Intuitive Statistics & Metacognition in Children and Adults

Madeline Pelz (mpelz@mit.edu)
Department of Brain and Cognitive Sciences, MIT,
77 Massachusetts Avenue, Cambridge, MA 02139 USA

Laura Schulz (lschulz@mit.edu)
Department of Brain and Cognitive Sciences, MIT,
77 Massachusetts Avenue, Cambridge, MA 02139 USA

Abstract

Across four experiments, we look at whether adults and children can represent the amount of information needed to distinguish different populations in the context of an intuitive statistical reasoning task requiring metacognitive monitoring and control. Consistent with a ground truth model of information gain, adults (N=60) modulated their information gathering with respect to the difficulty of the discrimination problem. Adults also adjusted their confidence threshold depending on task difficulty, allowing for more uncertain judgments when the discrimination was more difficult or gathering data was more costly (Experiments 1 and 2). In a simplified version of the task, children (N = 42, *M* = 7.3 years, range: 5.0-9.0) were also able to distinguish easy and difficult discrimination problems and judge that they needed more information to solve harder problems (Experiments 3 and 4).

Keywords: metacognition; development; information gathering; statistical reasoning; children, cognitive development

Introduction

Much of the power of human learning stems from our metacognitive abilities: we recognize when problems are difficult and can identify the contexts in which we need more information to answer our questions. There are many lenses through which to view metacognitive tasks, but one classic model describes a cognitive structure consisting of two separate processes: monitoring and control (Nelson, 1990). Monitoring consists of the ability to judge our competence in a certain task, while control enables a person to modify their behavior in order to optimize their performance.

There is a large literature on the developmental and neural underpinnings of metacognition (e.g., Fernandez-Duque, Baird, & Posner, 2000), much of it looking at the alignment between people's assessment of their abilities and their performance on tasks involving recall memory and retrieval. However, because this work is largely qualitative, it is difficult to assess the extent to which people's information seeking is precisely calibrated to their uncertainty.

However, in recent years, many researchers interested in uncertainty and information gain have explored the degree to which both adults (e.g., Gureckis & Markant, 2012; Loewenstein, 1994) and children (e.g., Kidd, Piantadosi, & Aslin, 2012; Ruggieri & Lombrozo, 2015) engage in efficient information search. Such work suggests that even children search efficiently, maximizing opportunities for information gain (e.g., Kidd, et al., 2012). Critically however, learners might explore rationally in the face of uncertainty without any metacognitive representation of the relative difficulty of different tasks or metacognitive control over their information search.

In the current study we look at whether adults and children explicitly represent the relative difficulty of statistical reasoning tasks, and can use this judgment to modulate their information seeking. Statistical reasoning offers a useful domain in which to test monitoring and control of information search because we can precisely quantify the discriminability of different contrasts and ask how sensitive learners are to differences in populations, and how learners might modulate their information sampling based on the difficulty of the problem.

Assessing the confidence with which we can determine a population from a sample is commonplace in scientific hypothesis testing, however, in the current study we aim to investigate if non-expert adults and children intuitively use this type of reasoning to guide information seeking.

Some recent work has suggested that when children are asked to distinguish the number of marbles in a box by shaking and listening, children as young as four modify their exploration in a graded way, tracking the discriminability of the stimulus (e.g., shaking longer when trying to discriminate nine marbles from eight than nine from two; Siegel, Magid, Tenenbaum & Schulz, 2014). This suggests that children modify their information search in quantitatively precise ways with respect to psychophysical stimuli. However, this task leaves open the question of whether children have a metacognitive representation of the differences in task difficulty. Here we ask this question in a more abstract domain, but one that preserves our ability to model graded differences in task difficulty in a quantitatively precise way.

Logic of the Task

In order to look at if and how adults and children represent the difficulty of statistical discrimination problems and use this judgment to modulate their information seeking, we use a task in which participants observe two boxes of balls (e.g., one filled with 90% red balls and 10% white balls, and the other with 90% white and 10% red, labeled hereafter as 90/10). Participants are told they will get to see a sample of balls drawn from one of the two boxes and are asked to estimate how many balls they would need to see to know from which of the two populations the sample was drawn. The difficulty of the discrimination problem depends on the overlap between the populations. Distinguishing 90/10 from 10/90 is relatively easy and should require only a small sample of balls; distinguishing 60/40 from 40/60 would be much harder and require a larger sample. Importantly, the participant only gets to select the *size of the sample*; they never see the specific balls that

make up the sample. In this way, the decision about how many balls to sample can only be based on the difficulty of the discrimination problem rather than the informativeness of the sample, or their increased certainty about the correct answer. Thus the task requires metacognitive monitoring (to know whether the task will be relatively easy or difficult) and control (to determine the appropriate sample size).

