may provide valuable insights. These behaviours have been examined primarily in metacognitive research in older children and adults[17,18] whose inaccurate responses in self-paced cognitive tasks are typically executed more slowly, and are associated with lower confidence[1,19,20] (results using speeded tasks can be found in refs. [21,22]). These data suggest that response latencies may reflect processes of evidence accumulation and decision making[16]. However, response latencies may also capture the operation of additional processes, such as stimulus processing or response execution[16], making it imperative to use methods that differentiate among them. Decision models, specifically drift–diffusion models, have been developed precisely to isolate the processes that contribute to response latencies in adults[16,23]. These models estimate three parameters from response latencies, including drift rate, boundary-separation and non-decision variables. The drift rate parameter captures the speed of evidence accumulation. For example, difficult perceptual tasks should result in a lower drift rate, indicating a lower quality of information and slower evidence accumulation. The boundary-separation parameter reflects the criterial amount of evidence needed to respond. For example, more difficult perceptual tasks may result in lower boundary separation (or distance between accurate and inaccurate responses) because individuals might recognize that a response must be eventually submitted, even when evidence is not optimal. The non-decision parameter captures variability in processes that are not related to decisions, such as the time it takes individuals to encode the stimuli or execute a response. Overall, these models formally specify that response latencies reflect decision and non-decision processes.

Drift–diffusion models have been used extensively in adults, but only rarely in developmental research and exclusively in older children[24–26]. For example, Ratcliff et al.[26] had third graders and young adults determine whether presented stimuli were or were not real words. In the easy condition, non-words were random letter strings, whereas, in the hard condition, they were pronounceable non-words. Both children and adults were less accurate in the hard condition compared with the easy condition. The non-decision parameter was greater in children compared with adults, indicating that age differences in response times may depend on processes that are not related to the assessment of available evidence. Critically, in

[1]Center for Mind and Brain, University of California, Davis, Davis, CA, USA. [2]Department of Psychology, University of California, Davis, Davis, CA, USA. [3]Human Development Graduate Group, University of California, Davis, Davis, CA, USA. ✉e-mail: ssleckey@ucdavis.edu; sghetti@ucdavis.edu

both children and adults, lower drift rate and boundary-separation parameters were observed in the hard condition compared with the easy condition. Thus, when faced with more difficult trials that might generate more uncertainty, both children and adults accumulated evidence more slowly, but also demanded less evidence to endorse a response option. We investigated whether toddlers may also accumulate evidence more slowly, but require less evidence, when faced with more difficult decisions.

Other behaviours, such as looking behaviours, can provide additional insights into toddlers' evidence accumulation and decision processes. For example, toddlers may visually inspect response options more closely when decisions are more difficult and more evidence is necessary, wavering between response options. Looking behaviours, including gaze switches between response options, have been associated with uncertainty monitoring in older children and adults[18,27]. In a recent study examining event-related potentials, Goupil and Kouider[13] found that a component resembling the error-related negativity (ERN) marking error monitoring in adults[28] is stronger when 12-month-old infants directed their first look after test stimulus onset to the distracter instead of the target stimulus. The necessary absence of a response selection (for example, pointing) given the age of the infants prevents us from establishing conclusively whether this initial eye gaze maps onto a response selection, because the initial attentional capture towards a stimulus may reflect a variety of additional factors (such as responses to stimulus complexity, ambiguity and novelty[29]). Our research examines how looking times and gaze switches map onto toddlers' response selection following the experimenter's instruction.

Here we conducted two identical experiments, of which experiment 2 was a preregistered replication (https://osf.io/nvf2j/), to examine how toddlers gather evidence when faced with situations that might generate uncertainty. We manipulated decision difficulty by asking toddlers to identify a target between two partially occluded images that were either similar (for example, an elephant and a bear) or dissimilar (for example, a cow and broccoli). Similar trials were expected to yield lower accuracy rates compared with dissimilar trials. Parallel versions of the tasks that included different exemplars of the same stimuli were used in a touchscreen task (to collect response latencies) and in an eye-tracker task (to collect eye movement data).

We expected longer average response latencies for inaccurate trials. Critically, using drift–diffusion models, we expected a slower drift rate for similar trials compared with dissimilar trials. We are aware that model fitting is typically pursued with numerous trials[30], which is difficult to achieve with toddlers. To mitigate this problem, we first fit the models at the group level, ensuring a large number of trials, and verified that we observed better model fit for the drift–diffusion model compared with a baseline model. We used parameter estimates at the individual level to verify that similar patterns of results were found and to conduct exploratory individual differences analyses. Specifically, we investigated whether evidence accumulation was associated with task-independent indices of toddlers' ability to report uncertainty. Toddlers begin to state 'I don't know' to express their ignorance between 2 and 2.5 years of age[31]. If parameters of evidence accumulation and decision processes capture aspects of an implicit ability to respond to states of knowledge or uncertainty, then toddlers who demonstrate more efficient evidence-accumulation abilities and demand more evidence may be more likely to have linked these signals to their experience and verbal expression of uncertainty.

Finally, we examined toddlers' looking times and gaze switches between response options. We expected toddlers to take longer to show a preference for their eventual choice and switch gaze between response options more frequently for more difficult trials, consistent with toddlers gathering more disambiguating information under conditions of greater uncertainty.

Results of experiment 1

Accuracy. The accuracy scores of the toddlers ranged from 0.56 to 1.00 (mean ± s.d., 0.84 ± 0.10) for the eye-tracker task and 0.53 to 1.00 (0.79 ± 0.13) for the touchscreen task (data and analytical code are provided at https://osf.io/t8p4g/). To verify that yoked items produced similar performance, we fit a logistic regression with trial accuracy (0 or 1) on the eye-tracker task as the dependent measure and trial accuracy on the touchscreen task as a fixed effect; trial and participant were entered as random effects. Accuracy on a given trial of the touchscreen task significantly predicted accuracy on the yoked trial of the eye-tracker task conducted one week apart ($\beta = 0.40$, $P = 0.03$, $d = 0.13$, 95% confidence interval (CI) = 0.02–0.77).

We found higher accuracy rates for dissimilar trials compared with similar trials in both tasks (eye-tracker: dissimilar, 0.88 ± 0.11, similar, 0.79 ± 0.16, $t_{72} = 4.43$, $P < 0.001$; $d = 0.52$, 95% CI = 0.05–0.14; touchscreen: dissimilar, 0.86 ± 0.14, similar, 0.71 ± 0.17, $t_{70} = 8.47$, $P < 0.001$; $d = 1.01$, 95% CI = 0.11–0.18). Thus, similar items were more difficult than dissimilar items across tasks.

Response latencies and drift–diffusion parameters. For the next set of analyses, we compared similar–accurate, dissimilar–accurate and inaccurate trials. Unless otherwise noted, inaccurate trials were collapsed across similar and dissimilar trials to have sufficient observations. This comparison enabled us to consider the effects of both accuracy and similarity. Post hoc analyses using both paired *t*-tests and multilevel models confirmed that there were no differences between dissimilar–inaccurate and similar–inaccurate trials (Extended Data Fig. 3, Supplementary Results 1).

Response latencies. We conducted a one-way (dissimilar–accurate versus similar–accurate versus inaccurate) repeated-measures analysis of variance (ANOVA) with average response latency as the dependent measure (an alternative multilevel model is provided in Supplementary Table 1a). We found a main effect of trial type ($F_{2,136} = 6.46$, $P = 0.002$, $\eta_p^2 = 0.09$, 95% CI = 0.01–0.18) such that response latencies were shorter for similar–accurate and dissimilar–accurate compared with inaccurate trials ($t_{68} = -3.28$, $P = 0.002$, $d = -0.39$, 95% CI = −1,152.95 to −280.66; and $t_{68} = -2.62$, $P = 0.01$, $d = -0.32$, 95% CI = −1,080.38 to −146.37, respectively). There was no statistically significant difference between similar–accurate and dissimilar–accurate trials ($t_{68} = -0.54$, $P = 0.59$, $d = -0.06$, 95% CI = −486.61–279.75; Fig. 1a). This suggests that overall response latencies are sensitive to trial accuracy. We found no statistically significant evidence that age in months was correlated with the differences between trial types (similar–accurate minus inaccurate: $r = 0.07$, $P = 0.54$, 95% CI = −0.17–0.30; dissimilar–accurate minus inaccurate: $r = 0.09$, $P = 0.44$, 95% CI = −0.15–0.32; dissimilar–accurate minus similar–accurate: $r = 0.03$, $P = 0.80$, 95% CI = −0.21–0.26).

Drift–diffusion parameters. Using all trials across the entire sample, we found that both drift rate (v) and separation (a) parameters seemed to be higher for dissimilar ($v = 0.51$, $a = 4.00$) compared with similar trials ($v = 0.28$, $a = 3.77$). The non-decision parameter did not seem to differentiate greatly between the two types of trials (dissimilar $t_0 = 0.62$; similar $t_0 = 0.58$). Compared with a baseline model in which similar and dissimilar trials were not differentiated, the main model provided a significantly better fit to the data (main model, Akaike information criterion (AIC) = 6,336.14, Bayesian information criterion (BIC) = 6,356.89; baseline model, AIC = 6,385.98, BIC = 6,406.73; $\chi^2_3 = 49.84$, $P < 0.001$). Quantile probability plots also suggested an acceptable fit of the main model (Extended Data Fig. 1a). In the individual-level models, the dissimilar condition yielded significantly greater estimates than the similar condition in both drift rate ($t_{67} = 7.57$, $P < 0.001$, $d = 0.92$, 95% CI = 0.31–0.53) and separation parameters ($t_{67} = 3.91$,

