Foundations of intuitive power analyses in children and adults

nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-022-01427-2

Foundations of intuitive power analyses in children and adults

Madeline C. Pelz, Kelsey R. Allen, Joshua B. Tenenbaum and Laura E. Schulz ✉

Decades of research indicate that some of the epistemic practices that support scientific enquiry emerge as part of intuitive reasoning in early childhood. Here, we ask whether adults and young children can use intuitive statistical reasoning and metacognitive strategies to estimate how much information they might need to solve different discrimination problems, suggesting that they have some of the foundations for 'intuitive power analyses'. Across five experiments, both adults ($N = 290$) and children ($N = 48$, 6–8 years) were able to precisely represent the relative difficulty of discriminating populations and recognized that larger samples were required for populations with greater overlap. Participants were sensitive to the cost of sampling, as well as the perceptual nature of the stimuli. These findings indicate that both young children and adults metacognitively represent their own ability to make discriminations even in the absence of data, and can use this to guide efficient and effective exploration.

In science, we use power analyses to estimate how much evidence is needed to test the hypotheses we are considering. For instance, if we suspect that a drug will cure 90% of patients while a placebo will work 30% of the time, we will conclude that a relatively small sample of patients will enable us to test the drug's efficacy; by contrast, if we believe the drug will cure only 50% of patients, we know we will need a larger sample to show an effect. Exact power analyses, of course, require formal training in statistics and an estimate of the size of the effect the researcher aims to demonstrate. But the intuitive foundation of these analyses—that it takes more evidence to distinguish some hypotheses than others—may be part of common sense reasoning more broadly. Decades of research suggest that some of the epistemic practices that support scientific enquiry emerge as part of intuitive reasoning in early childhood. Here we look at whether lay adults and young children can use these intuitive metacognitive abilities to represent how much information they need to distinguish between populations depending on the degree to which the populations overlap.

This question is grounded in a long tradition of work suggesting that children are sensitive to statistics in the environment and engage in selective exploration to maximize expected information gain. Infants attend to the transitional probability of events in both auditory[1] and visual[2] stimuli and can infer abstract rules from patterns of data (for example, the difference between ABA and ABB patterns[3]). Infants can also infer probable outcomes simply from the number of modal possibilities; thus, for instance, if three blue objects and one red one are spinning in an open container, infants look longer if the red object falls out than if a blue one does[4]. Additionally, infants are sensitive to the relationship between samples and populations: If most of the objects in a box are red, babies expect most of the objects pulled from the box to be red but suspend this inference if the objects are drawn from somewhere other than the box (that is, the experimenter's lap) or are drawn selectively by an experimenter searching through the box rather than randomly[5]. Infants also actively explore to reduce uncertainty. Seven- and 8-month-olds are more likely to look away from events that are very predictable or unpredictable relative to moderately surprising events[6] and 12-month-olds not only look longer at events that violate their intuitive theories, but intervene to explore the violations (for example, dropping objects that appeared to violate gravity; banging objects that appeared to violate solidity[7]).

Children's sensitivity to the relationship between hypotheses and evidence becomes increasingly sophisticated from toddlerhood to middle childhood. Toddlers are able to integrate observed data with their previous beliefs, and their inferences are not only sensitive to the content of a sample, but also whether that sample was drawn randomly or selectively[8]. When preschoolers observe events that cannot be explained by a known cause, they posit unobserved variables to explain the event[9,10], and selectively explain and explore events that violate their causal theories[11,12]. Preschoolers can also use the base rate of events to distinguish more and less probable hypotheses[13,14], isolate variables in a causal system to distinguish candidate causes[11], and integrate the evidence they observe with the testimony they hear from knowledgeable sources[15].

However, it is less clear to what extent young children have an explicit understanding of the relationship between the evidence they observe and the knowledge they will gain. Some studies suggest that at least some precursors to metacognition emerge early. Preschoolers correctly distinguish objects they can and cannot name[16], spend more time considering response options given uninformative versus informative task instructions[17], selectively withhold answers[18–22], ask for help[23] on items they struggle to remember and show more pupillary dilation and give higher confidence ratings to remembered items[24,25], see also refs. [18,22,26–28]. However, young children may know when they are more or less certain about information without using this knowledge to increase opportunities for learning. In self-paced learning tasks, 5-, 6- and 7-year-olds are more confident about correctly than incorrectly remembered items but only 6- and 7-year-olds accurately anticipate which items they will have trouble learning and dedicate more study time to these items[29]. Consistent with this, school-aged children often struggle with metacognitive tasks in the context of test-taking: children are overconfident in their memory[30,31], and choose randomly when given the chance to restudy test items rather than choosing on the basis of their previous performance[32]. Introspective self-reports of knowledge also improve from early school age to adolescence[33–37]. Thus,

Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. ✉e-mail: lschulz@mit.edu

children's ability to monitor their uncertainty and use it to regulate their behaviour appears to develop through middle childhood.

Despite these findings, it is hard to assess the extent to which children's information seeking is precisely and quantitatively calibrated to their uncertainty: almost all of the studies reviewed above have relied on children's qualitative judgements about task difficulty. The few studies that have used graded measures have looked at how long children attend to a task, or their confidence ratings, as they are actually working on the task and experiencing uncertainty online. Children might be able to make well-calibrated responses to their current state of knowledge without having fine-grained representations of task difficulty a priori. For instance, one recent paper found that 4- to 8-year-olds' exploration quantitatively tracked the difficulty of discrimination problems; when asked to identify the number of marbles in a box by shaking the box and listening, children's exploration time was independent of the number of marbles in the box but systematically tracked the difficulty of the simulated contrast (for example, the time children spent shaking a box containing nine marbles increased linearly when trying to discriminate those nine marbles from three, six or eight marbles in another box)[38]. This study suggests an impressively precise correlation between children's uncertainty and their exploration, but in this context it is unclear whether children decided how long to explore a priori on the basis of their representation of the difficulty of the task, or whether they used online monitoring to continuously gather more information until they were confident enough to guess.

In the current study, we are interested in cases in which learners cannot use uncertainty monitoring as a way to continuously adjust their exploration. Instead, we ask whether adults and children can represent the difficulty of discrimination problems and estimate, in advance, how much information they will need to make a good guess about the answer. To test this, we borrowed from a classic model in the infancy literature[39]. We introduced participants to two populations of coloured balls: in this case, boxes with inverse proportions of coloured and white balls (for example, one filled with 90% red balls and 10% white balls, and one with 90% white and 10% red, referred to hereafter as a 90/10 set). As noted, even infants can use an observed sample of balls to guess the population from which it was drawn. Here however, we never draw a sample of balls; Learners are simply asked how many balls they would need in the sample to tell the two populations apart. The difficulty of this discrimination problem depends, of course, on the overlap between the populations: distinguishing 90/10 from 10/90 is relatively easy and should require only a small sample of balls, while distinguishing 60/40 from 40/60 is much harder and requires a larger sample.

In a formal power analysis, a scientist might use an effect size estimated from previous findings, a significance level of $P=0.05$ and a power level of at least 0.8 to calculate the sample size needed for her study. In contrast, participants in the current task are given full knowledge of the populations that they are comparing so effect size is no longer an unknown variable. Measures of confidence also replace significance level and power in this context, as it is difficult to assign numerical values to each of these factors as they combine to structure the simulated sampling choices of each participant.

The participant never sees any balls drawn and must select the size of the sample a priori, so the decision about how many balls to draw depends on participants' assessments of the difficulty of the discrimination problem rather than their ability to solve it (that is, none of the problems can be solved given only the information provided). By varying the discriminability of the populations and comparing the results to a model of the difficulty of the discriminations, we can investigate the extent to which both adults and children modulate their information seeking in quantitatively precise ways that track the difficulty of the task.

Because we are interested in whether either adults or children have the foundations of intuitive power analyses, we start by testing adults (experiments 1–4b). We then focus on 6- to 8-year-olds (experiment 5 and replication) because the previous literature suggests that these are among the youngest ages at which explicit uncertainty monitoring and control emerge, and because pilot work suggested that 4- and 5-year-olds struggled with the task demands of the full quantitative version.