Computational Model

Because the question we are asking is quantitative in nature, we can formalize the structure of this sampling task using a computational model, and consider human behavior with respect to a model of information gain based on sampling in this scenario. Although it is not a cognitive model, it allows us to characterize people's tolerance for uncertainty as a function of the difficulty of each discrimination problem.

The two boxes that make up each discrimination task are randomly shuffled out of sight of the participant, so the model assumes a uniform prior between them. The participant does not have access to the exact content of the sample because it is placed into an opaque container, so the model sums across all possible samples that could be drawn from each box (e.g., a sample of two balls could contain two red balls, two white balls, or one of each), weighted by the probability of those samples. After calculating the probability that a sample was drawn from each box, the larger of the two probabilities is selected because after a particular draw is revealed to a participant, we assume that they will guess the sample was drawn from the box that has the same majority color as the sample that they see. Using this strategy, their probability of being correct is equal to the probability that that specific sample was drawn from the chosen box.

These probabilities are then combined into a weighted sum across samples, formalized as

$$\sum_{s} max\left\{\frac{p(s|Box1)}{p(s|Box1)+p(s|Box2)}, \frac{p(s|Box2)}{p(s|Box1)+p(s|Box2)}\right\} * p(s) \quad (1),$$

which can be interpreted as the confidence with which one could answer what box the samples were being pulled from based only on simulating the data that has been drawn.

As the discrimination difficulty between boxes increases, the informativeness of each sample decreases, leading to different curves for each proportion, as seen in Figure 1. One key question of this study is how people might adjust their confidence thresholds as both the difficulty of the discrimination problem and the cost of obtaining new information changes. One hypothesis is that people may have a given threshold of certainty that they want to reach before they make a guess, and that that threshold stays constant. For instance, someone may want to have a 90% chance of being correct about whether they are picking from one box or the other before making a guess, regardless of whether the problem is easy or hard.

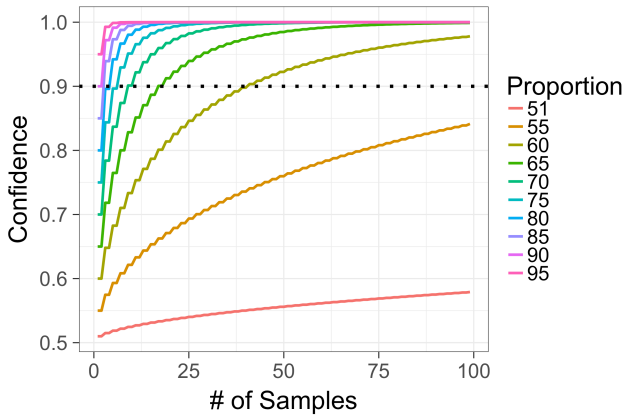

Figure 1. Formalization of the relationship between number of samples drawn and confidence in correctly guessing the box that it was drawn from. Dashed line represents a constant confidence threshold of 90%.

However, at high confidence thresholds, very large samples would be necessary to solve hard discrimination problems. Insofar as sampling evidence is costly, people might instead adjust their confidence threshold downward as the difficulty of the problems increases, becoming more willing to accept higher levels of uncertainty in more difficult situations. We tested participants in two conditions: one in which additional samples from the population could be taken at no cost, and one in which additional information came at a cost.

Experiment I

Participants & Method

Thirty adults were recruited and tested on Amazon Mechanical Turk. Four additional participants were excluded for failing to correctly answer check questions assessing attention and task understanding.

Participants were first shown a short video walking them through the setup of the task, in which two boxes with reversed proportions were shuffled behind a barrier so that participants did not know what box the samples would be drawn from. They then saw an animation of a hand picking out balls from behind one of the barriers and placing them into an opaque container. Following this demonstration, the contents of the container were revealed, and participants were asked to judge what box the sample had been drawn from. The training trial was done with a box with a ratio of 72/28 colored balls, and was designed to be an easy discrimination so that failure to make the discrimination could be used as an exclusion criterion.