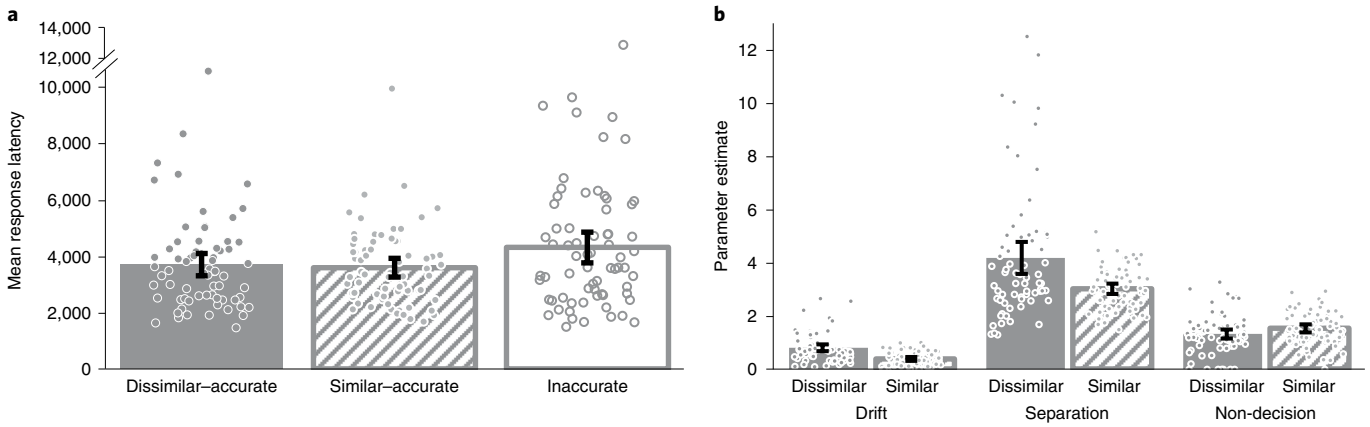


Fig. 1 | Response latencies and parameter estimates for the touchscreen task of experiment 1. a, Mean response latencies for the dissimilar–accurate, similar–accurate and inaccurate trials; $n = 71$ toddlers (dissimilar–accurate versus inaccurate trials: $t_{68} = -2.62$, $P = 0.01$, $d = -0.32$, 95% CI = −1,080.38 to −146.37; similar–accurate versus inaccurate trials: $t_{68} = -3.28$, $P = 0.002$, $d = -0.39$, 95% CI = −1,152.95 to −280.66; dissimilar–accurate versus similar–accurate trials: $t_{68} = -0.54$, $P = 0.59$, $d = -0.06$, 95% CI = −486.61–279.75). The *y* axis is broken between 10,000 and 12,000 ms. **b**, Mean drift-diffusion parameter estimates for the similar and dissimilar trials for the drift, separation and non-decision parameters; $n = 71$ toddlers (dissimilar versus similar trials for the drift parameter: $t_{67} = 7.57$, $P < 0.001$, $d = 0.92$, 95% CI = 0.31–0.53; dissimilar versus similar for the separation parameter: $t_{67} = 3.91$, $P < 0.001$, $d = 0.47$, 95% CI = 0.57–1.75; dissimilar versus similar for the non-decision parameter: $t_{70} = -2.14$, $P = 0.04$, $d = -0.25$, 95% CI = −0.41 to −0.01). The points represent individual data points. Data are jittered on the *x* axis to avoid stacking. The error bars show the 95% confidence intervals.

$P < 0.001$, $d = 0.47$, 95% CI = 0.57–1.75; Fig. 1b). Thus, the patterns of results for these parameters are similar to those obtained with the sample-level model (Supplementary Table 2a). The non-decision parameter was greater in the similar than the dissimilar condition ($t_{70} = -2.14$, $P = 0.04$, $d = -0.25$, 95% CI = −0.41 to −0.01). Supplementary analyses fitting a hierarchical drift–diffusion model (HDDM)[32] provided converging results (Supplementary Results 2).

We then used an individual difference approach and conducted a multiple regression analysis in which drift rate and separation parameters were entered simultaneously to predict the frequency that the toddlers use the 'I don't know' expression. The age and overall vocabulary of the toddlers were also included in the multiple regression. The model was significant ($R^2 = 0.24$, 95% CI = 0.08–0.40, $F_{4,63} = 5.08$, $P = 0.001$, $\eta^2 = 0.24$, 95% CI = 0.05–0.40) and both overall vocabulary ($\beta = 0.32$, $P = 0.01$, $d = 0.35$, 95% CI = 0.09–0.54) and the drift rate parameter ($\beta = 0.33$, $P = 0.004$, $d = 0.36$, 95% CI = 0.11–0.56) significantly predicted 'I don't know' use on the basis of the parental report. We found no statistically significant evidence that the separation parameter ($\beta = 0.02$, $P = 0.86$, $d = 0.02$, 95% CI = −0.20–0.24) and age ($\beta = 0.06$, $P = 0.58$, $d = 0.07$, 95% CI = −0.17–0.29) predicted 'I don't know' responses.

Eye movements. *Trajectory of the proportion of looking time to selected responses.* To investigate the differences in the trajectory of proportion of looking time towards the selected item as a function of trial type, we conducted a multilevel model, with fixed effects of trial type (dissimilar–accurate, similar–accurate and inaccurate) and time bin (0–1 s, 1–2 s and 2–3 s) and a random effect of individual. To estimate the significance of our model, we used a χ^2 difference test, testing its difference from a baseline model (intercept only). We found a main effect of trial type ($b = -0.08$, $t_{2,991} = -2.28$, $P = 0.02$, $d = -0.04$, 95% CI = −0.16 to −0.01), which was qualified by a trial type by time bin interaction (Fig. 2a; Extended Data Fig. 4a), indicating different patterns of preferential looking as a function of time. Our model was significantly different compared with the baseline model ($\chi^2_8 = 64.45$, $P < 0.001$). In time bin 1 (0–1 s), we found that there were no statistically significant differences among trial types (similar–accurate compared to inaccurate: $b = -0.09$, $P = 0.04$, Bonferroni-corrected P (P_{Bonf}) = 0.12, $d = -0.06$, 95% CI = −0.17–0.00; dissimilar–accurate compared to inaccurate: $b = -0.04$, $P = 0.39$, $P_{Bonf} = 1.00$, $d = -0.03$, 95% CI = −0.12–0.05; similar–accurate compared to dissimilar–accurate: $b = 0.05$, $P = 0.12$, $P_{Bonf} = 0.35$, $d = 0.05$, 95% CI = −0.01–0.11). In time bin 2 (1–2 s), we found that the proportion of looking times for dissimilar–accurate trials was significantly greater than those for inaccurate trials ($b = 0.13$, $P < 0.001$, $P_{Bonf} < 0.001$, $d = 0.12$, 95% CI = 0.06–0.19). Finally, in time bin 3 (2–3 s), we found that the proportion of looking times for both dissimilar–accurate trials and similar–accurate trials was significantly greater than those for inaccurate trials ($b = 0.17$, $P < 0.001$, $P_{Bonf} < 0.001$, $d = 0.14$, 95% CI = 0.09–0.25; and $b = 0.11$, $P = 0.01$, $P_{Bonf} = 0.02$, $d = 0.09$, 95% CI = 0.03–0.18, respectively). Overall, although toddlers ended up eventually showing a preference for the image that they chose (Extended Data Fig. 2), they showed this preference earliest in the trial for dissimilar–accurate decisions compared with inaccurate decisions.

Gaze switches. The preferential looking results suggested that dissimilar items required less deliberation. Our hypothesis is that toddlers explore options more in trials that require more evidence to make a decision. To directly test this hypothesis, we examined gaze switches between stimuli. We conducted a one-way (dissimilar–accurate versus similar–accurate versus inaccurate) repeated-measures ANOVA with the mean number of gaze switches as the dependent measure. We found a main effect of trial type ($F_{2,126} = 3.90$, $P = 0.02$, $\eta_p^2 = 0.06$, 95% CI = 0.0004–0.14) with fewer gaze switches for dissimilar–accurate compared with similar–accurate ($t_{63} = -4.15$, $P < 0.001$, $d = -0.84$, 95% CI = 0.18–0.51). There was no significant difference between similar–accurate and inaccurate ($t_{63} = 0.41$, $P = 0.68$, $d = 0.09$, 95% CI = −0.24–0.37) or between dissimilar–accurate and inaccurate ($t_{63} = -1.95$, $P = 0.06$, $d = 0.25$, 95% CI = −0.56–0.01; Fig. 2b; a post hoc complementary analyses of within-image gaze transitions is provided in Supplementary Results 3 and Supplementary Figs. 1 and 2a).

Results of experiment 2

Experiment 2 was a preregistered replication study. The effects of similarity on looking times, drift–diffusion parameters, and the association between drift parameters and mental state language were preregistered for this study. We predicted that the results of this experiment would replicate those observed in experiment 1.

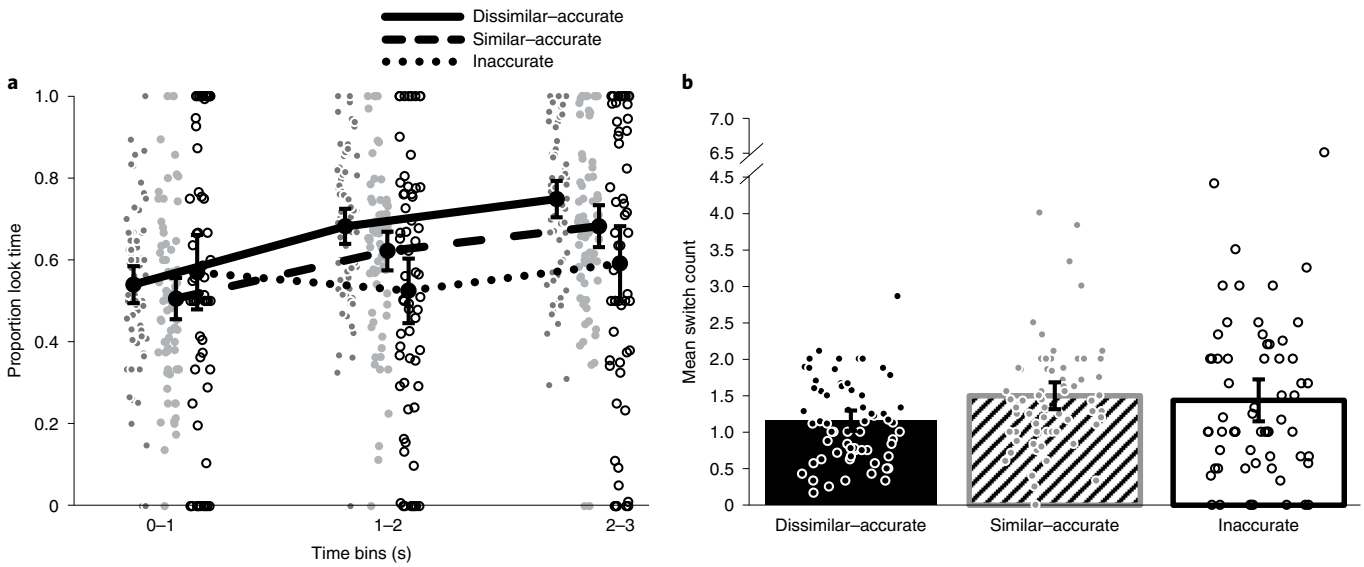


Fig. 2 | The proportion of looking times and switch counts for the eye-tracker task of experiment 1. a, The proportion of looking time towards the chosen image for the first 3 s of the trial in 1 s time bins for the dissimilar–accurate, similar–accurate and inaccurate trials; $n = 73$ toddlers (bin 1 dissimilar–accurate versus inaccurate trials: $b = -0.04$, $P = 0.39$, $P_{Bonf} = 1.00$, $d = -0.03$, 95% CI = −0.12–0.05; bin 1 similar–accurate versus inaccurate trials: $b = -0.09$, $P = 0.04$, $P_{Bonf} = 0.12$, $d = -0.06$, 95% CI = −0.17–0.00; bin 1 dissimilar–accurate versus similar–accurate trials: $b = 0.05$, $P = 0.12$, $P_{Bonf} = 0.35$, $d = 0.05$, 95% CI = −0.01–0.11; bin 2 dissimilar–accurate versus inaccurate trials: $b = 0.13$, $P < 0.001$, $P_{Bonf} < 0.001$, $d = 0.12$, 95% CI = 0.06–0.19; bin 2 dissimilar–accurate versus similar–accurate trials: $b = 0.06$, $P = 0.02$, $P_{Bonf} = 0.06$, $d = 0.07$, 95% CI = 0.01–0.11; bin 3 dissimilar–accurate versus inaccurate trials: $b = 0.17$, $P < 0.001$, $P_{Bonf} < 0.001$, $d = 0.14$, 95% CI = 0.09–0.25; bin 3 similar–accurate versus inaccurate trials: $b = 0.11$, $P = 0.01$, $P_{Bonf} = 0.02$, $d = 0.09$, 95% CI = 0.03–0.18). 0 s is stimulus onset. **b**, The mean number of gaze switches for dissimilar–accurate, similar–accurate and inaccurate trials; $n = 73$ toddlers (dissimilar–accurate versus inaccurate trials: $t_{63} = -1.95$, $P = 0.06$, $d = 0.25$, 95% CI = −0.56–0.01; similar–accurate versus inaccurate trials: $t_{63} = 0.41$, $P = 0.68$, $d = 0.09$, 95% CI = −0.24–0.37; dissimilar–accurate versus similar–accurate trials: $t_{63} = -4.15$, $P < 0.001$, $d = -0.84$, 95% CI = 0.18–0.51). The *y* axis is broken between 4.5 and 6.5 gaze switches. The points represent individual data points. The average values on the proportion of look time graph and the data points on both graphs are jittered on the *x* axis to avoid stacking. The error bars show the 95% confidence intervals.

