Cognition 126 (2013) 285–300

Contents lists available at SciVerse ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Rational variability in children's causal inferences: The Sampling Hypothesis

Stephanie Denison [a,*,1], Elizabeth Bonawitz [b], Alison Gopnik [b], Thomas L. Griffiths [b]

[a] University of Waterloo, Department of Psychology, 200 University Ave. West, PAS 4020, Waterloo, Ontario, Canada N2L 3G1
[b] Department of Psychology, University of California, Berkeley, United States

ARTICLE INFO

Article history:
Received 30 August 2011
Revised 26 October 2012
Accepted 29 October 2012
Available online 28 November 2012

Keywords:
Cognitive development
Causal learning
Sampling Hypotheses
Probability matching
Approximate Bayesian inference

ABSTRACT

We present a proposal—"The Sampling Hypothesis"—suggesting that the variability in young children's responses may be part of a rational strategy for inductive inference. In particular, we argue that young learners may be randomly sampling from the set of possible hypotheses that explain the observed data, producing different hypotheses with frequencies that reflect their subjective probability. We test the Sampling Hypothesis with four experiments on 4- and 5-year-olds. In these experiments, children saw a distribution of colored blocks and an event involving one of these blocks. In the first experiment, one block fell randomly and invisibly into a machine, and children made multiple guesses about the color of the block, either immediately or after a 1-week delay. The distribution of guesses was consistent with the distribution of block colors, and the dependence between guesses decreased as a function of the time between guesses. In Experiments 2 and 3 the probability of different colors was systematically varied by condition. Preschoolers' guesses tracked the probabilities of the colors, as should be the case if they are sampling from the set of possible explanatory hypotheses. Experiment 4 used a more complicated two-step process to randomly select a block and found that the distribution of children's guesses matched the probabilities resulting from this process rather than the overall frequency of different colors. This suggests that the children's probability matching reflects sophisticated probabilistic inferences and is not merely the result of a naïve tabulation of frequencies. Taken together the four experiments provide support for the Sampling Hypothesis, and the idea that there may be a rational explanation for the variability of children's responses in domains like causal inference.

1. Introduction

Human beings revise their beliefs throughout development, progressing towards an increasingly accurate portrayal of the world. Recent research suggests that young children perform this belief revision in a surprisingly systematic and rational way. In fact, a growing body of evidence suggests that children can revise their beliefs in a way that is consistent with Bayesian inference (Goodman et al., 2006; Gopnik, 2012; Gopnik et al., 2004; Kushnir & Gopnik, 2007; Schulz, Bonawitz, & Griffiths, 2007; Schulz & Gopnik, 2004; Xu & Tenenbaum, 2007). For example, Xu and Tenenbaum (2007) found that preschoolers can systematically integrate prior knowledge regarding the taxonomic structure of a domain with evidence provided by a speaker in order to apply the correct labels to a variety of objects in a word learning task. Similarly, Schulz et al. (2007) and Kushnir and Gopnik (2007) found that children's causal inferences rationally depend on both their prior beliefs and the observed evidence.

At first glance, the notion that preschoolers are capable of rationally updating their beliefs might seem

* Corresponding author.
E-mail address: stephanie.denison@uwaterloo.ca (S. Denison).
[1] Permanent address.

0010-0277/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.cognition.2012.10.010

incompatible with another striking feature of children's reasoning, namely its variability. Children will often express different beliefs or give a range of different answers to a question, even in the same testing session. This variability in responses might lead to some skepticism about children's reasoning abilities. For example, Piaget (1983) argued that children do not reason systematically about hypotheses until they reach the formal operational stage in late childhood. Since Piaget, some researchers have found evidence to corroborate this claim, demonstrating that children often appear to navigate randomly through a selection of different predictions and explanations (e.g. Siegler & Chen, 1998). In fact, Siegler has argued that such random variability may actually, in the long run, contribute to the learning process, comparing the learning process to such selection processes as biological evolution (Siegler, 1996). Nevertheless, his view is still that the variability itself is simply random rather than part of a rational process.

How can we reconcile the variability of children's responses with the apparent rationality of their inferences? Many rational accounts of children's behavior seem to at least implicitly assume that children are "Noisy Maximizers"—that they try to select the most likely hypothesis given the observed data, but they do so noisily (e.g. Kushnir & Gopnik, 2007; Sobel, Tenenbaum, & Gopnik, 2004). This noise is the result of cognitive load, context effects, or methodological flaws that lead children to stochastically produce errors. This accumulation of random noise accounts for the variability in children's responding. In this paper, we provide an alternative account of variability of children's responses—the "Sampling Hypothesis". On this view, at least some of the variability in children's responses may actually itself be rational. In particular, it may reflect an unconscious but systematic process that helps children select hypotheses that could explain the data they have observed.