Participants were then shown a sample of four characters and their boxes (Figure 2) to give them a sense of the space of possible contrasts in proportions. They were then told that for the rest of the games, it would be up to them to decide how many samples they wanted to draw in each set. Ten characters were then presented one at a time along with

their colored box and a white box with the reverse proportion, 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, so it was no more costly to sample 90 balls than 20 balls. Thus, we expected that participants' responses would largely reflect the difficulty of the discrimination problem, such that they would sample more balls as the problems became harder.

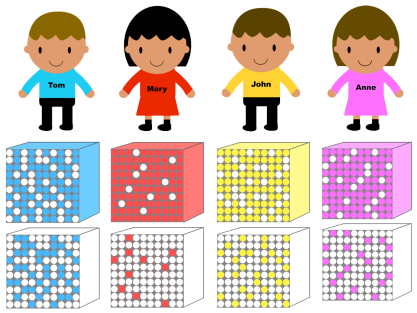

Figure 2. Characters and boxes of differing discrimination difficulty used in adult behavioral experiments 1 & 2.

Results

As clear in Figure 3, adults selected larger samples for more difficult discrimination problems (β= -0.36, *SE* = 0.06, *t* = -6.58, $p < 0.0001$, linear model). Although in principle, participants could have chosen a very large sample for all the discrimination problems, especially in this context where there was no explicit cost to sampling, the adults instead selected samples in a graded way, preferring more information for harder problems. Despite the fact that people chose larger samples for harder discrimination problems, they also accepted a lower certainty threshold for harder problems than easier problems rather than choosing the number of balls required to hold their threshold constant.

These results suggest that lay adults are "intuitive statisticians." They can use the difficulty of a discrimination problem to decide how much data they need to distinguish populations from samples, and they can do so without ever seeing the specific samples or gaining the specific information (the content of the sample) that would let them solve the discrimination problem. Intriguingly, although Experiment 1 imposed no costs to participants for sampling more data, people responded as if additional sampling were indeed costly, adjusting their confidence threshold downward. This is not unreasonable, given that sampling is typically costly in the real world. In Experiment 2, we explicitly add a cost to each additional sample to see if people continue to ask for more information for more difficult problems while also tolerating more uncertainty for more difficult discrimination problems.

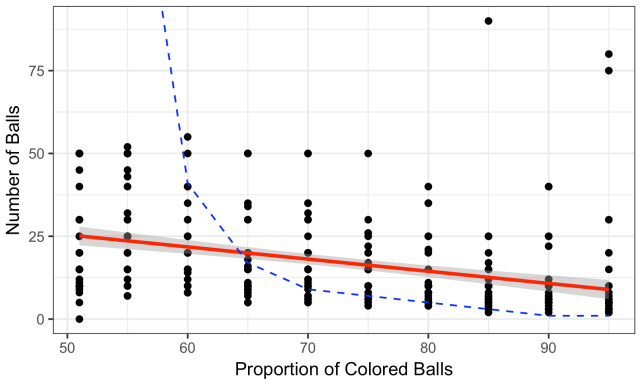

Figure 3. Adults request significantly fewer samples as discrimination difficulty decreases. Blue dashed line represents the number of samples needed to maintain a constant 90% confidence threshold.

Experiment 2

Participants & Method

A new sample of thirty adults were recruited and tested on Amazon Mechanical Turk. Two additional participants were excluded for failure to correctly answer check questions assessing attention and task understanding.

Experiment 2 was identical to Experiment 1 except that instead of being able to enter the total number of samples participants wanted for each proportion, they were asked after each individual sample if they would like to draw another ball or if they thought they had enough information to know which box was being sampled from.

Results

As in Experiment 1, adults selected larger samples for more difficult discrimination problems (β= -0.15, *SE* = 0.02, *t* = -6.6, $p < 0.0001$, linear model), Figure 4. In comparison to Experiment 1 however, the adults were even more conservative about their sampling: the average sample selected was smaller at every discrimination contrast

When a cost of sampling is added, this encourages participants to be more conservative in their sampling, and as it compresses the number of samples participants chose, it also compresses the range of certainty values.

When the model is used to transform the number of balls that the human participants chose to draw in each proportion to a confidence measure, it becomes apparent that participants are not relying on a single confidence threshold to make their judgments, but are instead modulating their confidence based on the difficulty of the task (Figure 5). Their tolerance for uncertainty increased with the difficulty of the discrimination problem regardless of the inclusion of an explicit cost of sampling.