Accuracy. Accuracy scores ranged from 0.53 to 1.00 (0.82 ± 0.12) for the eye-tracker task and 0.53 to 1.00 (0.76 ± 0.12) for the touchscreen task. As in experiment 1, accuracy on a given trial of the touchscreen task significantly predicted accuracy on the yoked trial of the eye-tracker task ($\beta = 0.87$, $P < 0.001$, $d = 0.31$, 95% CI = 0.54–1.33).

As in experiment 1, in the touchscreen task, we found higher accuracy rates for dissimilar items (0.82 ± 0.15) compared with similar items (0.71 ± 0.15) ($t_{66} = 5.43$, $P < 0.001$; $d = 0.66$, 95% CI = 0.07–0.15). In the eye-tracker task, we also found higher accuracy rates for dissimilar items (0.84 ± 0.16) compared with similar items (0.80 ± 0.15) ($t_{60} = 2.21$, $P = 0.03$; $d = 0.28$, 95% CI = 0.01–0.10).

Response latencies and drift–diffusion parameters. *Response latencies.* As in experiment 1, our repeated-measures ANOVA (an alternative multilevel model is shown in Supplementary Table 1b) showed a main effect of trial type ($F_{2,126} = 4.90$, $P = 0.01$, $\eta_p^2 = 0.07$, 95% CI = 0.01–0.16) such that response latencies were faster for similar–accurate and dissimilar–accurate trials compared with inaccurate trials ($t_{63} = -2.29$, $P = 0.03$, $d = -0.29$, 95% CI = −1,433.25 to −97.06; and $t_{63} = -2.50$, $P = 0.02$, $d = -0.31$, 95% CI = −1,934.08 to −213.99, respectively). No statistically significant difference was found between the similar–accurate and dissimilar–accurate trials ($t_{63} = 1.11$, $P = 0.27$, $d = 0.14$, 95% CI = −247.98–865.75; Fig. 3a). We found no statistically significant evidence that age in months was correlated with the differences between trial types (similar–accurate minus inaccurate: $r = 0.04$, $P = 0.79$, 95% CI = −0.21–0.28; dissimilar–accurate minus inaccurate: $r = -0.02$, $P = 0.86$, 95% CI = −0.26–0.23; dissimilar–accurate minus similar–accurate: $r = -0.08$, $P = 0.54$, 95% CI = −0.32–0.17).

Drift–diffusion parameters. Using all trials across the entire sample, we found that the drift rate and separation parameters were higher for the dissimilar condition ($v = 0.39$, $a = 4.20$) compared with the similar condition ($v = 0.24$, $a = 4.00$). The non-decision parameter was the same for both dissimilar and similar conditions ($t_0 = 0.53$). The AIC and BIC metrics and the χ^2 difference test revealed that our main model is a significantly better fit to the data (main model: AIC = 6,384.17, BIC = 6,404.58; baseline model: AIC = 6,406.21, BIC = 6,426.62; $\chi^2_3 = 22.04$, $P < 0.001$). Quantile probability plots suggested an acceptable fit (Extended Data Fig. 1b). Similar to the sample-level results and as shown in experiment 1, paired samples *t*-tests on individual level parameters revealed significant differences between the drift rate ($t_{63} = 6.07$, $P < 0.001$, $d = 0.76$, 95% CI = 0.16–0.31) and the separation parameter ($t_{63} = 2.27$, $P = 0.03$, $d = 0.28$, 95% CI = 0.04–0.69) such that the parameters for the dissimilar condition were greater than the parameter for the similar condition (Fig. 3b). There were no statistically significant differences between the non-decision parameters ($t_{65} = -0.28$, $P = 0.78$, $d = -0.03$, 95% CI = −0.29–0.22; Fig. 3b).

We found that our regression model—using the drift rate, separation parameter and general vocabulary to predict 'I don't know' use—was significant ($R^2 = 0.20$, 95% CI = 0.04–0.36, $F_{4,60} = 3.86$, $P = 0.01$, $\eta^2 = 0.20$, 95% CI = 0.02–0.35). In contrast to experiment 1, only overall vocabulary ($\beta = 0.46$, $d = 0.45$, $P = 0.001$, 95% CI = 0.21–0.71) significantly predicted 'I don't know' use; there was no statistically significant evidence that drift rate ($\beta = -0.04$, $P = 0.72$, $d = -0.05$, 95% CI = −0.28–0.19), separation parameter ($\beta = 0.09$, $P = 0.47$, $d = 0.09$, 95% CI = −0.15–0.32) and age ($\beta = -0.12$, $P = 0.36$, $d = -0.11$, 95% CI = −0.37–0.14) predicted 'I don't know' responses (an alternative multiple regression model

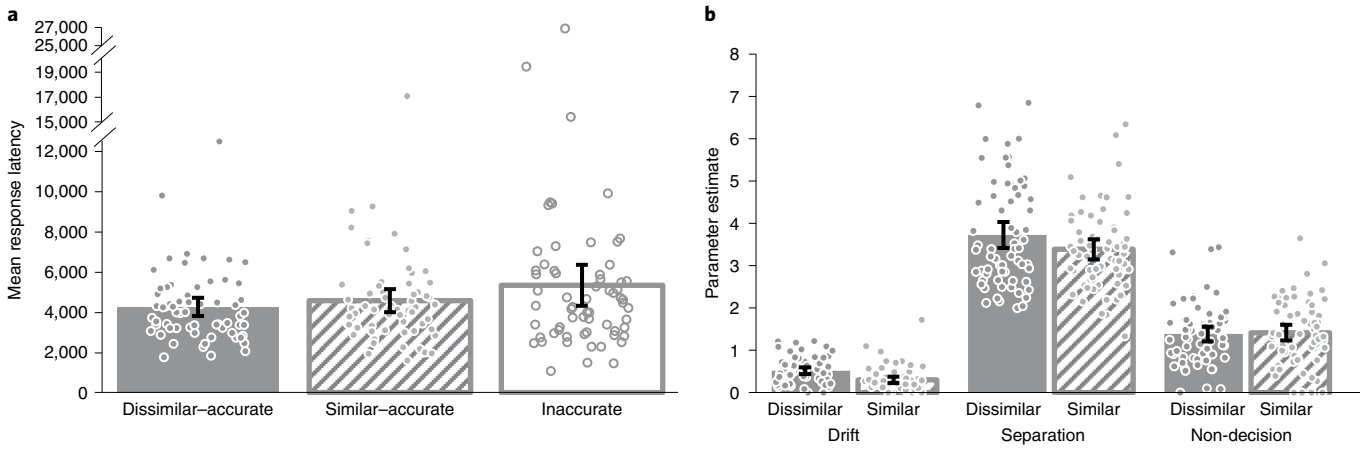


Fig. 3 | Response latencies and parameter estimates for the touchscreen task of experiment 2. a, The mean response latencies for the dissimilar–accurate, similar–accurate and inaccurate trials; $n = 67$ toddlers (dissimilar–accurate versus inaccurate trials: $t_{63} = -2.50$, $P = 0.02$, $d = -0.31$, 95% CI = −1,934.08 to −213.99; similar–accurate versus inaccurate trials: $t_{63} = -2.29$, $P = 0.03$, $d = -0.29$, 95% CI = −1,433.25 to −97.06; dissimilar–accurate versus similar–accurate trials: $t_{63} = 1.11$, $P = 0.27$, $d = 0.14$, 95% CI = −247.98–865.75). The *y* axis is broken between 12,000 ms and 15,000 ms and between 19,000 ms and 25,000 ms. **b**, Mean drift-diffusion parameter estimates for similar and dissimilar trials for the drift, separation and non-decision parameters; $n = 67$ toddlers (dissimilar versus similar trials for drift parameter: $t_{63} = 6.07$, $P < 0.001$, $d = 0.76$, 95% CI = 0.16–0.31; dissimilar versus similar for separation parameter: $t_{63} = 2.27$, $P = 0.03$, $d = 0.28$, 95% CI = 0.04–0.69; dissimilar versus similar for non-decision parameter: $t_{65} = -0.28$, $P = 0.78$, $d = -0.03$, 95% CI = −0.29–0.22). The points represent individual data points. Data are jittered on the *x* axis to avoid stacking. The error bars show the 95% confidence intervals.

showing an association between separation and mental state language is shown in Supplementary Table 3).

Eye movements. *Trajectory of proportion looking times to selected responses.* As in experiment 1, our multilevel model revealed a main effect of trial type ($b = -0.09$, $t_{2,311} = -2.27$, $P = 0.02$, $d = -0.05$, 95% CI = −0.17 to −0.01), which was qualified by a significant trial type by time-bin interaction (Fig. 4a, Extended Data Fig. 4b). Our model was significantly different compared with the baseline model ($\chi^2_8 = 17.87$, $P = 0.02$). Follow-up analyses further confirmed the results from experiment 1. In time bin 1 (0–1 s), there were no statistically significant differences among trial types (similar–accurate compared to inaccurate: $b = -0.08$, $P = 0.10$, $P_{Bonf} = 0.30$, $d = -0.06$, 95% CI = −0.17–0.01; dissimilar–accurate compared to inaccurate: $b = -0.09$, $P = 0.05$, $P_{Bonf} = 0.14$, $d = -0.07$, 95% CI = −0.18–0.00; similar–accurate compared to dissimilar–accurate: $b = -0.01$, $P = 0.68$, $P_{Bonf} = 1.00$, $d = -0.01$, 95% CI = −0.09–0.06). In time bin 2 (1–2 s), we found that only dissimilar–accurate trials resulted in higher proportions of looking times compared with inaccurate trials ($b = 0.11$, $P = 0.003$, $P_{Bonf} = 0.01$, $d = 0.10$, 95% CI = 0.04–0.18). Finally, in time bin 3 (2–3 s), we found that the proportion of looking times for both dissimilar–accurate and similar–accurate trials was significantly higher compared with inaccurate trials ($b = 0.12$, $P = 0.005$, $P_{Bonf} = 0.01$, $d = 0.10$, 95% CI = 0.03–0.20; and $b = 0.12$, $P = 0.006$, $P_{Bonf} = 0.02$, $d = 0.10$, 95% CI = 0.04–0.20, respectively), but they were not statistically different from one another ($b = 0.002$, $P = 0.96$, $P_{Bonf} = 1.00$, $d = 0.001$, 95% CI = −0.06–0.07).