Results

Computational framework. To succeed in this task, participants need to decide how informative a sample of a certain size would be without being able to see the contents of the sample. Vul et al.[40] offers one way to formalize this sampling behaviour, suggesting that when ideal Bayesian inference is intractable and there is a cost associated with each sample the globally optimal solution might be to instead rely on a very small number of samples from the posterior distribution[40]. We proposed that this strategy might successfully capture behaviour in our sampling task, so building on this previous work we adapted this framework to model an agent who represents all of the possible contents of a sample of a particular size N (for example, in a sample of $N=3$ balls, it could contain three white balls, three coloured balls, one white and two coloured balls, or two white and one coloured ball), then weights them on the basis of the likelihood of drawing that particular sample from the population. We then use a cumulative distribution function to model the sampling process,

$$\sum_{i=0}^{x} \binom{N}{i} p^i (1-p)^{(N-i)}$$

with N being the total number of samples drawn (restricted to even-numbered samples using a ceiling operator), x representing the number of samples that are coloured and p representing the probability that a given sample is a particular colour (that is, the probability associated with a given box). When most of the possible samples of a given size point to one of the two options, then participants should decide the sample is big enough; if not, then participants should request a larger sample. We therefore represent the utility of a particular sample size N as

$$\sum_{i=\mathrm{ceil}(N/2)}^{N} \binom{N}{i} p^i (1-p)^{(N-i)}$$

When this utility model is mapped to the number of samples an agent should request, it suggests that participants should sample exponentially, resulting in the ideal learner requesting almost the entirety of the potential samples for the hardest discriminations. This exponential strategy is ideal in a situation in which there is no cost for additional samples, but ignores the important role that cost might play when approximating utility. Drawing additional samples would in theory take time and energy; to address this, we modify the utility function to include a cost of sampling proportional to the number of samples taken, formally represented as: $N \times c$. The full utility model, for a given sample size N is then

$$\sum_{i=\mathrm{ceil}(N/2)}^{N} \binom{N}{i} p^i (1-p)^{(N-i)} - N \times c$$

When even a moderate cost of sampling is added to the model, it drastically affects the shape of the ideal learner's sampling curve, changing it from an exponential to an inverted U-shape (Fig. 1). This makes the prediction that for both the easiest and most difficult discriminations, a small number of samples should be drawn, with the most samples being drawn for the moderately difficult discriminations in between. This framework echoes the work of Vul

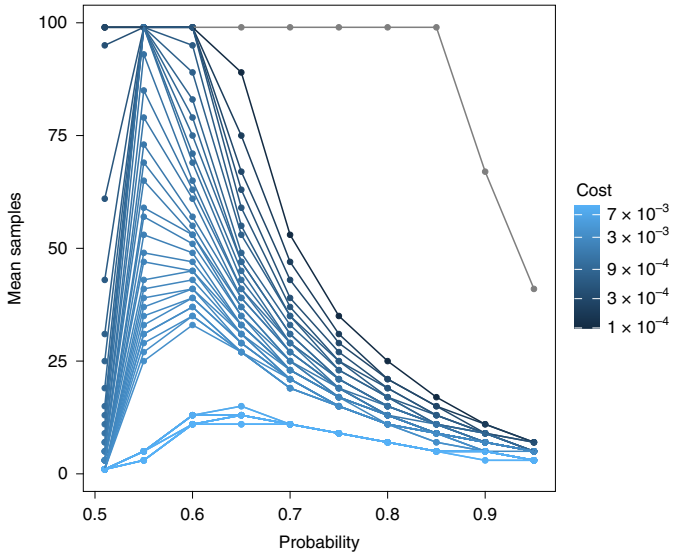

Fig. 1 | Ideal sampling across varied costs. The grey line at the top indicates the predicted samples for each proportion if there were no cost to sampling. The remaining lines indicate the ideal number of samples for each set of proportions across different costs of sampling according to the globally optimal model with added cost (1,000 model runs). When each sample incurs even a small cost, the sampling curve changes from an exponential to an inverted U-shape.

et al.[40] in which they demonstrate that choosing a very small number of samples (and in some cases, even a single sample), is a more optimal strategy when taking cost into consideration[40]. To compare the results from this globally optimal model to adults' sampling behaviour, we used the behavioural model described above.

Experiment 1. Adults were shown a short animation that walked them through the setup of the task, in which images of two boxes of balls with inverse proportions were shown being shuffled behind a barrier so the location of each was unknown. Then a hand reached into one of the boxes and picked balls one at a time and placed them into an opaque bowl without revealing the colour of each ball. After the container was filled, the contents were revealed and participants were asked to judge which of the two boxes the sample had been drawn from. The training trial was done with a 72/28 ratio set, and was designed to be an easy discrimination. Failure to choose the correct box was used as an exclusion criterion.

Participants were then shown four characters, each with a set of two boxes, to give them a sense of the range of contrasts in the task and were told 'This time, you will be deciding how many balls I should put in the bowl. For each friend, think about how tricky it will be to figure out which box I am picking from. For some friends you might need more balls to decide which box they are picked from, and for some friends you might need fewer. Try not to ask for more balls than you need.'

Ten sets of boxes each with a new character were presented one at a time, along with a question asking 'How many balls do you think I need to put in the bowl for you to know whether the balls came from my box or (the current character)'s box?' Participants simply had to type in the number of samples that they thought they would need to discriminate each pair (Fig. 2a). The proportions in the test stimuli varied from very easy discriminations (95/5) to very difficult ones (51/49) (Fig. 2b). Adults were not told the specific ratios; they had to estimate the difficulty of the discrimination problem from the visual display. Nor were adults told the exact number of balls in the box. One hundred balls were visible on the face of the box but the boxes were presented as three-dimensional sketches so, assuming participants thought the boxes were cubes, they might have inferred that there were roughly 1,000 balls in each box. The instructions indicated sampling without replacement and in principle this could be modelled by a hypergeometric distribution (that is, the probability of a coloured ball would change after each draw). However, given the inferred size of the population, sampling with or without replacement would make little difference thus for simplicity, we modelled the probability of drawing a coloured ball as a binomial distribution. Participants were not given any feedback on their responses and the order in which the ten boxes were presented was randomized for each participant.

As each participant made a judgement for each of the ten pairs of boxes, we used a linear mixed effects model to analyse the relationship between the difficulty of the discrimination problem and the number of samples requested by participants. Proportion of coloured balls was a fixed effect, with participant as a random intercept. *P* values were obtained by likelihood ratio tests comparing the full model with a null model that left out the effect of proportion. Adults were sensitive to the difficulty of the discrimination problem and asked for more balls as the discrimination problems got more difficult ($\chi^2(1) = 66.19$, $P < 0.001$), requesting 0.37 ± 0.04 (standard error (s.e.), 95% CI (0.28, 0.45)) more balls for each decreasing proportion (for example, from 60/40 to 59/41), Fig. 3a). Quantile–quantile plots of model residuals were used to verify the normality of the data. Although there were no explicit costs associated with sampling more balls (that is, it was as easy to request 80 samples as it was ten), adults did not sample exponentially more balls in the hardest discriminations. Adults' sampling behaviour appeared linear rather than exponential or an inverted U-shape, suggesting that participants were unlikely to be relying fully on the strategy described above, despite evidence that they were in fact incorporating an implicit cost of sampling.

Experiment 2. To verify that costs indeed affect participants' responses, we replicated experiment 1 with 30 additional adults on MTurk, but made the cost of sampling explicit by requiring participants to push a button to request each additional sample (for example, rather than typing the numeral 10 they had to click the button ten times), while still being blind to the content of each sample. As expected, participants were in fact sensitive to the cost of each additional sample, and the addition of explicit cost led participants to sample more conservatively across the board. We again used a linear mixed effects model with proportion of coloured balls as a fixed effect and participant as a random intercept. The results replicated those in experiment 1: participants' responses were graded with respect to the difficulty of the discrimination, asking for more samples for the more difficult discriminations and fewer for the easier discriminations ($\chi^2(1) = 57.16$, $P < 0.001$), requesting 0.15 ± 0.02 (s.e.), 95% CI (0.11, 0.19) more balls for each decreasing proportion (Fig. 3b). Quantile–quantile plots of model residuals were used to verify the normality of the data.

Despite the fact that adults' sampling behaviour indicates that they again considered the cost of sampling, participants' estimates of how many samples to draw for each discrimination do not reflect the U-shaped curve predicted by the globally optimal model, and continue to appear linear. What other factors might be influencing participants to sample in this way? One possibility for this discrepancy is that while the model predicts that people should take almost zero samples for the most difficult case, in experiments 1 and 2 we instructed adults to request a sample for every discrimination problem that was presented. This might have led them to make a pragmatic assumption that they should sample in every round even if in some cases they knew that it would be nearly impossible to be certain about the correct answer.

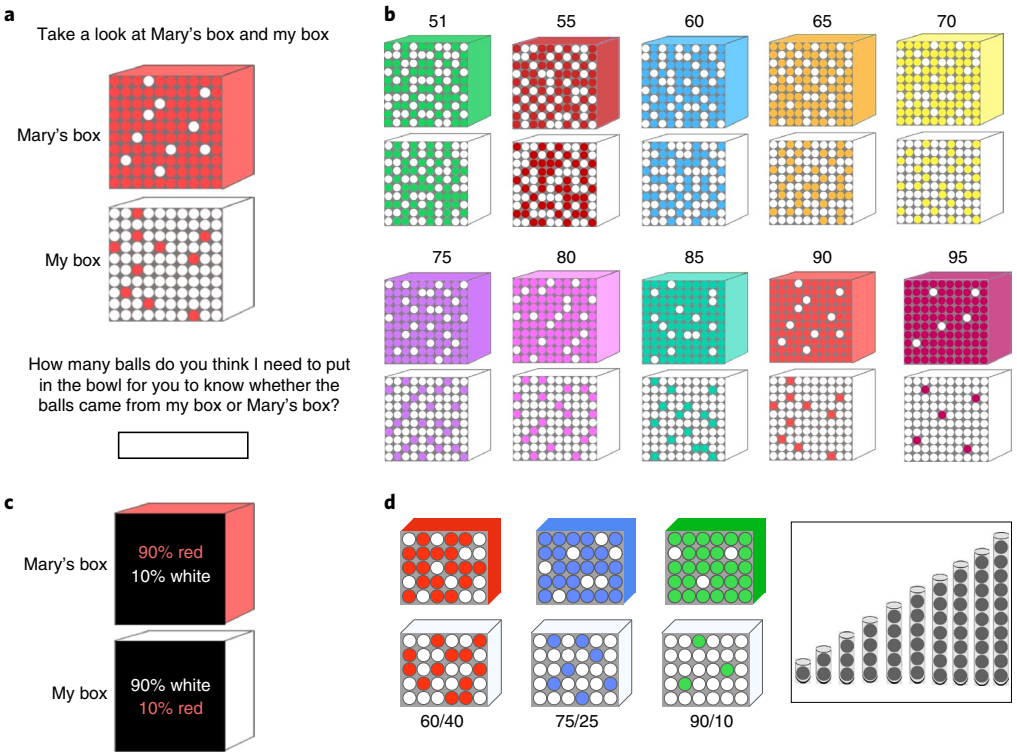

Fig. 2 | Stimuli used in behavioural tasks. a, Screenshot of how test questions were presented to online adult participants in experiments 1 and 2. **b**, The proportions in the test stimuli varied from very easy discriminations (95/5) to very difficult ones (51/49), and were randomly shuffled for each participant. **c**, Screenshot of how test questions were presented to adult participants online in experiment 4b to avoid perceptual noise. **d**, Schematic of the simplified three-box task in experiment 5. Participants were presented with puppets and physical boxes filled with small plastic balls in ratios of 60/40, 75/25 and 90/10, and images of tubes ranging from the height of one ball to the height of ten balls.