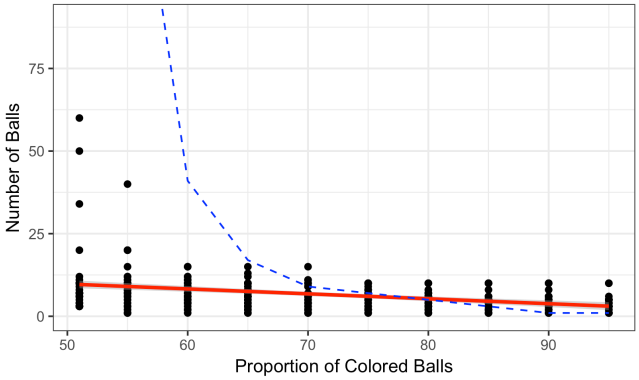

Figure 4. Adults select larger samples for more difficult discrimination problems even when required to pay a cost for each sample. Blue dashed line represents the number of samples needed to maintain a constant 90% confidence threshold.

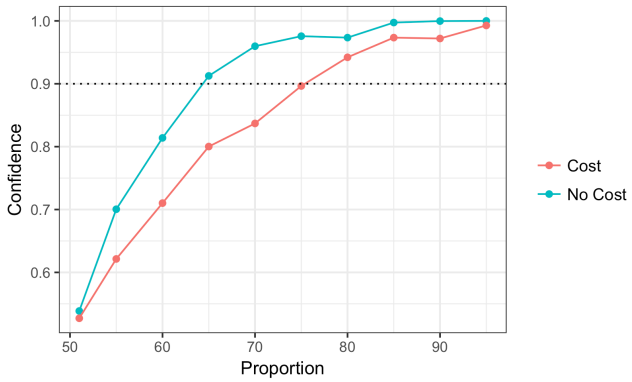

Figure 5. Adults in the cost (red) and no-cost (blue) conditions both show varying certainty thresholds across difficulty of discrimination. Dashed black line represents constant certainty threshold of 90%.

Experiment 3

In Experiment 3, we ask whether even young children monitor the relative difficulty of discrimination tasks and know to ask for more information for more difficult problems. The games that children played were similar to those used in the adult task, but much simpler. We asked children to make a qualitative distinction between easy tasks and hard tasks and asked whether children knew to ask for more information for the hard tasks. To engage the children, the tasks were embedded in a social context in which the children's job was to help four different puppets distinguish their boxes of toys from the experimenter's boxes so each box could be returned to its rightful owner. The boxes used for the children had 30 balls visible rather than the 100 used in the adult online game. Consistent with comparable work on children's understanding of uncertainty monitoring and information search (Nelson, et al., 2014; Ruggieri, et al., 2015), we tested a relatively wide age-range: five to nine-year-olds.

Participants

Children (N=25, M=6.9 years, range: 5.1-8.9, 48% girls) were recruited from an urban children's museum. For this and the following experiment, while most of the children were white and middle class, a range of ethnicities and socioeconomic backgrounds reflecting the diversity of the local population (47% European American, 24% African American, 9% Asian, 17% Latino, 4% two or more races) and the museum population (29% of museum attendees receive free or discounted admission) were represented throughout.

Training

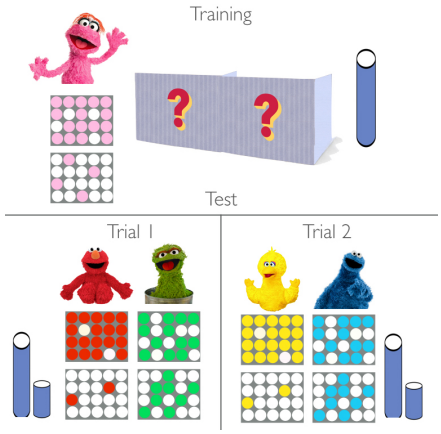

Figure 6. Task structure for behavioral experiments with children (Experiments 3 and 4).

Children were tested individually in a private room at a children's museum. During the training, the experimenter introduced the child to a puppet and two 12.5cm x 12.5cm boxes. Both had a transparent face, with the red box showing 20 red balls and 10 white balls and the white box showing 10 white balls and 20 red balls.