Gaze switches. As in experiment 1, our repeated-measures ANOVA revealed a main effect of trial type ($F_{2,112} = 3.19$, $P = 0.05$, $\eta_p^2 = 0.05$, 95% CI = 0–0.14) with more gaze switches for inaccurate compared with dissimilar–accurate ($t_{56} = -2.21$, $P = 0.03$, $d = -0.29$, 95% CI = −0.82 to −0.04). However, the number of gaze switches for similar–accurate was intermediate and not statistically significantly different from inaccurate ($t_{56} = -1.44$, $P = 0.15$, $d = -0.20$, 95% CI = −0.74–0.12) or dissimilar–accurate ($t_{56} = 1.28$, $P = 0.20$, $d = 0.19$, 95% CI = −0.07–0.31; Fig. 4b; additional gaze switch analyses within each image are provided in Supplementary Figs. 1 and 2b).

A post hoc analysis showed that the spontaneous behaviours of the toddlers that suggested uncertainty varied as a function of trial type (Supplementary Results 4).

Discussion

Here we investigated whether toddlers exhibit behaviours that might reflect responses to states of uncertainty and information gathering before committing to a decision. We have proposed that toddlers' overt decisions may depend on implicit error signals that manifest in their behaviours during the time between stimulus presentation and response selection. These signals provide cues for uncertainty reports in older children and adults[15] and may serve as precursors of metacognitive abilities if young children learn to attend to them during cognitive tasks, or they might reflect core metacognitive abilities in infants and young children[33]. We devised a task that requires little instruction. By administering yoked versions of the task using an eye-tracking and a touchscreen system, we could characterized children's information gathering and hesitation during perceptual discrimination and the corresponding decisions. We extend previous research that examined either infants' spontaneous signals of evidence accumulation with no overt decision[13,34] or overt behavioural responses without examination of intervening processes[12,35].

Toddlers responded more slowly during incorrect responses compared with during correct responses, suggesting that they were sensitive to decision difficulty. This finding is consistent with recent research showing that even 18-months-olds can distinguish the accuracy of their responses by searching longer for a hidden toy in the correct location compared with the incorrect location[13]. We did not observe any differences in response latencies between similar and dissimilar correct responses. Although this result might suggest that toddlers could only slow their responding when they generally failed to identify an appropriate response, results from the drift–diffusion model provided a more nuanced picture. The drift rate parameter was lower in similar trials compared with dissimilar trials, consistent with toddlers extracting lower-quality evidence under conditions of reduced discriminability (and possibly higher uncertainty). However, we also found greater values for the separation parameter with dissimilar trials, suggesting that more evidence

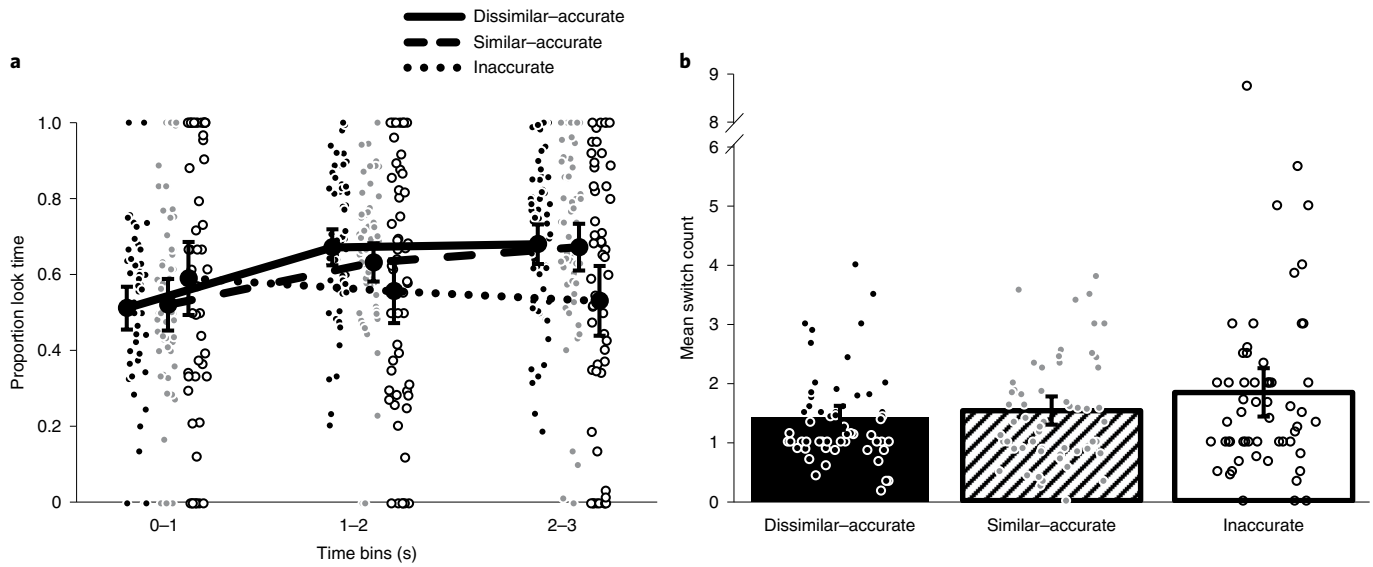


Fig. 4 | The proportion of looking times and switch counts for the eye-tracker task of experiment 2. a, The proportion of looking time towards the chosen image for the first 3 s of the trial in 1 s time bins for the dissimilar–accurate, similar–accurate and inaccurate trials; $n = 61$ toddlers (bin 1 dissimilar–accurate versus inaccurate trials: $b = -0.09$, $P = 0.05$, $P_{Bonf} = 0.14$, $d = -0.07$, 95% CI = −0.18–0.00; bin 1 similar–accurate versus inaccurate trials: $b = -0.08$, $P = 0.10$, $P_{Bonf} = 0.30$, $d = -0.06$, 95% CI = −0.17–0.01; bin 1 dissimilar–accurate versus similar–accurate trials: $b = -0.01$, $P = 0.68$, $P_{Bonf} = 1.00$, $d = -0.01$, 95% CI = −0.09–0.06; bin 2 dissimilar–accurate versus inaccurate trials: $b = 0.11$, $P = 0.003$, $P_{Bonf} = 0.01$, $d = 0.10$, 95% CI = 0.04–0.18; bin 3 dissimilar–accurate versus inaccurate trials: $b = 0.12$, $P = 0.005$, $P_{Bonf} = 0.01$, $d = 0.10$, 95% CI = 0.03–0.20; bin 3 similar–accurate versus inaccurate trials: $b = 0.12$, $P = 0.006$, $P_{Bonf} = 0.02$, $d = 0.10$, 95% CI = 0.04–0.20; bin 3 dissimilar–accurate versus similar–accurate trials: $b = 0.002$, $P = 0.96$, $P_{Bonf} = 1.00$, $d = 0.002$, 95% CI = −0.06–0.07). 0 s is the stimulus onset. **b**, The mean number of gaze switches for the dissimilar–accurate, similar–accurate and inaccurate trials; $n = 61$ toddlers (dissimilar–accurate versus inaccurate trials: $t_{56} = -2.21$, $P = 0.03$, $d = -0.29$, 95% CI = −0.82 to −0.04; similar–accurate versus inaccurate trials: $t_{56} = -1.44$, $P = 0.15$, $d = -0.20$, 95% CI = −0.74–0.12; dissimilar–accurate versus similar–accurate trials: $t_{56} = 1.28$, $P = 0.20$, $d = 0.19$, 95% CI = −0.07–0.31). The *y* axis is broken between 6 and 8 gaze switches. The points represent individual data points. The average values of the proportion of look time graph and data points on both graphs are jittered on the *x* axis to avoid stacking. Error bars are 95% confidence intervals.

was necessary to endorse a response on dissimilar trials; this finding might indicate that, when evidence is expected to be accessible and clear, as is the case with dissimilar trials, responses are rendered after thorough examination. Thus, the contribution of these opposing processes explains why comparable average response latencies were observed across conditions and underscores the use of parameter estimates. The pattern of results for drift and separation parameters is strikingly consistent with results observed in adults and older children facing a lower or higher level of decision difficulty[26]. The separation parameter has been shown to be sensitive to other manipulations resulting in different levels of accuracy, consistent with the idea that when a higher level of evidence is expected, the separation parameter is higher compared with when a lower level of evidence is expected. For example, when accuracy is emphasized, the separation parameter is greater compared with when speed is emphasized, which results in lower overall accuracy[36,37]. Our task was self-paced and the length of overall response times did not suggest that toddlers interpreted the task as emphasizing speed. Future research should nevertheless examine the emergence of speed–accuracy trade-offs in early childhood.

We acknowledge that we included a much smaller number of trials per participant than is recommended for drift–diffusion modelling[30]. The confirmation that the pattern of results at the individual level matches the pattern at the sample level suggests that the results were still reliable, even with fewer trials. Furthermore, the replication of our results across experiments provides some reassurance that the measure captured meaningful information. Finally, we also conducted a HDDM, which enabled us to obtain estimates from a small number of trials per participant. The HDDM model confirmed our pattern of results, revealing higher drift and separation parameters for the dissimilar trials.

We also explored whether drift parameters could be used as an individual difference measure to predict toddlers' use of language indicating ignorance or uncertainty[31] (that is, 'I don't know'). In experiment 1, we found that drift rate positively predicted frequency of 'I don't know'. This result suggests that toddlers who are more efficient at extracting evidence also show a greater ability to express their states of ignorance, linking implicit behavioural signals with overt expressions of uncertainty. However, this result was not replicated in experiment 2 (an alternative model is shown in Supplementary Table 3 that suggests instead an association with the separation boundary parameter). It is also possible that, contrary to our initial hypothesis, the separation parameter capturing the evidence necessary to commit to a decision may be more meaningfully associated with behaviours that signal uncertainty than the drift parameter. Our post hoc examination of the association between drift–diffusion parameters and the language of general mental states—above and beyond general vocabulary—in experiment 2, as well as our exploration of the association between drift–diffusion parameters and the spontaneous uncertainty behaviours performed during the session, are suggestive of this alternative possibility. Future research should examine this further.

Future research should also use multiple sessions to increase the number of trials and test whether this provides a stronger basis for assessment of associations with other variables including the use of 'I don't know' in toddlers. It is nevertheless possible that the relation between the two constructs is tenuous at this age, or it encompasses decision processes more generally (instead of evidence accumulation specifically) regardless of the reliability of the estimated parameters.