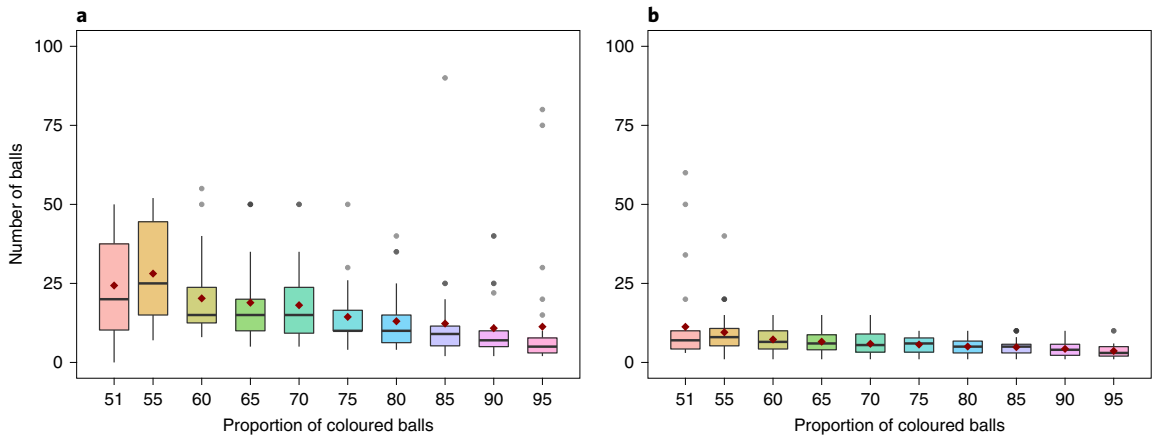

Fig. 3 | Adult sampling behavior for each of the ten discriminations (ranging from 51/49 to 95/5) in experiments 1 and 2. a, Participants' sampling behaviour in experiment 1 tracked the difficulty of discriminations; adults ($N = 30$) requested 0.37 ± 0.04 (standard error (s.e.), 95% CI (0.28, 0.45) more samples for more difficult discriminations ($\chi^2(1) = 66.19$, $P < 0.001$). Red diamonds indicate the mean number of samples requested. Box plots are centred on the median number of samples, with the bounds of each box marking the first and third quartiles, and the whiskers extending to 1.5× the interquartile range. **b**, The number of balls adults requested for each of the ten discriminations in experiment 2, when the cost of sampling was explicit. Participants' sampling behaviour tracked the difficulty of discriminations; adults ($N = 30$) requested 0.15 ± 0.02 (s.e.), 95% CI (0.11, 0.19) more samples for each decreasing proportion ($\chi^2(1) = 57.16$, $P < 0.001$). Red diamonds indicate the mean number of samples requested. Box plots are centred on the median number of samples, with the bounds of each box marking the first and third quartiles, and the whiskers extending to 1.5× the interquartile range.

Experiment 3a. To investigate whether adults have the metacognitive awareness to rationally opt out of the most difficult discriminations if given the opportunity, experiment 3a closely replicated experiment 1 with the addition of one sentence to the end of the task instructions: 'To opt out of this trial, please enter '0'.' Fifty additional adults were recruited on Amazon MTurk and tested online.

When explicitly given permission to opt out of each trial, some participants took the option to quit while others still requested

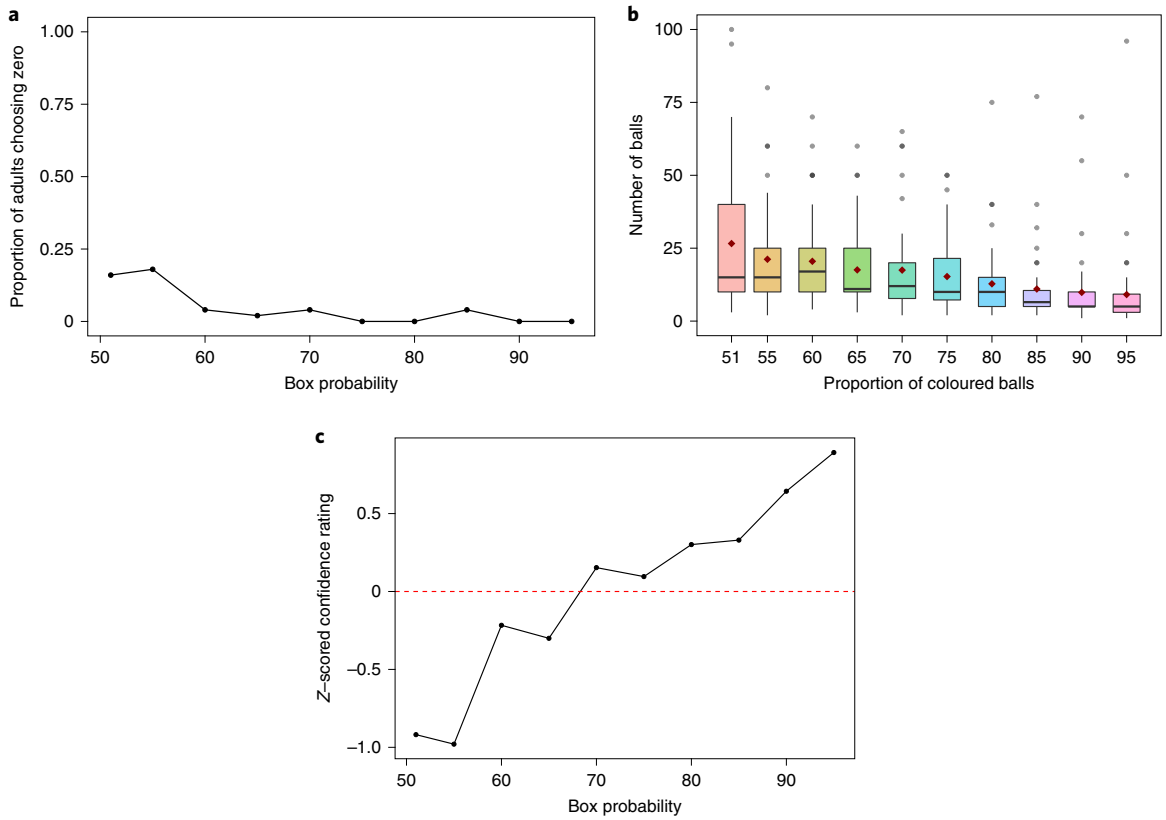

Fig. 4 | 'Measures of adults' metacognitive awareness before and after sampling in experiments 3a and 3b. a, Adult participants ($N=50$) made rational decisions about when to quit, choosing to quit more often for the more difficult discriminations. **b**, Number of balls adults requested for each of the ten discriminations in experiment 3a with all instances where adults sampled zero balls (that is, chose to quit) removed. Participants' sampling behaviour once again tracked the difficulty of discriminations; adults ($N=50$) requested 0.28 ± 0.04 (s.e.), 95% CI (0.21, 0.36) more samples for each decreasing proportion ($\chi^2(1)=56.94$, $P<0.001$). Red diamonds indicate the mean number of samples requested. Box plots are centred on the median number of samples, with the bounds of each box marking the first and third quartiles, and the whiskers extending to 1.5× the interquartile range. **c**, *Z*-scored confidence ratings across discrimination difficulty. Participants' ($N=50$) ratings were above zero only for the easier discriminations (ratios of 70/30 to 95/5).

samples in a similar pattern as in experiments 1 and 2. Forty percent of participants (20/50) chose to use the quitting option; those who quit did so rationally, choosing to quit more often for the most difficult discriminations (Fig. 4a). Even when all the '0' (quitting) responses were removed, adults were sensitive to the difficulty of the discrimination problem, requesting more samples for more difficult discriminations. According to a linear mixed effects model with proportion of coloured balls as a fixed effect, and participant as a random intercept ($\chi^2(1)=56.94$, $P<0.001$), adults requested 0.28 ± 0.04 (s.e.), 95% CI (0.21, 0.36) more balls for each decreasing proportion, Fig. 4b). Quantile–quantile plots of model residuals were used to verify the normality of the data. But how did the 60% of participants who chose not to quit determine how many samples to request in the more difficult discriminations?