The experimenter said 'I'm going to play a trick on Sam and mix up these boxes behind this wall so he doesn't know which side his box is on." The experimenter placed the boxes behind a barrier made of two 35cm x 25cm cardboard screens and shuffled them from side to side. The two barriers were then separated with one box hidden behind each so that the child could tell that each box was behind a barrier but could not tell which box was where. Then the experimenter said "Here's how you can figure out which box is Sam's: I'm going to open up one of these boxes and take balls out one at a time and put them into this tube." She introduced a 3cm diameter, 30cm tall tube that was opaque on the side facing the child and clear on the side facing the experimenter, and began moving balls one at a time from the box to the tube without revealing the color of each ball to the child until the tube was filled to the top. Then the experimenter turned the tube of balls around so that the child could see the contents. She asked the child "Do you think I took these balls from Sam's box, or from

my box?" All children successfully identified the correct box. Then the experimenter brought out a second, smaller tube (3cm diameter x 10cm height) and asked "Suppose we had used this smaller tube for that same game. Would it have been easier or harder to guess which box the balls came from?" All children said that the smaller tube would have made the task more difficult.

Test

Following the training task, the puppet and the boxes from the training trial were moved out of sight, and children were shown two additional puppets, one with a 90/10:10/90 set of boxes, and one with a 60/40:40/60 set. The experimenter placed the puppets and their set of boxes one at a time onto the table as she said "This is [name]. He also brought a box of balls to the museum. Just like I had a white box with the same colors inside as Sam's box, I have a white box with the same colors as [name]'s box inside, but my box always has more white balls." Children were then told "Some of my friends' boxes are easier to tell apart from my box, and some are harder. Which of my friends' boxes do you think is easier to tell apart from my box?" to draw their focus to the contrasting proportions inside the character's boxes.

Children were then introduced to one large tube that could hold approximately 20 balls, and one small tube that could hold about 5 balls inside. The experimenter said "See this big tube? This tube can hold a lot of balls inside, so if it's hard to tell my friend's box apart from my box, it might be good to use the big tube. If it's easy to tell apart my box from my friend's box, you might only need to look at a couple of balls and you could use the little tube." Children were then asked, "Can you help me decide which tube to use for which friend's game?" and handed the tubes to place in front of the boxes. This test trial was then repeated with two different puppets who also had 90/10 and 60/40 box sets in different colors.

Results

As predicted, more children selected the large tube for the puppet with the difficult discrimination (60/40) and the small tube for the puppet with the easier discrimination (90/10) across both test trials (Wilcoxon signed-rank, $Z = 3.41$, $p < 0.001$). There was no effect of age on children's accuracy ($t = 1.03$, $p = 0.313$), although in this sample children did not begin to succeed until nearing age six.

These results suggest that children distinguish the relative difficulty of these statistical discrimination tasks and recognize that the more overlap there is between the populations, the larger the sample they will need to distinguish them. However, the task instructions in Experiment 3 leave open some doubt about whether children succeeded at both metacognitive monitoring and control or whether they succeeded only at the former. Children may have successfully identified which discrimination was more difficult but then rather than recognizing that they needed more information to make the harder discrimination, they may have simply followed the instructions to assign the larger tube to the more difficult problem. In order to look at whether children genuinely understand that the more difficult tasks requires more information, we ran an additional experiment in which we did not explicitly make the link between the difficulty of discrimination and the amount of information they needed to solve the problem. This allowed us to assess whether children could make this inference themselves.

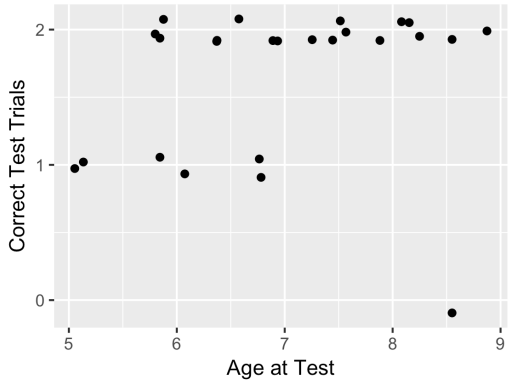

Figure 7. Children select the larger tube to provide more information for the more difficult comparison.

Experiment 4

Participants & Method

Children ($N = 18$, $M = 7.6$ years, range: 6.0 – 9.0, 50% girls) were recruited from an urban children's museum. Although there was no effect of age in Experiment 3, the few five-year-olds tested performed at chance, thus in Experiment 4, we restricted the sample to six to nine-year-olds.