The toddlers' looking behaviours also suggested information seeking. Although toddlers looked longer overall at the image

that they ultimately selected compared with the non-selected image (Extended Data Fig. 2), it took them longer to settle on the selected picture during inaccurate or similar trials. Toddlers also demonstrated more gaze switches between picture pairs for similar pictures. In the adult literature, it has been shown that the gaze switches between response options while making decisions are associated with lower confidence in the decision[18]. Adults will therefore actively try to gather more information about a decision that they are uncertain about. Similarly, in our experiments, toddlers seem to actively gather extra information when faced with uncertain decisions, owing to the lower quality of available evidence, before committing to a decision. However, gaze switching could also initially reflect uncertainty and then promote the active process of gathering information. Future research is needed to disentangle these possibilities.Overall, we showed that, while deliberating, toddlers exhibit a repertoire of behaviours that reflect gathering and assessment of evidence and decision-making processes, which are also engaged in overt uncertainty monitoring[27,38]. The developmental process that links these behavioural responses in toddlers to a later emerging ability to experience and report on subjective uncertainty is unclear at present. We recognize that these behaviours have not been uniformly accepted as indicators of metacognitive processes. For example, these behavioural indicators may have stemmed from associative learning in non-human animals and, therefore, require little to no access to representations of uncertainty[9]. Furthermore, it has been suggested that true metacognition can only be identified when children possess conceptual prerequisites about uncertainty[39]. Nevertheless, research in adults and children shows that response times and eye movements–and the underlying processes—are used as cues for explicit reports of uncertainty[19,38,40]. Indeed, current theories of metacognition in adults posit that decision confidence is a second-order judgement that is based on the quality and quantity of evidence, computed separately from the decision itself[41,42]. Future research is needed to link the signatures of behavioural hesitation and information gathering observed here to the explicit uncertainty monitoring that is observed just a year later[1].

In summary, we utilized a perceptual discrimination paradigm with a similarity manipulation that was hypothesized to elicit more decision uncertainty for similar trials compared with dissimilar trials. This paradigm enabled us to examine toddlers' behaviours before they made an overt decision. The ability to delay responding when uncertain and selectively gather disambiguating information may bolster young children's rapid early learning across domains.

Methods

Our experiments were approved by the Institutional Review Board of the University of California, Davis. Informed consent was obtained from all of the parents, and the toddlers received a small book after each visit as a token of appreciation. On the basis of pilot data on an independent sample and previous research investigating perceptual similarity in toddlers[34], we anticipated a medium effect size of similarity in the current studies. Both samples were sufficient to detect a small-to-medium main effect ($F = 0.15$, $\eta_p^2 = 0.02$) with 80% power and an alpha of 0.05 between our three conditions (similar–accurate, dissimilar–accurate and inaccurate).

Participants. *Experiment 1.* Eighty toddlers (aged 25–34 months; mean, 28.85 months; 39 females) participated in this study. The household incomes of the families were as follows: less than US$15,000 ($n = 1$), US$15,000–25,000 ($n = 4$), US$25,000–40,000 ($n = 7$), US$40,000–60,000 ($n = 16$), US$60,000–90,000 ($n = 16$), more than US$90,000 ($n = 33$) and unreported ($n = 3$). Four toddlers were reported by their parents to be African American, 10 were Asian, 4 were American Indian or Alaskan, 56 were Caucasian, 3 were Native Hawaiian and the parents of 3 did not report a race. This sample included the first half of a larger longitudinal study investigating the development of uncertainty monitoring. Families were recruited from a local northern California community.

Some of the toddlers completed only one of the tasks owing to inattention in the other; six did not complete the eye-tracker task and six did not complete the touchscreen task. Of the remaining toddlers, one performed below chance in both the eye-tracker task and the touchscreen task and two performed below chance in the touchscreen task, leaving 73 participants in the eye-tracker analyses and 71 participants in the touchscreen analyses.

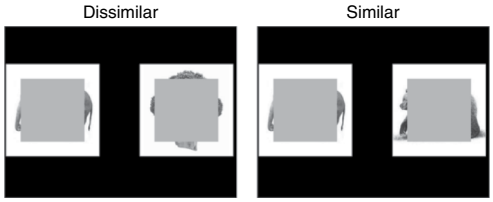


Fig. 5 | Example of dissimilar and similar trials. Children were asked to point to (eye-tracker task) or touch (touchscreen task) a target item depicted in one of the drawings (for example, 'where is the elephant?').

Experiment 2. Experiment 2 involved a preregistered replication of experiment 1. The replication experiment was preregistered on the Open Science Framework and can be found at https://osf.io/nvf2j/. This study was preregistered on 31 July 2018.

Eighty toddlers (aged 25–33 months; mean, 29.09 months; 43 Females) participated in this study. The household incomes of the families were as follows: less than US$15,000 ($n = 3$), US$25,000–40,000 ($n = 9$), US$40,000–60,000 ($n = 8$), US$60,000–90,000 ($n = 13$), more than US$90,000 ($n = 46$) and unreported ($n = 1$). Eight toddlers were reported by their families to be African American, 14 were Asian, 1 was American Indian or Alaskan, 51 were Caucasian and the parents of 6 did not report a race.

Some of the toddlers completed only one of the tasks owing to being uncooperative in the other or computer error; 13 did not complete the eye-tracker task (11 were uncooperative and 2 due to computer error) and 6 did not complete the touchscreen task (6 were uncooperative). Of the remaining toddlers, 6 performed below chance in the eye-tracker task and 7 performed below chance in the touchscreen task, leaving 61 participants in the eye-tracker analyses and 67 participants in the touchscreen analyses.

Materials. Both experiments used the same materials and procedure.

Perceptual task. The present task was a perceptual discrimination task administered on both a touchscreen monitor and an eye-tracker monitor. We collected data using these two methods during the same task to obtain precise response latencies (using a touchscreen version) and eye movement data (using an eye-tracker version). Toddlers touched the screen on the touchscreen task and pointed to their selected image on the eye-tracker task. Stimuli included photographs of common objects and animals[43], which were presented partially occluded by grey squares superimposed over the centre of each image (Fig. 5). Stimuli were selected on the basis of age-of-acquisition norms[44], and we verified with parents whether their child was familiar with the stimuli labels; the rare trials with reportedly unfamiliar labels were removed from the analyses.

During each trial, toddlers saw two of the occluded images side by side. They were required to select the target image by touching it (touchscreen task) or by pointing to it (eye-tracker task). There were 20 trials in each task—10 trials of perceptually and semantically similar items (for example, an elephant and a bear) and 10 trials of perceptually and semantically dissimilar items (for example, an elephant and broccoli; Fig. 5). Similar and dissimilar trials were presented in a pseudorandom order and the allocation of individual stimuli to similar or dissimilar trials was counterbalanced. Details of all of the possible stimuli pairs are provided at https://osf.io/t8p4g/. The eye-tracker and touchscreen tasks were yoked such that the same stimuli pairs were presented on both, but with different exemplars (for example, an elephant may be paired with a bear on both tasks, but the photographs depicted two different elephants and bears). This design was meant to equate processing demands across tasks while limiting practice effects.

Language measures. Parents were asked to report on a five-point scale, from 'never/not yet' to 'often' how frequently the toddler uses the phrase 'I don't know' in every day conversations. Parents completed this assessment during their laboratory visit. To assess general language ability, we used the MacArthur Bates Communicative Development Inventory-III (ref. [45]). This questionnaire asks parents about their toddler's vocabulary, grammar, semantics, pragmatics and comprehension. For this study, we used the toddler's vocabulary score. Parents marked which of 100 words toddlers verbally express and we calculated the proportion of words the toddler expressed by dividing the number of checked off words by 100.

Procedure. Participation involved two visits, scheduled one week apart. On each visit, an experimenter played with the child for about 5 min outside the testing rooms until the child was comfortable before beginning the research tasks. Participants completed the eye-tracker task on the first visit and the touchscreen task on the second visit. All of the participants completed the tasks in this order, because pilot testing revealed that toddlers who experienced the touchscreen task first were more prone to leaning forward to touch the eye-tracker monitor, which interfered with successful data recording.

The touchscreen task. The touchscreen task was administered on an upright monitor. Toddlers sat in front of the monitor on a child-sized chair and were instructed to respond by touching the screen. Just before starting the task, toddlers were told "Now we are going to find some things that are hiding behind boxes!". For each trial, the experimenter said "Can you find the (target)?" on a blank screen and then pressed a button to present the trial. As soon as the child touched a side of the screen, the task advanced to a blank screen before starting the next trial. If a child refused to respond, the experimenter keyed in a separate code and moved on to the next trial. Response latencies were taken from the touchscreen task.

The eye-tracker task. The eye-tracker task was administered on a Tobii T-120 17 inch eye-tracker monitor. Toddlers sat on their parents' laps approximately 60 cm from the monitor. The stimuli were 10 cm × 10 cm (visual angle, 9.53°) with 4.45 cm (visual angle, 4.24°) between them. Tobii Studio's standard infant calibration was used; a cartoon cat was presented at five points on the screen accompanied by a sound effect. The experiment proceeded when the gaze of the toddlers was captured at all five points. Default Tobii fixation filter settings were used for eye-movement data reduction (velocity threshold, 35 px per sample; distance threshold, 35 px; minimum fixation duration, 83 ms). Parents wore dark sunglasses to ensure that their eye movements were not unintentionally recorded. Parents were instructed to refrain from speaking to their child during the task, and to hold their child to prevent them from leaning forward or moving excessively.

The eye-tracker task was identical to the touchscreen task, except that the toddler pointed to the chosen image and the experimenter immediately keyed in their response, advancing the task to a blank screen with a fixation cross in the middle. As in the touchscreen task, the experimenter keyed in a separate code if a child refused to respond. Looking times and gaze switches were taken from the eye-tracking task.

Analytical approach. The similarity manipulation was apparent to the experimenter collecting the data and, as such, data collection was not performed in a blinded manner. However, the experimenters were blinded to the hypotheses tested in the study. Data analyses were not performed blinded to the conditions of the experiments. Data distribution was assumed to be normal but this was not formally tested.

Drift–diffusion modelling. We tested our drift–diffusion models using the RStudio (v.3.3.1, 2016) package RWiener[46] using response times generated from the touchscreen task. Each trial condition (similar and dissimilar) was fit separately with response thresholds representing accuracy (accurate versus inaccurate). We estimated a total of six parameters as follows: drift parameter (v), separation parameter (a) and non-decision parameter (t_0) for the similar and dissimilar trials. The start point parameter z_0 was fixed to 0.5, as there were no a priori differences that should occur for evidence accumulation towards a correct versus incorrect response. As the number of trials per participant was small compared with what is typically used to fit drift–diffusion models[30], our initial analysis estimated parameters using all trials across the entire sample. Fit was assessed by comparing AIC and BIC fit indices and using a χ^2 difference test to compare our main model to a baseline model (model with no similarity conditions). We then fit the models for each individual participant and assessed whether the average of these individual participants would yield similar findings to that of the initial sample-level model. Furthermore, we confirmed the results by utilizing a HDDM, which is robust to a small number of trials per participant (Supplementary Information).