One possibility is that participants chose a threshold of certainty that they hoped to attain for each sampling problem, and then asked for the number of balls that would be required in each of the scenarios to reach that threshold. Although this threshold-based model seems to be an intuitive strategy, it predicts drastic exponential sampling behaviour that does not align with our behavioural results (Supplementary Fig. 1). Despite this, the fact that a subset of participants knew to give up for the more difficult discriminations suggests that they had metacognitive awareness about the cases in which they had a lower likelihood of success. Even for other participants who did not choose to quit, this ability might be reflected in a measure of their reported confidence in simulating the sampling outcomes in these more difficult cases.

Experiment 3b. To measure participants' metacognitive awareness of their ability to simulate the outcome of each discrimination task, experiment 3b again replicated experiment 1, but instead of offering participants the option to give up, we added a question after each sampling judgement to ask participants how confident they would be in picking the correct box if they saw the number of samples they had just requested. Participants used a slider to indicate their confidence between 0 (not confident at all) and 100 (extremely confident), with additional labels at 25 (slightly confident), 50 (moderately confident) and 75 (very confident). After *z*-score normalization, participants' confidence ratings were only positive for the relatively easier discriminations (70/30 to 95/5), suggesting that although participants' sampling tracked the difficulty of each discrimination, they were also aware that their ability to succeed in the task would not be constant across the different discriminations (Fig. 4c).

Experiment 4a. In addition to having lower confidence in the more difficult discrimination problems, the perceptual nature of our stimuli may have also affected participants' sampling behaviour. Previous research has suggested that perceptual estimations of proportion are often distorted, and can affect decision-making reliant on these stimuli (see ref. [41] for review). To investigate this in the context of our specific stimuli, we ran an experiment in which we asked adult participants ($N=100$) to estimate the proportion of coloured balls in each of the ten different distributions used in this study. Participants were shown the same images of pairs of boxes of varying proportions as in experiments 1–3, but each image was

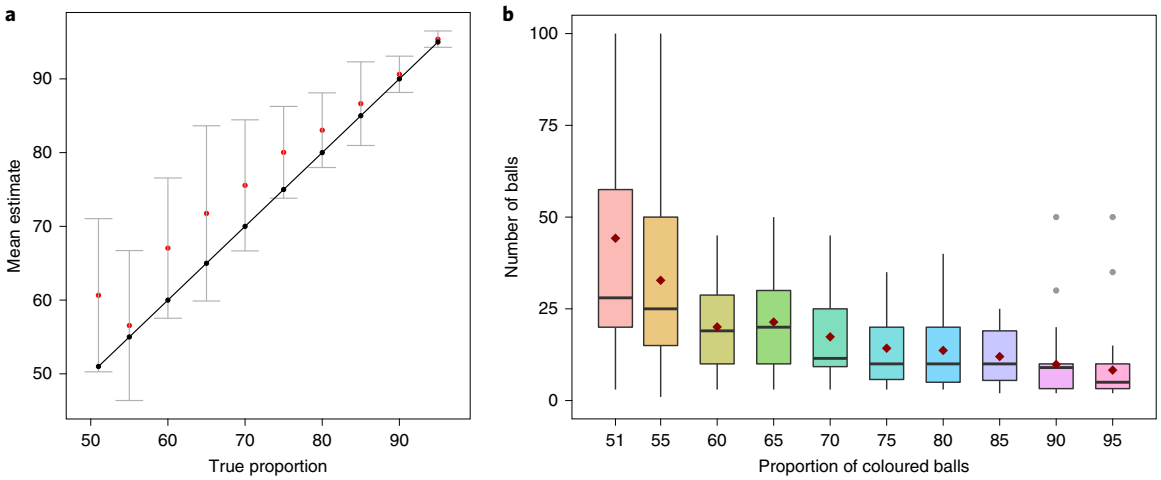

Fig. 5 | Adult estimates and sampling behavior using numerical rather than visual proportions. a, Adults' ($N=100$) estimates of the proportion of coloured balls in each set of boxes (experiment 4a). Participants' estimates were both closer to the true proportions and less variable as the proportion of coloured balls increased and the boxes became easier to distinguish from each other. The mean and standard deviations of these estimates were used to inform the perceptual noise model. The black line indicates the true proportion, while the red points are centred on the mean estimate for each proportion ± s.e.m. **b**, Number of balls adults requested for each of the ten discriminations (ranging from 51/49 to 95/5) when shown numerical rather than visual proportions in experiment 4b. Participants ($N=30$) again asked for more samples for more difficult discriminations, but their sampling behaviour was exponential ($\beta=-0.03$, $P<0.001$, adjusted $R^2=0.28$, 95% CI = (−0.04, −0.02), AIC = 702.33, as fit by a linear model) rather than linear ($\beta=-0.66$, $P<0.001$, adjusted $R^2=0.25$, 95% CI = (−0.80, −0.54), AIC = 2,541.29) when perceptual noise was not a factor. Red diamonds indicate the mean number of samples requested. Box plots are centred on the median number of samples, with the bounds of each box marking the first and third quartiles, and the whiskers extending to 1.5× the interquartile range.

only shown for 2 seconds before it disappeared so participants could not simply count the number of coloured balls. After each image was shown, participants were asked to estimate the proportion of coloured balls in the top box of the set. Participants' estimates were both closer to the true proportions and less variable as the proportion of coloured balls increased and the boxes became easier to distinguish from each other (Fig. 5a).

Perceptual noise model. In response to these findings, we modified the globally optimal model to use the means and standard deviations from the participants' estimates rather than the true proportions. The model structure remained the same, but for each discrimination problem, 100 samples of p were sampled from a distribution with the mean and standard deviation of the participants' estimates of the proportions on that trial, truncated so that $P>0.5$. Each of these 100 P values was then used to generate predictions for the number of samples to draw for the corresponding discrimination problem, which were then averaged to obtain a single model prediction. The addition of this 'perceptual noise' flattened the curve of the model prediction and successfully captured the linear shape of the human judgements in experiment 1 (with a fit cost of 0.0024 and a root-mean-square error (r.m.s.e.) of 3.93), experiment 2 with an additional cost of sampling (fit cost of 0.0091, r.m.s.e. = 1.99), as well as the sampling behaviour of participants in experiment 3 after removing the trials in which participants quit (fit cost = 0.0027, r.m.s.e. = 2.97) (Fig. 6a–c). This adjusted model offers a better fit overall than the model that does not include perceptual noise (experiment 1 r.m.s.e. = 3.88, experiment 2 r.m.s.e. = 4.10, experiment 3 r.m.s.e. = 3.75).

Experiment 4b. To further investigate the influence of perceptual noise on participants' sampling decisions, we ran a study with 30 additional adults modelled after experiment 1, but instead of representing the proportion of coloured and white balls inside the box by showing participants an image of the front of the box with the coloured balls visible, we presented them as percentages in written text (Fig. 2c). Simply by removing the visual nature of the stimuli, participants' sampling behaviour aligned with the exponential curve predicted by the no-cost model, rather than linearly as they had been before. We tested this by fitting a linear model to participants log-transformed ratings across the ten different proportions ($\beta=-0.03$, $P<0.001$, adjusted $R^2=0.28$, 95% CI = (−0.04, −0.02), Akaike information criterion (AIC) = 702.33) as well as to the raw data ($\beta=-0.66$, $P<0.001$, adjusted $R^2=0.25$, 95% CI = (−0.80, −0.54), AIC = 2,541.29). These results lend further support to the hypothesis that the linearity of adults' sampling in the previous experiments is influenced by the perceptual nature of our stimuli (Fig. 5b).

Experiment 5 and replication. The results of experiments 1–4 indicate that adults can successfully adjust their sampling behaviour on the basis of the difficulty of the discrimination problem, that they were sensitive to small changes in the cost of sampling and that they had the metacognitive awareness to know when the problem was too difficult to solve and it was a better choice to quit. In experiment 5, we ask whether 6- to 8-year-old children are also able to quantitatively track the discrimination difficulty of a problem and adjust their information gathering in response. We focus on this age range because (as reviewed above), the literature suggests that these are among the youngest ages at which children demonstrate competency in tasks requiring metacognitive monitoring of the gaps in their knowledge as well as control and planning of their future actions to minimize those gaps. In experiment 5 and its preregistered replication (osf.io/uafcq) we tested children's abilities in both a simplified version of the adult task and in a full quantitative task that very closely followed the structure of experiment 1.