The materials used in Experiment 4 were identical to Experiment 3, as were the explanation of the game, the training trial, and the introduction of the two 90/10 and 60/40 puppets for the test trial. When introducing the large and small tubes, the experimenter said "I have two tubes, one is big and can hold a lot of balls inside which would give us a lot of information about which box I picked the balls from, and one is small and can only hold a couple of balls inside, which would give us just a little bit of information." The connection between discrimination difficulty and information was not mentioned explicitly.

Results

As in Experiment 3, more children selected the large tube for the puppet with the more difficult discrimination and the smaller tube for the puppet with the easier discrimination across both test trials (Wilcoxon signed-rank, $Z = 2.33$, $p < 0.05$). Again, there was no effect of age on children's performance ($t = 1.117$, $p = 0.281$). In this study children could not succeed by simplifying identifying which task was harder and which was easier on each trial; children additionally had to recognize that they needed to collect more samples on the harder problem than the easy one. Children's success on this task suggests that they can both

monitor the difficulty of these discrimination problems and regulate their choices to maximize information gain.

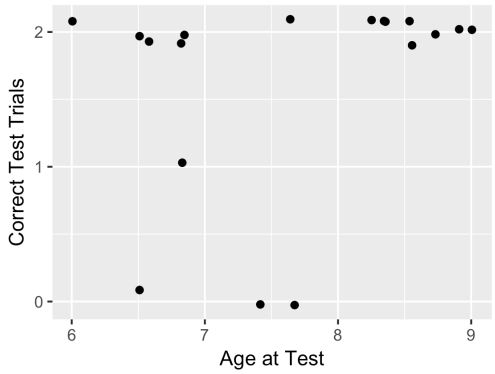

Figure 8. Children choose to gather more information for the harder discrimination even when this connection was not explicitly discussed.

Discussion

Across four experiments, we showed that both naïve adults and children represent the relative difficulty of discriminating populations and recognize that larger samples are required to discriminate populations with greater overlap. Consistent with a ground truth computational model, adults were sensitive to the relative difficulty of discrimination tasks, showing a graded increase in the amount of information they requested as problems became progressively more difficult. Adults were also sensitive to the cost of sampling information, both in contexts in which those costs were made explicit and those in which they were not. Thus they adjusted their confidence threshold downward, tolerating more uncertainty as discrimination problems became more difficult. Our results also suggest that children as young as six distinguish easy and difficult discrimination problems and know that they need larger samples to succeed in more difficult discriminations.

Although considerable work suggests that even infants represent the relationship between samples and populations (Xu & Garcia, 2008; Xu & Denison, 2009; Gweon, Tenenbaum, & Schulz, 2010), and in this sense are "intuitive statisticians," to our knowledge this is the first study to ask whether children represent these relationships metacognitively, distinguishing 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. In future work, we might ask whether children can make not just qualitative distinctions about the information required to distinguish populations but, like adults, graded inferences about the number of samples they would need as discrimination problems become more difficult.

However, the current work suggests that even children's intuitive statistics extends beyond the ability to recognize probabilistic relationships between samples and populations. 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. These results also suggest that even young children engage in metacognitive monitoring of the relative difficulty of discrimination problems and adjust their pursuit of information in response to this difficulty, suggesting that young children understand something about how to allocate resources to affect their knowledge state and allow for more effective learning.

Acknowledgments

We thank the Boston Children's Museum and all of the families that participated in this research, and the NSF Center for Brains Minds and Machines for funding. We also thank our reviewers for helpful suggestions about improvements to the computational model.

References

Fernandez-Duque, D., Baird, J. A., & Posner, M. I. (2000). Executive attention and metacognitive regulation. *Consciousness and cognition*, 9(2), 288-307.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464-481.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *Proceedings of the National Academy of Sciences*, *107*(20), 9066-9071.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one*, *7*(5), e36399.

Loewenstein, G. (1994). The psychology of curiosity: A review and reinterpretation. *Psychological bulletin*, *116*(1), 75.

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation*(Vol. 26, pp. 125-173). Academic Press.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203-216.

Siegel, M., Magid, R., Tenenbaum, J., & Schulz, L. (2014, January). Black boxes: Hypothesis testing via indirect perceptual evidence. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, *105*(13), 5012-5015.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, *112*(1), 97-104.