Statistical analyses. All statistical tests used were two-sided. We tested our multilevel models for looking times using the RStudio (v.3.3.1, 2016) package nlme[47]. By utilizing a multilevel model, we were able to examine trial-level data and avoid casewise deletions that would have occurred due to participants not having data for all combinations of time bins and trial types. The multilevel model included fixed effects of time bin (0–1 s, 1–2 s and 2–3 s, dummy coded against 0–1 s) and trial type (inaccurate, dissimilar–accurate and similar–accurate, dummy coded against inaccurate) and a random effect of individual.

Data processing. *Experiment 1.* Before data analysis, trials were removed for which no answer was provided. This resulted in 13 (0.88%) total trials across 8 participants for the touchscreen task and 33 (2.23%) trials total across 16 participants for the eye-tracker task. We also eliminated trials for which parents reported that their child was not familiar with the target word. This resulted in a total of 51 (3.45%) trials across 26 participants in the touchscreen task and a total of 56 (3.78%) trials from 26 participants in the eye-tracker task.

Once these trials were removed, the accuracy in each task was calculated and any participants with a chance accuracy of 0.50 or below were eliminated from the analyses. One participant performed below chance (<0.50) on both tasks, and two performed below chance on the touchscreen task.

For response latency analyses, we removed any trials that were probably fast responding errors (that is, responses produced before processing the stimuli and trials during which the participant was inattentive). The criteria for these eliminations were trials with response latencies of less than 700 ms in duration and trials in which the z-scored response latencies across each individual participant were ±3 s.d. The 700 ms cut-off was determined on the basis of research that indicated that the average time for 5-year-old children to make a motor response is approximately 750 ms (ref. [48]). Given that this experiment involved a younger sample, we deemed that 700 ms was a reasonable response latency cut-off. This resulted in 35 (2.36%) trials across 35 participants being eliminated. Short-response-latency removals included 2 incorrect similar trials and 1 correct dissimilar trial. Long-response-latency removals included 5 correct dissimilar trials, 10 correct similar trials, 4 incorrect dissimilar trials and 13 incorrect similar trials. When analyses were conducted with these trials included, the results remained unchanged. For the drift–diffusion model estimations only, trials with response latencies of greater than 15 s were also removed on the basis of previous research showing that drift–diffusion models can be used reliably for response latencies up to 15 s (ref.[30]); 18 (1.23%) trials across 9 participants were eliminated due to this criterion. Furthermore, individual drift–diffusion parameters were considered to be outliers if z-scores were ±3 s.d. or more and were removed from analyses. This resulted in 8 parameter values across 4 participants (1 separation parameter, 1 drift parameter, 1 similar separation parameter, 1 similar drift parameter, 2 dissimilar separation parameters and 2 dissimilar drift parameters).

For all eye movement analyses, we used Tobii Studio software to create areas of interest (AOIs). These AOIs encompassed individual square images surrounding targets and lures, so that each trial had a target AOI and a lure AOI. For the looking time measure, we calculated the total proportion of looking time towards one of the AOIs in time bins of 1 s. For accurate trials, this was calculated as the total duration of looking time towards the target AOI in that time window divided by the total duration of looking time towards both AOIs. For inaccurate trials, the proportion was the total duration of looking time towards the lure AOI in the time window divided by the total duration of looking towards both AOIs. We were interested in the toddlers' looking patterns during the decision process and before committing to a response; as the toddlers had an average response latency of 3.8 s we examined the first 3 s of the trial. We calculated gaze switches from one stimulus to the other during a trial and defined a switch as a fixation on an AOI that was preceded by a fixation on the other AOI, including cases in which there were fixations on other (non-AOI) areas of the screen in between fixations to AOIs. The first fixation to an AOI in a trial was not counted as a switch, so the minimum number of switches in a trial was zero. There was a subset of trials in which no fixations to AOIs were recorded—a total of 158 trials (10.68%) from 36 participants. These trials were excluded from the analyses.

Experiment 2. The same process of trial eliminations as described experiment 1 was followed in experiment 2. A total of 13 (0.88%) trials across 8 participants for the touchscreen task and a total of 14 (1.06%) trials across 5 participants in the eye-tracker task were removed due to participants not providing answers. A total of 61 (4.12%) trials across 28 participants in the touchscreen task and 52 (3.95%) trials from 26 participants in the eye-tracker task were eliminated on the basis of the participants not knowing the target word. Seven participants were removed due to performing below chance on the touchscreen task, and six were removed for performing below chance on the eye-tracker task. For response-latency analyses, 38 (2.57%) trials across 37 participants were eliminated. Short-response-latency removals included one incorrect similar trial and one correct dissimilar trial. Long-response-latency removals included 16 correct dissimilar trials, 14 correct similar trials, 2 incorrect dissimilar trials and 4 incorrect similar trials. In the drift–diffusion analyses, 28 (1.89%) trials across 13 participants were removed for having response latencies longer than 15 s. Moreover, 9 parameter values across 6 participants were eliminated for being outliers (1 separation parameter, 1 drift parameter, 2 similar separation parameters, 2 similar drift parameters, 1 dissimilar separation parameter, 1 dissimilar drift parameter and 1 dissimilar non-decision parameter). Similar to experiment 1, we chose to analyse the first 3 s of trials for look time analyses as the average response time was 4.6 s. Finally, a total of 178 (13.52%) trials from 37 participants was eliminated from eye movement analyses owing to participants not looking at either AOI during the trial.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated and analysed during the current studies are available at the Open Science Framework repository (https://osf.io/t8p4g/).

Code availability

The code generated and used during the current studies are available at the Open Science Framework repository (https://osf.io/t8p4g/).

Received: 3 July 2019; Accepted: 11 June 2020;
Published online: 20 July 2020

References

1. Coughlin, C., Hembacher, E., Lyons, K. E. & Ghetti, S. Introspection on uncertainty and judicious help-seeking during the preschool years. *Dev. Sci.* **18**, 957–971 (2015).

2. Jaswal, V. K., Croft, A. C., Setia, A. R. & Cole, C. A. Young children have a specific, highly robust bias to trust testimony. *Psychol. Sci.* **21**, 1541–1547 (2010).
3. Schulz, L. E. & Bonawitz, E. B. Serious fun: preschoolers engage in more exploratory play when evidence is confounded. *Dev. Psychol.* **43**, 1045–1050 (2007).
4. Stahl, A. E. & Feigenson, L. Observing the unexpected enhances infants' learning and exploration. *Science* **348**, 91–94 (2015).
5. Balcomb, F. K. & Gerken, L. Three-year-old children can access their own memory to guide responses on a visual matching task. *Dev. Sci.* **11**, 750–760 (2008).
6. Lyons, K. E. & Ghetti, S. The development of uncertainty monitoring in early childhood. *Child Dev.* **82**, 1778–1787 (2011).
7. Nelson, T. O. & Narens, L. in *Metacognition: Knowing About Knowing* vol. 13 (eds Metcalfe, J. & Shimamura, A. P.) 1–25 (MIT Press, 1994).
8. Dehaene, S., Lau, H. & Kouider, S. What is consciousness, and could machines have it? *Science* **358**, 486–492 (2017).
9. Hampton, R. R. Multiple demonstrations of metacognition in nonhumans: converging evidence or multiple mechanisms? *Comp. Cogn. Behav. Rev.* **4**, 17–28 (2009).
10. Smith, J. D. & Washburn, D. A. Uncertainty monitoring and metacognition by animals. *Curr. Dir. Psychol. Sci.* **14**, 19–24 (2005).
11. Metcalfe, J. in *Handbook of Metamemory and Memory* (eds Dunlosky, J. & Bjork, R. A.) 29–46 (Psychology Press, 2008).
12. Goupil, L., Romand-Monnier, M. & Kouider, S. Infants ask for help when they know they don't know. *Proc. Natl Acad. Sci. USA* **113**, 3492–3496 (2016).
13. Goupil, L. & Kouider, S. Behavioral and neural indices of metacognitive sensitivity in preverbal infants. *Curr. Biol.* **26**, 3038–3045 (2016).
14. Desender, K., Boldt, A. & Yeung, N. Subjective confidence predicts information seeking in decision making. *Psychol. Sci.* **29**, 761–778 (2018).
15. Fandakova, Y. et al. Changes in ventromedial prefrontal and insular cortex support the development of metamemory from childhood into adolescence. *Proc. Natl Acad. Sci. USA* **114**, 7582–7587 (2017).
16. Ratcliff, R. & McKoon, G. The diffusion decision model: theory and data for two choice decision tasks. *Neural Comput.* **20**, 873–922 (2008).
17. Roderer, T. & Roebers, C. M. Can you see me thinking (about my answers)? Using eye-tracking to illuminate developmental differences in monitoring and control skills and their relation to performance. *Metacogn. Learn.* **9**, 1–23 (2014).
18. Folke, T., Jacobsen, C., Fleming, S. M. & De Martino, B. Explicit representation of confidence informs future value-based decisions. *Nat. Hum. Behav.* **1**, 0002 (2016).
19. Ackerman, R. & Koriat, A. Response latency as a predictor of the accuracy of children's reports. *J. Exp. Psychol. Appl.* **17**, 406–417 (2011).
20. Lyons, K. E. & Ghetti, S. I don't want to pick! Introspection on uncertainty supports early strategic behavior. *Child Dev.* **84**, 726–736 (2013).
21. Pleskac, T. J. & Busemeyer, J. R. Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol. Rev.* **117**, 864–901 (2010).
22. Heitz, R. P. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Front. Neurosci.* **8**, 150 (2014).
23. Ratcliff, R., Gomez, P. & McKoon, G. A diffusion model account of the lexical decision task. *Psychol. Rev.* **111**, 159–182 (2004).
24. Metin, B. et al. ADHD performance reflects inefficient but not impulsive information processing: a diffusion model analysis. *Neuropsychology* **27**, 193–200 (2013).
25. Mulder, M. J. et al. Basic impairments in regulating the speed-accuracy tradeoff predict symptoms of attention-deficit/hyperactivity disorder. *Biol. Psychiatry* **68**, 1114–1119 (2010).
26. Ratcliff, R., Love, J., Thompson, C. A. & Opfer, J. E. Children are not like older adults: a diffusion model analysis of developmental changes in speeded responses. *Child Dev.* **83**, 367–381 (2012).
27. Roderer, T. & Roebers, C. M. Explicit and implicit confidence judgments and developmental differences in metamemory: an eye-tracking approach. *Metacogn. Learn.* **5**, 229–250 (2010).
28. Gehring, W. J., Coles, M. G., Meyer, D. E. & Donchin, E. A brain potential manifestation of error-related processing. *Electroencephalogr. Clin. Neurophysiol. Suppl.* **44**, 261–272 (1995).
29. Aitken, P. P. & Hutt, C. The effects of stimulus incongruity upon children's attention, choice, and expressed preference. *J. Exp. Child Psychol.* **19**, 79–87 (1975).
30. Lerche, V., Voss, A. & Nagler, M. How many trials are required for parameter estimation in diffusion modeling? A comparison of different optimization criteria. *Behav. Res. Methods* **49**, 513–537 (2017).
31. Harris, P. L., Bartz, D. T. & Rowe, M. L. Young children communicate their ignorance and ask questions. *Proc. Natl Acad. Sci. USA* **114**, 7884–7891 (2017).
32. Vandekerckhove, J., Tuerlinckx, F. & Lee, M. D. Hierarchical diffusion models for two choice response times. *Psychol. Methods* **16**, 44–62 (2011).
33. Goupil, L. & Kouider, S. Developing a reflective mind: from core metacognition to explicit self-reflection. *Curr. Dir. Psychol. Sci.* **28**, 403–408 (2019).
34. Arias-Trejo, N. & Plunkett, K. The effects of perceptual similarity and category membership on early word-referent identification. *J. Exp. Child Psychol.* **105**, 63–80 (2010).
35. Geurten, M. & Bastin, C. Behaviors speak louder than explicit reports: implicit metacognition in 2.5-year-old children. *Dev. Sci.* **22**, e12742 (2018).
36. Voss, A., Rothermund, K. & Voss, J. Interpreting the parameters of the diffusion model: an empirical validation. *Mem. Cogn.* **32**, 1206–1220 (2004).
37. Voss, A., Nagler, M. & Lerche, V. Diffusion models in experimental psychology: a practical introduction. *Exp. Psychol.* **60**, 385–402 (2013).
38. Koriat, A. & Ackerman, R. Choice latency as a cue for children's subjective confidence in the correctness of their answers. *Dev. Sci.* **13**, 441–453 (2010).
39. Carruthers, P. Basic questions. *Mind Lang.* **33**, 130–147 (2018).
40. Urgolites, Z. J., Smith, C. N. & Squire, L. R. Eye movements support the link between conscious memory and medial temporal lobe function. *Proc. Natl Acad. Sci. USA* **115**, 7599–7604 (2018).
41. Fleming, S. M. & Daw, N. D. Self-evaluation of decision-making: a general Bayesian framework for metacognitive computation. *Psychol. Rev.* **124**, 91–114 (2017).
42. Yeung, N. & Summerfield, C. Metacognition in human decision-making: confidence and error monitoring. *Philos. Trans. R. Soc. B* **367**, 1310–1321 (2012).
43. Yassa, M. A. et al. Pattern separation deficits associated with increased hippocampal CA3 and dentate gyrus activity in nondemented older adults. *Hippocampus* **21**, 968–979 (2011).
44. Kuperman, V., Stadthagen-Gonzalez, H. & Brysbaert, M. Age-of-acquisition ratings for 30,000 English words. *Behav. Res. Methods* **44**, 978–990 (2012).
45. Fenson, L. et al. Reply: measuring variability in early child language: don't shoot the messenger. *Child Dev.* **71**, 323–328 (2000).
46. Wabersich, D., & Vandekerckhove, J. The RWiener Package: an R package providing distribution functions for the Wiener diffusion model. *R J.* **6**, 49–56 (2014).
47. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & R Core Team nlme: linear and nonlinear mixed effects models. R package version 3.1-137 https://cran.r-project.org/web/packages/nlme/nlme.pdf (2018).
48. Droit-Volet, S. Stop using time reproduction tasks in a comparative perspective without further analyses of the role of the motor response: the example of children. *Eur. J. Cogn. Psychol.* **22**, 130–148 (2010).