Three inclusion criteria were used in both experiment 5 and the replication. First, children had to determine which of two populations (80/20 or 20/80 pink to white balls) a sample of ten pink and three white balls was drawn from. Next, children were shown three pairs of boxes (pairs presented in random order) with coloured and white balls in ratios of 60/40, 75/25 and 90/10 and asked which of the three pairs of boxes would be the easiest to tell apart from each

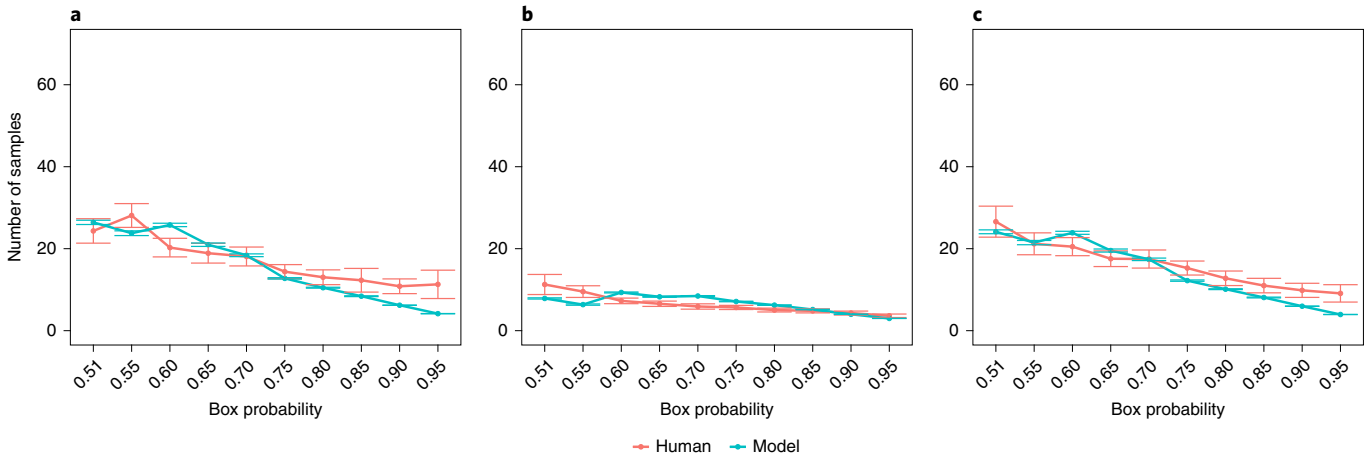

Fig. 6 | Perceptual noise model fits. a–c, Perceptual noise model fits for experiment 1 ($N=30$) (**a**), experiment 2 ($N=30$) (**b**) and experiment 3 ($N=50$) (**c**). This model had the same structure as the globally optimal model, but instead of using the true proportions of the box, used the means and standard deviations of estimates made by adult participants in experiment 4a. Data are presented as mean values ± s.e.m.

other and which would be the hardest, without any discussion of how sample size might relate to the difficulty of the discrimination problem (Fig. 2d). Finally, before children moved on to the full quantitative task they were asked to name three different numbers between one and 100 and then asked which of those numbers were the highest. Children who failed any of these three inclusion criteria were excluded from analysis.

Children were first introduced to two plastic boxes, one filled with an 80/20 ratio of pink to white balls and one with a 20/80 ratio of pink to white balls. The experimenter placed the boxes behind a barrier and shuffled them from side to side. The barrier was then separated into two pieces with one box hidden behind each so that the child could tell that the experimenter was only pulling samples from one of the boxes, but could not tell which box was behind which barrier. The experimenter then took balls one at a time from the box and placed them into a tube without revealing the colour of the balls. When the tube was filled, the experimenter revealed that the tube contained ten pink balls and three white balls and the child was asked from which population the sample originated. Three puppets were then introduced, each with pairs of boxes in 60/40, 75/25 and 90/10 proportions presented in a random order. To make sure children understood the structure of the game, they were then asked which of the three pairs of boxes would make the guessing game the hardest and which would make the game the easiest, without any discussion about how sample size might relate to difficulty.

After passing these inclusion criteria, children began the test trial. Children were shown a set of printed images representing ten different sized tubes, ranging from one that could hold only a single ball to one that could hold ten balls (Fig. 2d). The experimenter told the children that they were going to play the same game with each of the three puppets, but that they could choose which size tube to use for each game so that once the tube was filled, they would be able to figure out which box the balls had been sampled from. She then asked the children to assign one of the ten tubes to each game. To encourage children to think about the contrast between the different games, each tube could only be chosen once.

In both experiment 5 and its replication, most children correctly ranked the three discriminations, asking for more samples from the most difficult discrimination and the fewest from the easiest (experiment 5, Friedman chi-squared 15.25; $P<0.001$, replication, Friedman chi-squared 17.333, $P<0.001$, Fig. 7a,b, respectively). There was no effect of age on choosing the correct rank order experiment 5 estimate 0.75, z value = 1.187, $P=0.235$, 95% CI = (−0.44, 2.12), effect size = 0.32, replication: estimate 1.39, z value = 1.871, $P=0.06$, 95% CI = (0.13, 3.14), effect size 0.36, by binomial regression). Fifteen of 24 children (62.5%) chose the correct rank order in experiment 5; 17 of 24 children (70.8%) chose the correct rank order in the preregistered replication.

After children completed the task with the three sets of boxes, the same group of children were then given the full quantitative task similar to the task completed by adults. To continue to the next task, children were asked to list three numbers that fall between 1 and 100 and identify which of the numbers they had listed was the largest to ensure that they understood the size of the samples that they requested.

Children were then told that there were even more different sets of boxes that they could make guesses for. The tasks transitioned from real physical boxes and balls to digital images presented on a laptop. Rather than assigning a tube of a certain size, children were asked to say exactly how many balls each character would need on each trial to figure out which box the sample had been drawn from, between 1 and 100 (children's responses were anchored to avoid responses such as 'a truck full' or 'a gajillion'). The question was repeated for each of the ten different discrimination tasks, ranging from the most difficult (51/49) to the easiest (95/5), in a pseudo-random order. Otherwise, the task was identical to the adult task except that it was administered by the experimenter rather than online and instead of typing the number or tapping a key to request a sample, children were asked to say the number out loud.

Despite much larger variance in children's responses compared to adults (and the anchoring on 1 to 100 samples) 6–8-year-olds were sensitive to the difficulty of discriminations across the ten different contrasts in both experiment 5 and its replication, ($\chi^2(1)=39.79$, $P<0.001$) and $\chi^2(1)=52.09$, $P<0.001$), respectively, choosing to sample 0.72 ± 0.11 (s.e.), 95% CI (0.5, 0.93) more balls for each decreasing proportion in experiment 5 and 0.78 ± 0.10 (s.e.), 95% CI (0.58, 0.98) more in the replication, according to a linear mixed effects model with proportion as a fixed effect and participant as a random intercept, Fig. 7c,d, respectively). Quantile–quantile plots of model residuals were used to verify the normality of the data. Children's sampling behaviour was also successfully captured by the perceptual noise model in both experiment 5 and its replication (fit cost 0.0009, r.m.s.e. 10.39 and fit cost 0.0004, r.m.s.e. = 10.97, respectively, Fig. 8a,b, respectively).

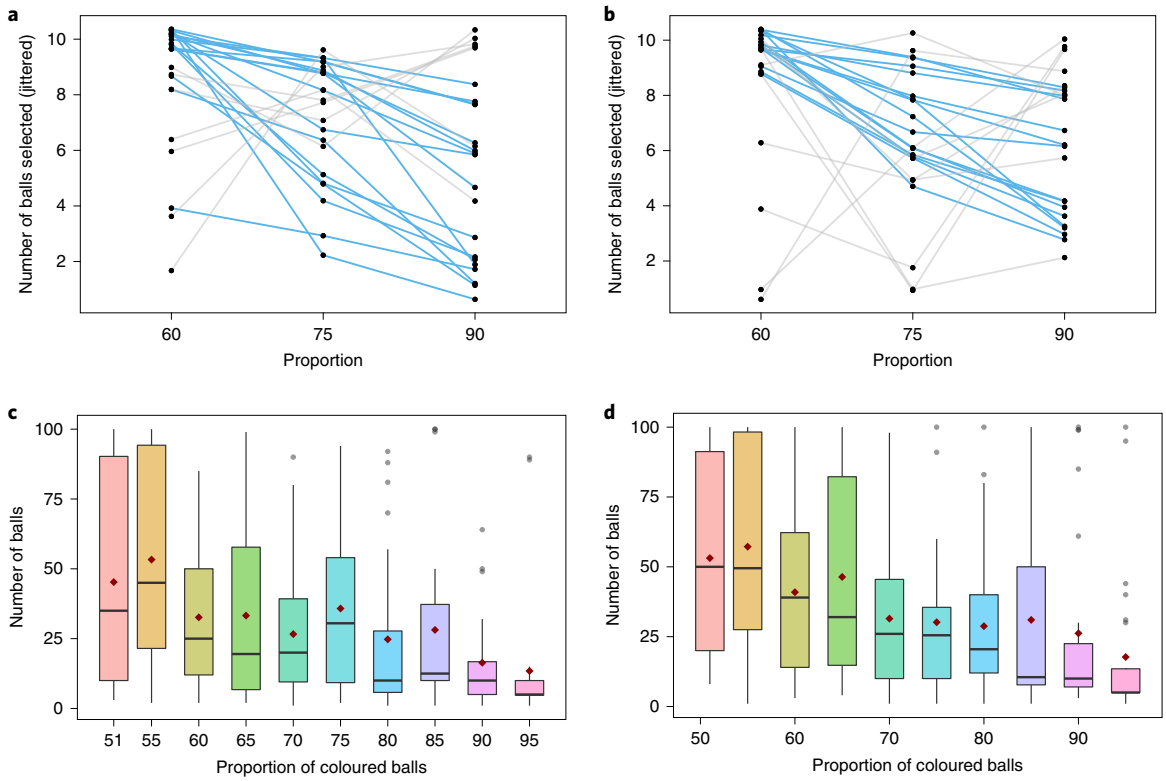

Fig. 7 | Children's sampling behavior in experiment 5 and its replication. a,**b**, In experiment 5 (**a**) and its replication (**b**), children requested the largest sample for the 60/40 discrimination, the smallest sample for the 90/10 discrimination and a sample in between those values for the 75/25 discrimination, suggesting that they modulated their information seeking on the basis of the difficulty of the task at hand ($N = 25$ in each experiment 5 and its replication, experiment 5 Friedman chi-squared 15.25, $P < 0.001$, replication: Friedman chi-squared 17.333, $P < 0.001$). Children who correctly rank-ordered all three tasks (15 of 24 in experiment 5 and 17 of 24 in the replication) are highlighted in blue, while other participants are shown in grey. **c**,**d**, In both experiment 5 (**c**) and its replication (**d**), 6- to 8-year-olds ($N = 24$ were sensitive to the difficulty of discriminations across the ten different contrasts, ($\chi^2(1) = 39.79$, $P < 0.001$) and $\chi^2(1) = 52.09$, $P < 0.001$), respectively, choosing to sample 0.72 ± 0.11 (s.e.), 95% CI (0.5, 0.93) more balls for each decreasing proportion in experiment 5 and 0.78 ± 0.10 (s.e.), 95% CI (0.58, 0.98) more in the replication. Red diamonds indicate the mean number of samples requested. Box plots are centred on the median number of samples, with the bounds of each box marking the first and third quartiles and the whiskers extending to 1.5× the interquartile range.