Acknowledgements

This research was supported by a grant from the National Science Foundation (NSF; BCS1424058) to S.G. Any opinions, findings and conclusions or recommendations expressed in this manuscript are those of the authors and do not necessarily reflect the views of the NSF. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

S.G. developed the study concept. S.G., S.L. and E.H. finalized the study design. S.L. and E.H. performed data collection. S.L., D.S., A.K. and E.G.J. contributed to data analysis and interpretation under the supervision of S.G. S.L., D.S., E.H. and S.G. drafted the manuscript. All of the authors provided revisions and approved the final version of the manuscript for submission.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-020-0913-y.

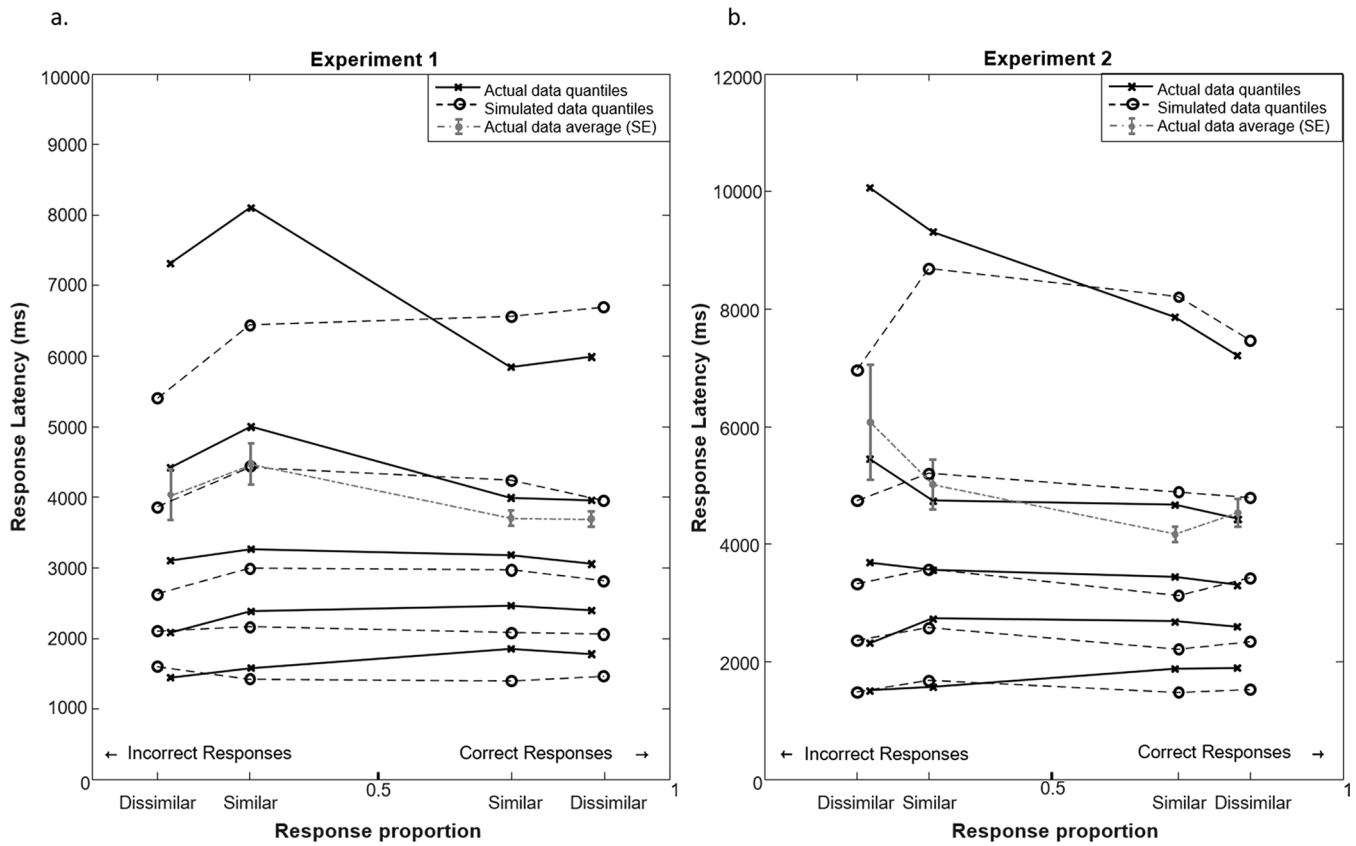
Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-0913-y.

Correspondence and requests for materials should be addressed to S.L. or S.G.

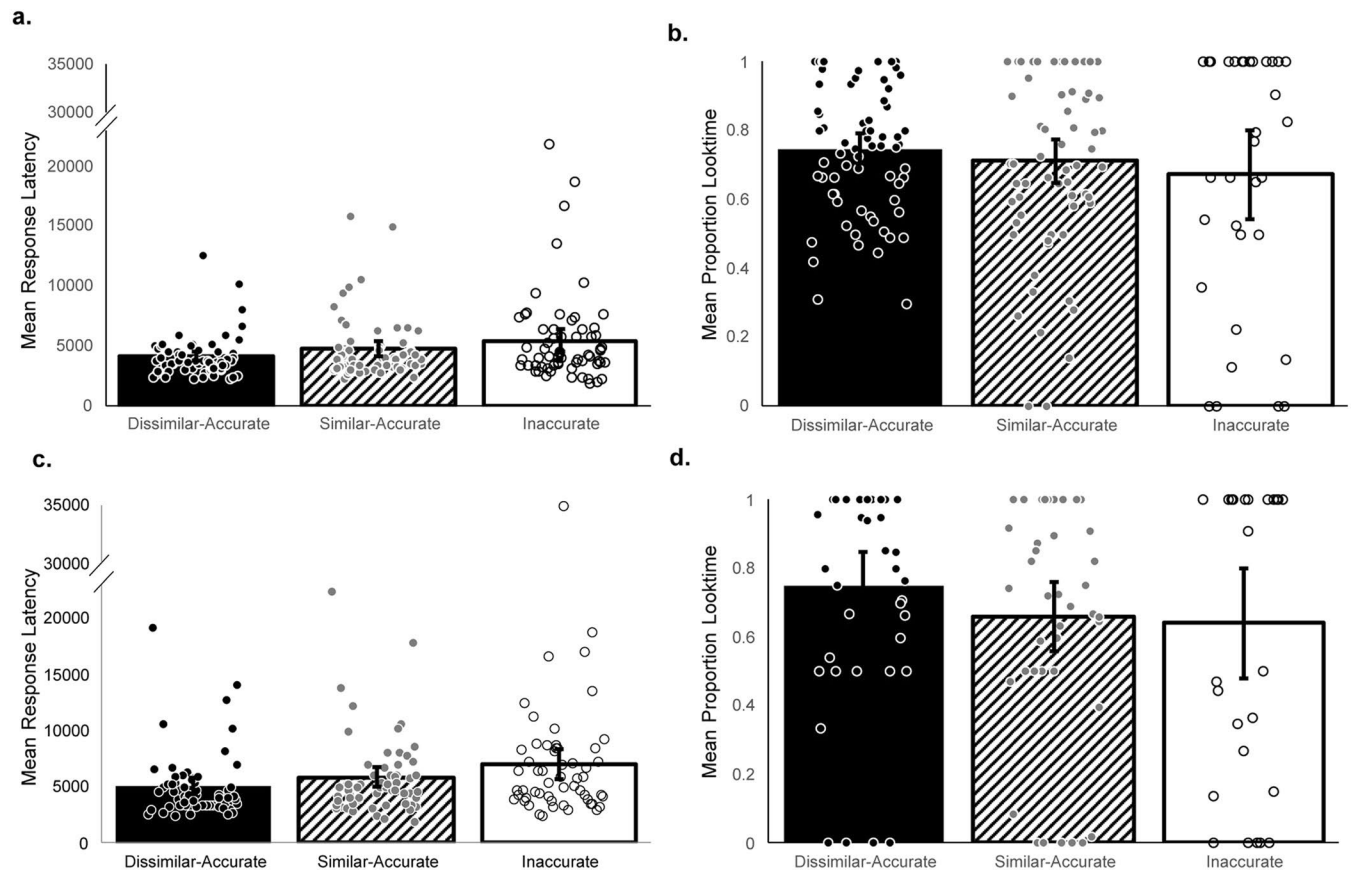
Peer review information Primary Handling Editor: Marike Schiffer.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Extended Data Fig. 1 | Quantile plots for drift-diffusion model. Lines with x markers are plotted based on observed data and dashed lines with o markers are the simulated data produced by our complete model in Experiment 1 (**a**) and Experiment 2 (**b**). Graphs show the .1, .3, .5 (median), .7, and .9 quantiles (stacked vertically) plotted against response proportion for each of the two conditions (dissimilar and similar). Similar/Dissimilar labels are placed at the level on x-axis corresponding to the response proportion for that type of trial. Correct response proportions are plotted to the right, and incorrect response proportions are plotted to the left. Predicted values qualitatively resemble observed values, indicating good fit of our drift-diffusion models to the data.