Discussion

Considerable previous work suggests that even infants represent the relationship between samples and populations[5,8,39], and in this sense are 'intuitive statisticians'. Building on this, the current work suggests that children's intuitive statistical reasoning can extend far beyond this ability. Here we demonstrate that both adults and children can represent the probabilistic relationships between samples and populations metacognitively, judging the relative amount of evidence required to distinguish easier and harder discrimination problems in the absence of any specific information about the sample being drawn. Children and lay adults intuitively recognize something comparable to the kind of inference we make in science: that the more overlap there is between populations, the more statistical power it takes to distinguish them. While previous work has shown that even young children engage in online monitoring and control[22,42], the current study demonstrates that children can represent their uncertainty a priori and adjust their behaviour quantitatively in response to gradations in problem difficulty.

Across each of our experiments, both lay adults and children represented the relative difficulty of discriminating populations and recognized that larger samples were required to discriminate populations with greater overlap. In experiment 1, adults were able to track the difficulty of statistical discriminations and adapt their sampling behaviour in response. Nonetheless, their behaviour did not align with the U-shaped globally optimal model of sampling described in previous work[40]. In experiment 2, we confirmed that this was not because adults were insensitive to the cost of sampling; adults requested fewer samples across the board when costs were explicit. However, they still asked for more samples than predicted by the model for more difficult discriminations. Arguably, this might have been because adults did not realize they could ask for zero balls (that is, 'quit') on the most difficult problems. When given explicit permission to quit (experiment 3) participants that decided to quit did so for the most difficult discriminations, which aligned with their self-reported confidence in making the discriminations. Participants felt confident in their ability to succeed in the easier discriminations (between 70/30 and 94/5), but did not feel confident for the more difficult discriminations (51/49 to 65/35).

One puzzling aspect of participants' behaviour in experiment 3 is why, when given the explicit option of quitting on the hardest discrimination problems, 40% of participants elected to do so and 60% did not. Here, we asked separate groups of participants either to sample or to quit, or to estimate their confidence when sampling. Thus, one possibility is that the group of participants who chose to quit were (appropriately) less confident in their ability to make estimates from the sample than those who did not. Alternatively, those who quit might have placed a higher estimate on the cost of sampling, or at least the combined cost a sample large enough to be informative. It is also possible that other temperamental factors (a dislike of quitting in general on characterological grounds, or

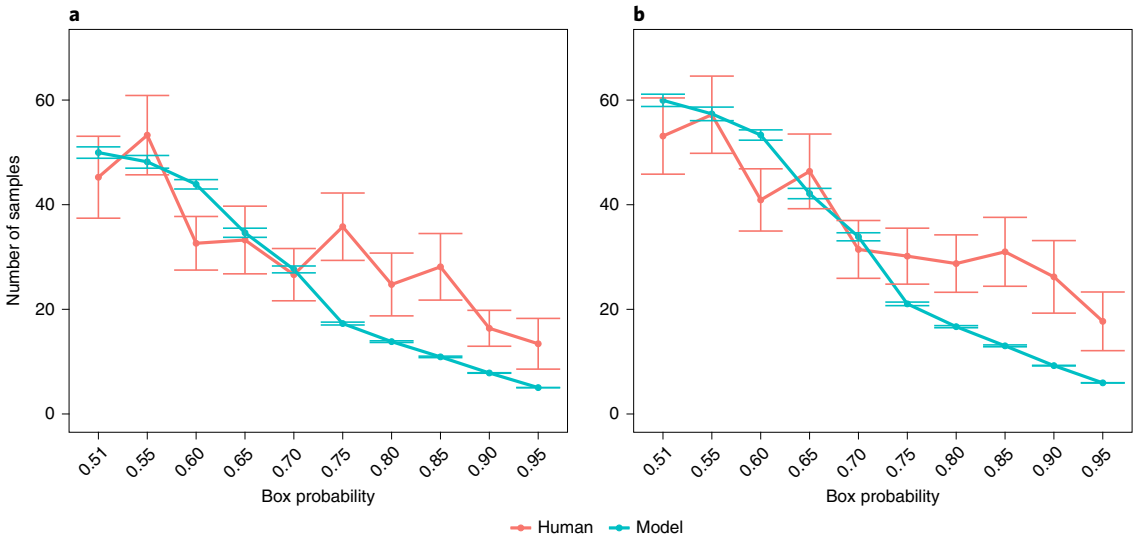

Fig. 8 | Perceptual noise model fits for Experiment 5 and its replication. a,**b**, Experiment 5 (**a**) and its replication (**b**). This model had the same structure as the globally optimal model, but instead of using the true proportions of the box, used the means and standard deviations of estimates made by adult participants in experiment 4a. Data are presented as mean values ± s.e.m.

a greater preference for making risky guesses under uncertainty) might have motivated some of the participants' behaviour. Finally, in this design participants never actually had to perform the task at all; participants were asked to choose a sample but they did not need to guess the population. In contexts where there is no penalty for making a wrong guess, and especially in this context where they did not have to guess at all, participants may reasonably feel that they might as well try something rather than nothing. Future research might distinguish the factors that affect quitting behaviour, and extend this work to children as well.

In each of experiments 1–3, adults' sampling behaviour appeared to be increasing linearly as the discriminations became more difficult, while the globally optimal model suggests exponentially larger samples would be necessary for the most difficult cases. In experiment 4a we asked adults to estimate the numerical proportion of each set of boxes from the visual depictions used in earlier experiments, and then ran the model on those estimated proportions rather than on the true proportions. This simple addition changed the sampling estimates from exponential to linear. To verify this experimentally, in experiment 4b we removed the perceptual factor by showing participants the proportion of balls contained in each box as text (for example, '70% red balls and 30% white balls') instead of as a visual representation of the balls themselves. In this experiment, adults' sampling reflected an exponential trend, further supporting the impact that perceptual noise had on sampling behaviour in the previous experiments. Despite the exponential nature of participants' sampling behaviour in experiment 4b, the number of samples they requested still does not align with what would be expected if they had maintained a fixed confidence threshold across each discrimination (Supplementary Fig. 2).

Another question that stems from our findings is whether there are other factors beyond the cost of each sample and perceptual noise that might have influenced adults to request a relatively small number of samples for even the most difficult discrimination problems. Previous work[42,43] has suggested that people are relatively insensitive to sample size, and in particular, trust that small samples will reflect the population they are drawn from more than is warranted. While participants in our task were sensitive to relatively small variations in the ratio of objects when contrasted across populations and knew to ask for more information for harder discriminations, they drew much smaller samples (consistent with a belief in the law of small numbers) than was predicted by the globally optimal model. However, adults' confidence ratings suggested that while they might have been biased towards selecting small samples, they were not fully misled about their chances of success: they (accurately) expressed very low confidence that they would succeed on the most difficult discrimination problems. The interplay between the assumption that small samples will be representative of their population and rational changes to confidence in the probability of success from these small samples is another interesting topic for future study.

The current results also extend our understanding of the metacognitive abilities of children in middle childhood. In experiment 5, children's sampling behaviour also tracked the difficulty of each discrimination problem, with children requesting more samples for the populations with the most overlap. To our knowledge, this study is one of very few to quantitatively map the link between children's ability to represent the difficulty of a problem and the amount of information they need to solve it[38]. In contrast to earlier work suggesting that school-age children struggle with metacognitive tasks such as allocating study time on the basis of their past performance[32], but consistent with more recent work suggesting relatively sophisticated metacognition in young children (for example, ref. [22]) our findings suggest that young school-age children can represent both the difficulty of problems and the amount of information they will need to solve them. In the current study, 6 to 8-year-olds were able to make proactive, graded judgements about how much information they would need to solve a problem, and engaged in both metacognitive monitoring and control across fine-grained discriminations. While this demonstrates impressive metacognitive abilities, children were more likely to oversample in the easier discrimination cases, leading to more divergence from model predictions than adult sampling.

In contrast to these findings, one recent study that assessed how children and adults chose to sample information to determine the location of a target object found that children undersampled with respect to their ideal learner model[44]. A few important differences may have led to these disparate results. In the Jones et al. task, children sampled possible locations one at a time, and could see the explicit cost of each additional sample as well as the explicit reward they would get if they could identify the object's correct position. In the current study, the cost of sampling was implicit, and rather

than being able to monitor cost and confidence online, children were asked to make a single, proactive judgement about how many samples they would need for each discrimination. Relative to adults, children may weigh costs more heavily when they have to collect samples one by one but underestimate costs when estimating the total amount of information needed all at once. Recall also that children in our task were anchored on a scale from 1–100 samples. Again, relative to adults, sampling one at a time may have led children to stop early but anchoring children on a range extending all the way to 100 may have induced them to consider large samples. Future work might explore how children's estimation of cost might differ in proactive versus online sampling within the same task as well as how anchoring affects children's sampling behaviours.

While the current task provided a context in which we could precisely quantify how sampling behaviours changed with task difficulty, everyday decision-making often lacks such clear and concrete displays of information. In less controlled environments, we are forced to more roughly assess and estimate likelihoods when choosing to act. In the current study, target populations and their inverse pairs were also presented in succession, giving participants the additional possible benefit of seeing related problems in the context of those that differed only in their relative difficulty. This relative ease may limit the applicability of these findings to real-world estimation problems. Future work could ask how participants might respond to discriminations such as those used in the current study when presented in isolation, as well as whether people's ability to track how much information they need extends to more diverse domains.

Collectively, the current results indicate that adults and children as young as six distinguish easy and difficult discrimination problems, and can make sophisticated graded inferences about the number of samples they need as discrimination problems become more difficult. Beyond the impressive ability to simulate how much information we might need to solve a problem, the current findings suggest that this complex metacognitive ability insensitive to the cost of sampling and the modality in which we perceive the population. Our findings also suggest that we are able to judge which subset of problems we can complete with confidence, and have the capacity to use this information to decide how to act (or whether to quit) on tasks that extend past our current abilities. Future work might further explore the development of our metacognitive abilities across different tasks and domains and the ways that they inform and direct our information seeking behaviour.

Methods

Participants. This research complied with all ethical regulations of and was approved by the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology. Informed written consent (for adults and parents of children who participated) and verbal assent (for children) was obtained from all participants.

Computational modelling. Models were written and run using Python. Model predictions were fit to both adults' and children's behavioural data in R by finding the value of the cost of sampling c that minimized the sum of squared errors between the model's predicted number of samples and the number of samples requested by participants.

Experiment 1. G*Power was used to estimate that with an effect size of 0.4 (based on pilot testing) to reach a power level of 0.9 we would need a sample size of $N=30$. Thirty-three adults were recruited on Amazon Mechanical Turk and tested online. Participants were required to be above 18 years of age to participate, but data on exact age and sex were not collected for online participants. Participants were compensated according to minimum wage in Massachusetts for the average time spent on the task. Three participants were excluded from analysis for failing to correctly answer an attention check, for a final sample size of $N=30$.

Experiment 2. Thirty-two adults were recruited on Amazon Mechanical Turk and tested online. Participants were required to be above 18 years of age to participate, but data on exact age and sex were not collected for online participants. Participants were compensated according to minimum wage in Massachusetts for the average time spent on the task. Two participants were excluded from analysis for failing to correctly answer an attention check, for a final sample size of $N=30$ (consistent with experiment 1).

Experiment 3a. For experiment 3, we increased the sample size because allowing participants to enter '0' would mean we would have fewer data points to analyse and we wanted to be able to have similar power with potentially fewer data points per participant. Fifty-four adults were recruited on Amazon Mechanical Turk and tested online. Participants were required to be above 18 years of age to participate, but data on exact age and sex were not collected for online participants. Participants were compensated according to minimum wage in Massachusetts for the average time spent on the task. Three participants were excluded from analysis for failing to correctly answer an attention check, and one was excluded for failing to answer all test questions, for a final sample size of $N=50$.

Experiment 3b. Sixty-one adults were recruited on Amazon Mechanical Turk and tested online. Participants were required to be above 18 years of age to participate, but data on exact age and sex were not collected for online participants. Participants were compensated according to minimum wage in Massachusetts for the average time spent on the task. Eleven participants were excluded from analysis for failing to correctly answer an attention check, for a final sample size of $N=50$ (consistent with experiment 3a).

Experiment 4a. One hundred and forty-one adults were recruited on Amazon Mechanical Turk and tested online. Participants were required to be above 18 years of age to participate, but data on exact age and sex were not collected for online participants. Participants were compensated according to minimum wage in Massachusetts for the average time spent on the task. Forty-one participants were excluded from analysis for failing to correctly answer an inclusion question (successfully reporting that a box with 100% coloured balls contained a proportion of 100% and not any other proportion), for a final sample size of $N=100$.

Experiment 4b. Sixty-six adults were recruited on Amazon Mechanical Turk and tested online. Participants were required to be above 18 years of age to participate, but data on exact age and sex were not collected for online participants. Participants were compensated according to minimum wage in Massachusetts for the average time spent on the task. Twelve participants were excluded from analysis for failing to correctly answer an attention check. Due to the fact that this experiment was less visually interesting than the previous experiments, an additional check question was included in which participants saw a set of boxes with 100% of one colour of balls and 0% of the other, and were excluded if they did not request a single sample for that trial. Twenty-four participants were excluded on the basis of this check question for a final sample size of $N=30$ (consistent with experiments 1 and 2).

Experiment 5 and replication. With a goal of reaching a power level of 0.9 for the Friedman's test used for the ranking task, we performed simulations on bootstrapped samples drawn from pilot data. A sample of 24 participants was sufficient to reach this threshold, as recorded in a preregistration (osf.io/uafcq, registered July 2018). Children were recruited from an urban children's museum and tested in a private room off the museum floor; 27 children were recruited for experiment 5. Participants were given a selection of stickers to thank them for participating. Data collection was not performed blind to the order of the discrimination problems presented to the children. One was excluded from analysis due to a camera malfunction that prevented the data from being coded, one due to parent-reported verbal disability and one for failure to identify the easiest and most difficult discriminations in the simplified task for a final $N=24$ (15 female, mean age 7; 6; range 6; 0–8; 11). For the replication, an additional sample of 30 children were recruited, with four excluded on the basis of their failure to identify the easiest and most difficult discriminations in the simplified task, and two who failed to complete the task with a final sample of $N=24$ (15 female, mean age 7; 2; range 6; 1–8; 10).

Materials. For experiments 1–4, the testing materials were presented using the online survey platform Qualtrics and were run asynchronously online.

For experiment 5, four cloth hand puppets (pink, red, blue and green) were used. Four clear plastic boxes ($12\times12\times14\,cm^3$) were lined with paper that matched the colour of most of the balls inside (coloured paper for the characters' boxes, white for the experimenter's) on all sides except the front so that participants could see the ratio of coloured to white plastic balls (2 cm diameter) inside. A barrier was glued into place so that the first two layers of plastic balls were held in place in the same location for all participants and it looked as though the entire box was filled. For the set of boxes used during the training trial, two additional hidden compartments were built behind the barrier of each box to hold the right number and colour of balls needed to place into the tube during the warm-up game so that the experimenter could close her eyes but still sample a consistent sample across participants. The cardboard tube used during training (3 cm diameter, 26 cm in height) was cut vertically so that one side could be covered with clear

packing tape to create one opaque side and one clear side. During sampling, two white cardboard barriers (each 46×36 cm) were used to cover each of the boxes and block the experimenter's hands while sampling balls from the box into the tube. For the simplified task, laminated images of the different sizes of tubes that children could choose from (2 cm width, varying in height by how many balls it could contain) were presented on a plastic tray (25×38 cm), and during the full quantitative task, digital images of the ten sets of boxes were presented using Keynote on a 13 inch laptop screen.

All children were tested in a private testing room in a children's museum with both the experimenter and the parent or guardian present, with the child sitting across from the experimenter at a small child-sized table. All sessions were recorded and children's tube assignment and verbal responses were coded from video by an experimenter blind to trial order.

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

Data availability

The data that support the findings of this study are available on the Open Science Foundation project page found at https://osf.io/gdp68. Source data are provided with this paper.

Code availability

Code used for data analysis is available from the corresponding author upon request.