Extended Data Fig. 2 | Response latencies and mean proportion looktime in the eye tracker task. Mean response latencies for dissimilar-accurate, similar-accurate, and inaccurate trials for Experiment 1 (**a**) and Experiment 2 (**c**). Mean proportion looktime for the time bin prior to the average response latency for dissimilar-accurate (2–3 seconds for Experiment 1, 4-5 seconds for Experiment 2), similar-accurate (3-4 seconds for Experiment 1, 4-5 seconds for Experiment 2), and inaccurate (4-5 seconds in Experiment 1, 5-6 seconds for Experiment 2) trials for Experiment 1 (**b**) and Experiment 2 (**d**). Points represent individual data points. Data points on both graphs are jittered on the horizontal axis to avoid stacking. Error bars are 95 percent confidence intervals.

Effect of Similarity on Reaction Time Inaccurate Trials

a

								95% CI	
		SD	b	SE	df	*d*	*t*	*p*	Lower	Upper
Experiment 1										
Random Effect										
	Intercept	1881.222							1437.824	2461.355
Fixed Effect										
	Intercept		4103.318	427.499	202	0.58	9.59	<.001	3260.385	4946.251
	Similar		307.439	422.931	202	0.04	0.73	0.46	-526.486	1141.364

b

								95% CI	
		SD	b	SE	df	*d*	*t*	*p*	Lower	Upper
Experiment 2										
Random Effect										
	Intercept	3186.005							2267.415	4476.739
Fixed Effect										
	Intercept		6008.099	812.630	229	0.49	7.39	<.001	4406.911	7609.288
	Similar		-899.329	885.559	229	-0.07	-1.02	0.311	-2644.214	845.556

Extended Data Fig. 3 | Effect of Similarity on Reaction Time Inaccurate Trials. Multilevel model results showing the effect of similarity on reaction times for inaccurate trials for Experiment 1 (**a**) and Experiment 2 (**b**). Results displayed here are for models dummy coded relative to dissimilar-inaccurate trial type.

Looking Time Multilevel Model

a

	SD	b	SE	df	*d*	*t*	*p*	95% CI Lower	95% CI Upper
Experiment 1									
Random Effect									
Intercept	0.053							0.035	0.082
Fixed Effect									
Intercept		0.580	0.032	2991	0.327	18.05	<.001	0.517	0.643
Dissimilar Accurate		-0.036	0.037	2991	-0.018	-0.98	0.33	-0.109	0.037
Similar Accurate		-0.086	0.038	2991	-0.041	-2.28	0.02	-0.160	-0.012
Time Bin 2		-0.017	0.044	2991	-0.007	-0.39	0.70	-0.103	0.069
Time Bin 3		-0.017	0.046	2991	-0.007	-0.38	0.71	-0.107	0.073
Dissimilar Accurate*Time Bin 2		0.164	0.052	2991	0.057	3.17	0.002	0.063	0.265
Dissimilar Accurate*Time Bin 3		0.206	0.054	2991	0.069	3.81	<.001	0.100	0.312
Similar Accurate*Time Bin 2		0.154	0.052	2991	0.053	2.94	0.003	0.052	0.257
Similar Accurate*Time Bin 3		0.194	0.055	2991	0.064	3.55	<.001	0.087	0.301

b

Experiment 2									
Random Effect									
Intercept	0.064							0.045	0.092
Fixed Effect									
Intercept		0.627	0.035	2311	0.367	17.71	<.001	0.558	0.697
Dissimilar Accurate		-0.093	0.041	2311	-0.047	-2.27	0.02	-0.174	-0.013
Similar Accurate		-0.081	0.042	2311	-0.040	-1.93	0.05	-0.162	0.001
Time Bin 2		-0.037	0.048	2311	-0.016	-0.78	0.44	-0.130	0.056
Time Bin 3		-0.058	0.050	2311	-0.024	-1.17	0.24	-0.156	0.039
Dissimilar Accurate*Time Bin 2		0.200	0.057	2311	0.072	3.51	<.001	0.088	0.311
Dissimilar Accurate*Time Bin 3		0.218	0.059	2311	0.076	3.68	<.001	0.102	0.334
Similar Accurate*Time Bin 2		0.146	0.057	2311	0.053	2.54	0.01	0.033	0.259
Similar Accurate*Time Bin 3		0.203	0.060	2311	0.069	3.38	<.001	0.086	0.321

Extended Data Fig. 4 | Looking Time Multilevel Model. Multilevel model results for Experiment 1 (**a**) and Experiment 2 (**b**). Results displayed here are for models dummy coded relative to inaccurate trial type and time bin 1. Models were evaluated for significance with a chi-squared difference test. Both models were statistically different from an intercept only model (Experiment 1: X^2 (8) = 64.45, $p < .001$; Experiment 2: X^2 (8) = 17.87, $p = .02$).

nature research

Corresponding author(s): Sarah Leckey, Dr. Simona Ghetti

Last updated by author(s): Apr 29, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The eye tracker task was administered on a Tobii T-120 17-in eye-tracker monitor. Experiments were coded using Tobii Studio, version 3.0.9. For the touchscreen task experiments were coded using Direct RT version 2006.2.0.28.
Data analysis	All statistical analyses were done using R version 3.3.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data is available on Open Science Framework at https://osf.io/t8p4g/.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The studies used a quantitative cross-sectional design.
Research sample	Experiment 1: Eighty toddlers ages 25-34 months (M = 28.85, 39 females) participated in this study. Families' household incomes were less than $15,000 (n = 1), $15,000-$25,000 (n = 4), $25,000-$40,000 (n = 7), $40,000-$60,000 (n = 16), $60,000-$90,000 (n = 16), more than $90,000 (n = 33) and unreported (n = 3). Four toddlers were African American, 10 were Asian, 4 were American Indian or Alaskan, 56 were Caucasian, 3 were Native Hawaiian, and 3 did not report a race. Experiment 2: Eighty toddlers ages 25-33 months (M = 29.09, 43 Females) participated in this study. Families' household incomes were less than $15,000 (n = 3), $25,000-$40,000 (n = 9), $40,000-$60,000 (n = 8), $60,000-$90,000 (n = 13), more than $90,000 (n = 46) and unreported (n = 1). Eight toddlers were African American, 14 were Asian, 1 were American Indian or Alaskan, 51 were Caucasian, and 6 did not report a race. This sample was chosen because it was the first wave of a longitudinal study examining the development of uncertainty monitoring from 2-3 years of age.
Sampling strategy	The sample was a convenience sample. Based on pilot data on an independent sample and previous research investigating similarity in toddlers (Arias-Trejo & Plunkett, 2010), we anticipated a medium effect size of similarity in the current study. Our sample is sufficient to detect a small to medium main effect (f = .15, np2 =.02) with 80% power and an alpha of .05 between our three conditions (similar-accurate, dissimilar-accurate, and inaccurate). Since the studies are part of a larger longitudinal study, data collection ended when we had reached the predetermined number of participants.
Data collection	An eye-tracking computer and touch screen computer were utilized for data collection of the variables. In addition, parents filled out language and demographic surveys with pen and paper. Children sat on the parents lap during the eye-tracking task and for the touchscreen task, only the experimenter and the child were present. The researcher was aware that there was an experimental condition, but did not know which trials were under which condition.
Timing	Data collection occurred between October 2014 and December 2017. Some pilot testing occurred before these dates.
Data exclusions	Trials were excluded in which no answer was provided (Experiment 1: 13 (.88%) total trials across 8 participants for the touchscreen task and 33 (2.23%) trials total across 16 participants in the eye tracker task. Experiment 2: Thirteen (.88%) total trials across 8 participants for the touchscreen task and 14 (1.06%) trials total across 5 participants in the eye tracker task) and in which parent indicated that the child was not familiar with the target word (Experiment 1: 51 (3.45%) trials total across 26 participants in the touchscreen task and 56 (3.78%) trials total from 26 participants in the eye tracker task. Experiment 2: Sixty-one (4.12%) trials across 28 participants in the touchscreen task and 52 (3.95%) trials from 26 participants in the eye tracker task). Once these eliminations were made, children with < 50% accuracy scores were removed from analyses (Experiment 1: One participant performed below chance (<.50) on both tasks, and two performed below chance on the touchscreen task. Experiment 2: Seven participants were removed due to performing below chance on the touchscreen task, and 6 were removed for performing below chance on the eye tracker task). For response time analyses, trials with short response times (less than 700 ms) and trials with z-scored response latencies across individual participants were +/- 3 standard deviations were removed (Experiment 1: This resulted in 35 (2.36%) trials across 35 participants. Experiment 2: 38 (2.57%) trials across 37 participants). For drift diffusion analyses, trials with response latencies > 15 seconds were removed due to the limitation of these models only being reliable for faster trials based on previous research (Experiment 1: 18 (1.23%) trials across 9 participants. Experiment 2: 28 (1.89%) trials across 13 participants). Individual drift parameters were removed for being outliers of +/- 3 standard deviations (Experiment 1: 8 parameter values across 4 participants. Experiment 2: 9 parameter values across 6 participants). Finally, for looking time analyses, trials were removed if the eye-tracker reported no fixations on either of the stimuli (Experiment 1: 158 trials (10.68%) total from 36 participants. Experiment 2: 178 (13.52%) trials total from 37 participants).
Non-participation	No participants dropped out/declined participation.
Randomization	Studies were within subject, no randomization was required.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology
☐	☐ Animals and other organisms
☐	☒ Human research participants
☐	☐ Clinical data

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	Participants were recruited from a database of names of families who had previously expressed interest in participating in child development studies. These families were originally contacted about interest when the child was an infant by state birth records.
Ethics oversight	This study was approved by the Institutional Review Board of the University of California, Davis.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*

Cell population abundance | *Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*

Gating strategy | *Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type | *Indicate task or resting state; event-related or block design.*

Design specifications | *Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*

Behavioral performance measures | *State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s) | *Specify: functional, structural, diffusion, perfusion.*

Field strength | *Specify in Tesla*

Sequence & imaging parameters | *Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*

Area of acquisition | *State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software | *Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*

Normalization | *If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*

Normalization template | *Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*

Noise and artifact removal | *Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring | *Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings | *Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*

Effect(s) tested | *Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016) | *Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*

Correction | *Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.