Received: 25 October 2020; Accepted: 8 July 2022;
Published online: 05 September 2022

References

1. Saffran, J. R., Aslin, R. N. & Newport, E. L. Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928 (1996).
2. Johnson, M. H., Posner, M. I. & Rothbart, M. K. Components of visual orienting in early infancy: contingency learning, anticipatory looking, and disengaging. *J. Cogn. Neurosci.* **3**, 335–344 (1991).
3. Marcus, G. F., Vijayan, S., Rao, S. B. & Vishton, P. M. Rule learning by seven-month- old infants. *Science* **283**, 77–80 (1999).
4. Téglás, E. et al. Pure reasoning in 12-month-old infants as probabilistic inference. *Science* **332**, 1054–1059 (2011).
5. Xu, F. & Denison, S. Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition* **112**, 97–104 (2009).
6. Kidd, C., Piantadosi, S. T. & Aslin, R. N. The Goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE* **7**, e36399 (2012).
7. Stahl, A. E. & Feigenson, L. Observing the unexpected enhances infants' learning and exploration. *Science* **348**, 91–94 (2015).
8. Gweon, H., Tenenbaum, J. B. & Schulz, L. E. Infants consider both the sample and the sampling process in inductive generalization. *Proc. Natl Acad. Sci. USA* **107**, 9066–9071 (2010).
9. Bullock, M., Gelman, R. & Baillargeon, R. in *The Developmental Psychology of Time* (ed. Friedman, W. J.) 209–254 (Academic Press, 1982).
10. Saxe, R., Tenenbaum, J. & Carey, S. Secret agents: Inferences about hidden causes by 10-and 12-month-old infants. *Psychological Sci.* **16**, 995–1001 (2005).
11. Bonawitz, E. B., van Schijndel, T. J., Friel, D. & Schulz, L. Children balance theories and evidence in exploration, explanation, and learning. *Cogn. Psychol.* **64**, 215–234 (2012).
12. Legare, C. H. Exploring explanation: explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Dev.* **83**, 173–185 (2012).
13. Cook, C., Goodman, N. D. & Schulz, L. E. Where science starts: spontaneous experiments in preschoolers' exploratory play. *Cognition* **120**, 341–349 (2011).
14. Sobel, D. M., Tenenbaum, J. B. & Gopnik, A. Children's causal inferences from indirect evidence: backwards blocking and Bayesian reasoning in preschoolers. *Cogn. Sci.* **28**, 303–333 (2004).
15. Sobel, D. M. & Kushnir, T. Knowledge matters: how children evaluate the reliability of testimony as a process of rational inference. *Psychological Rev.* **120**, 779 (2013).
16. Marazita, J. M. & Merriman, W. E. Young children's judgment of whether they know names for objects: the metalinguistic ability it reflects and the processes it involves. *J. Mem. Lang.* **51**, 458–472 (2004).
17. Patterson, C. J., Cosgrove, J. M. & O'Brien, R. G. Nonverbal indicants of comprehension and noncomprehension in children. *Developmental Psychol.* **16**, 38 (1980).
18. Balcomb, F. K. & Gerken, L. Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental Sci.* **11**, 750–760 (2008).
19. Lyons, K. E. & Ghetti, S. I don't want to pick! Introspection on uncertainty supports early strategic behavior. *Child Dev.* **84**, 726–736 (2013).
20. Hembacher, E. & Ghetti, S. Don't look at my answer: subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological Sci.* **25**, 1768–1776 (2014).
21. Kim, S., Paulus, M., Sodian, B. & Proust, J. Young children's sensitivity to their own ignorance in informing others. *PLoS ONE* **11**, e0152595 (2016).
22. Ghetti, S., Hembacher, E. & Coughlin, C. A. Feeling uncertain and acting on it during the preschool years: a metacognitive approach. *Child Dev. Perspect.* **7**, 160–165 (2013).
23. Coughlin, C., Hembacher, E., Lyons, K. E. & Ghetti, S. Introspection on uncertainty and judicious help-seeking during the preschool years. *Developmental Sci.* **18**, 957–971 (2015).
24. Lyons, K. E. & Ghetti, S. The development of uncertainty monitoring in early childhood. *Child Dev.* **82**, 1778–1787 (2011).
25. Paulus, M., Proust, J. & Sodian, B. Examining implicit metacognition in 3.5-year-old children: an eye-tracking and pupillometric study. *Front. Psychol.* **4**, 145 (2013).
26. Call, J. & Carpenter, M. Do apes and children know what they have seen? *Anim. Cognition* **3**, 207–220 (2001).
27. Chouinard M. M. Children's questions: a mechanism for cognitive development. *Monogr. Soc. Res. Child Dev.* https://doi.org/10.1111/j.1540-5834.2007.00412.x (2007).
28. Whitebread, D. et al. The development of two observational tools for assessing meta-cognition and self-regulated learning in young children. *Metacognition Learn.* **4**, 63–85 (2009).
29. Destan, N., Hembacher, E., Ghetti, S. & Roebers, C. M. Early metacognitive abilities: the interplay of monitoring and control processes in 5-to 7-year-old children. *J. Exp. Child Psychol.* **126**, 213–228 (2014).
30. Flavell, J. H., Friedrichs, A. G. & Hoyt, J. D. Developmental changes in memorization processes. *Cogn. Psychol.* **1**, 324–340 (1970).
31. Koriat, A., Sheffer, L. & Ma'ayan, H. Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *J. Exp. Psychol.: Gen.* **131**, 147 (2002).
32. Metcalfe, J. & Finn, B. Metacognition and control of study choice in children. *Metacognition Learn.* **8**, 19–46 (2013).
33. Hofer, B. K. & Pintrich, P. R. The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning. *Rev. Educ. Res.* **67**, 88–140 (1997).
34. Lockl, K. & Schneider, W. The effects of incentives and instructions on children's allocation of study time. *Eur. J. Developmental Psychol.* **1**, 153–169 (2004).
35. Roebers, C. M. & Howie, P. Confidence judgments in event recall: developmental progression in the impact of question format. *J. Exp. Child Psychol.* **85**, 352–371 (2003).
36. Schneider, W. & Pressley, M. *Memory Development Between Two and Twenty* (Psychology Press, 2013).
37. Veenman, M. V., Van Hout-Wolters, B. H. & Afflerbach, P. Metacognition and learning: conceptual and methodological considerations. *Metacognition Learn.* **1**, 3–14 (2006).
38. Siegel, M. H., Magid, R. W., Pelz, M., Tenenbaum, J. B. & Schulz, L. E. Children's exploratory play tracks the discriminability of hypotheses. *Nat. Commun.* **12**, 3598 (2021).
39. Xu, F. & Garcia, V. Intuitive statistics by 8-month-old infants. *Proc. Natl Acad. Sci. USA* **105**, 5012–5015 (2008).
40. Vul, E., Goodman, N., Griffiths, T. L. & Tenenbaum, J. B. One and done? Optimal decisions from very few samples. *Cogn. Sci.* **38**, 599–637 (2014).
41. Zhang, H. & Maloney, L. T. Ubiquitous log odds: a common representation of probability and frequency distortion in perception, action, and cognition. *Front. Neuroscience* https://doi.org/10.3389/fnins.2012.00001 (2012).
42. Tversky, A. & Kahneman, D. Belief in the law of small numbers. *Psychological Bull.* **76**, 105–110 (1971).
43. Tversky, A. & Kahneman, D. Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974).
44. Jones, P. R. et al. Efficient visual information sampling develops late in childhood. *J. Exp. Psychol.: Gen.* **148**, 1138–1152 (2019).

Acknowledgements

Thank you to the Boston Children's Museum and the families who participated in this research. This material is based on work supported by the National Science Foundation Graduate Research Fellowship under grant no. 1745302 to M.C.P., and National Science Foundation grant no. 1231216 to L.E.S. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.C.P. and L.E.S. conceived and designed the behavioural studies. K.R.A. and J.B.T. contributed the computational modelling. The manuscript was written primarily by M.C.P. and L.E.S. with input and comments from K.R.A.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-022-01427-2.

Correspondence and requests for materials should be addressed to Laura E. Schulz.

Peer review information *Nature Human Behaviour* thanks Yingying Yang, Matteo Lisi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature research

Corresponding author(s): Laura E. Schulz

Last updated by author(s): May 5, 2022

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 our Editorial Policies 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	Data was collected from online survey platform Qualtrics for adults, and was coded from video for children. Children were shown both physical stimuli and digital stimuli using Keynote.
Data analysis	We analyzed data with custom R scripts and our computational models were implemented with custom Python code.

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 and 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

The data that support the findings of this study are available at osf.io/gdp68

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

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

Study description	Data are quantitative: adults and children were shown different pairs of boxes across different discrimination difficulties, and asked for a certain number of samples based on what they saw.
Research sample	Children ages 6-8 years were recruited from an urban children's museum. The sample was representative of the museum's patronage, but due to the cost of entry to the museum the sample is likely to have been less representative of the city's population overall.
Sampling strategy	Pilot studies were conducted to assess sample size. A larger sample of adults were included for Experiment 3 because we were aiming to assess behaviors (e.g. quitting) that were not relevant to every individual in the sample rather than comparison to the computational model.
Data collection	Adult participants were recruited from Amazon Mechanical Turk (MTurk) and tested using the online survey platform Qualtrics. Children were tested in person at an urban children's museum, and testing sessions were recorded with a video camera to be coded after each session. Because participants were young children, family members were present along with the experimenter during data collection. The experimenter was not blind to experimental design or hypothesis.
Timing	Experiment 1 and 2: 1/2018, Experiment 3a: 5/2019, Experiment 3b: 7/2020, Experiment 4a and 4b: 4/2018-6/2018, replication 6/2018-7/2018
Data exclusions	Across all 4 experiments, twenty-four participants were excluded from analysis due to failure to pass an attention check or inclusion question, three did not answer all of the questions, two for experimenter error/technological issues, and one for a parent-reported verbal disability that prevented the participant from answering the questions.
Non-participation	None.
Randomization	Presentation of test trials was randomized for all adult studies through Qualtrics, and children were presented with three different pseudo-random orders using Keynote.

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 and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above.
Recruitment	Parents/guardians visiting a local children's museum were approached and asked to participate. Adults were recruited on Amazon Mechanical Turk. All adults gave written consent and children assented to participate.

Ethics oversight	Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (COUHES) approved all experiments.

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