The basic idea behind Bayesian inference is that a learner begins with a set of hypotheses of varying probability (the prior distribution). Then the learner evaluates these hypotheses against the evidence, and using Bayes rule, updates the probability of the hypotheses based on the evidence. This yields a new set of probabilities, the posterior distribution. But, for most problems, the learner can't actually consider every possible hypothesis—searching exhaustively through all the possible hypotheses rapidly becomes computationally intractable. Consequently, applications of Bayesian inference in computer science and statistics approximate these calculations using Monte Carlo methods. In these methods, hypotheses are sampled from the appropriate distribution rather than being exhaustively evaluated. A system that uses this sort of sampling will be variable—it will entertain different hypotheses apparently at random. However, this variability will be systematically related to the probability distribution of the hypotheses—more probable hypotheses will be sampled more frequently than less probable ones. The Sampling Hypothesis thus provides a way to reconcile rational reasoning with variable responding.

We present four experiments examining whether variability in children's inferences in a causal task might reflect this kind of sampling. We first describe the computational accounts that motivate the Sampling Hypothesis and highlight some connections to research with adults that are consistent with this hypothesis. We then review earlier research on children's variability, particularly the phenomenon of probability matching in reinforcement learning. This is followed by four experiments, designed to distinguish the Sampling Hypothesis from noisy maximizing and from simple reinforcement learning.

1.1. Belief revision and sampling

Demonstrating that people revise their beliefs in a way that is consistent with Bayesian inference does not necessarily imply that children or adults actually work through the steps of Bayes' rule in daily life. Evaluating all possible hypotheses each time new data are observed would not be feasible from either a formal or a practical standpoint, given the large number of hypotheses that would need to be considered. One way to think about how the mind may be approximating Bayesian inference is to start with good engineering solutions to this problem. Techniques for approximating Bayesian inference have already been developed in computer science and statistics, raising the possibility that human minds might also be using some version of these strategies.

One strategy for implementing Bayesian inference is Monte Carlo approximation, which is based on the idea of sampling from a probability distribution. Using sophisticated Monte Carlo algorithms, it is possible to generate samples from the posterior distribution without having to evaluate all of the hypotheses assigned probability by that distribution (Robert & Casella, 1999). Following this approach, people might be approximating Bayesian inference by evaluating a small sample of the many possible hypotheses that could account for the observed data. Formally, this sample should be drawn from the posterior distribution, $p(h|d)$, which indicates the degree of belief assigned to each hypothesis h given the observed data d. Recent work has shown how Monte Carlo methods that approximate this posterior distribution can account for human behavior in a range of tasks (Levy, Reali, & Griffiths, 2009; Sanborn, Griffiths, & Navarro, 2010; Shi, Feldman, & Griffiths, 2008). Other results suggest that people might be basing their decisions on just a few samples from appropriate probability distributions (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Mozer, Pashler, & Homaei, 2008). Indeed, in many cases an optimal solution is to take only one sample (Vul, Goodman, Griffiths, & Tenenbaum, 2009).

Sampling a hypothesis from a distribution necessarily involves a degree of randomness. However, the process is not entirely random in the conventional sense of giving equal probability to each alternative as when we flip a coin or roll a die. Hypotheses with high probability under the distribution will be sampled more often than those with lower probability. This strategy allows the learner to entertain a variety of hypotheses and in the long run, ensures that they will give more consideration to likely hypotheses but will not overlook a lower probability hypothesis that could turn out to be correct. The Sampling Hypothesis thus suggests that at least some of the variability that appears in children's responses should be systematic—determined by the posterior distribution over hypotheses.

If children are selecting hypotheses by sampling from a distribution, certain hallmarks of sampling should be present in their behavior. The signature of sampling is the fact that aggregating over numerous samples should return the original distribution. If instead learners generate a single "best guess," but do so noisily, then aggregating over numerous samples should result in an inaccurate reflection of the distribution, characterized by an overweighting of the most likely hypothesis. This leads to the key prediction of the Sampling Hypothesis: Response variability should reflect the posterior distribution of hypotheses. Of course, there may be additional noise in children's responses—because children may indeed stochastically produce errors in responding. However, if at least some of the variability in children's responding is captured by the Sampling Hypothesis, then responses should noisily reflect the posterior distribution, rather than noisily maximizing.

The idea that children might be selecting hypotheses by sampling from a probability distribution is related to two other phenomena: the "wisdom of crowds" effect (Galton, 1907; Surowiecki, 2004) and probability matching (Estes, 1950; Estes & Suppes, 1959). In the remainder of this section, we summarize the literature on these phenomena and relate them to the Sampling Hypothesis. We close the section by laying out the predictions that motivate our four experiments.

1.1.1. The wisdom of crowds

Galton (1907) observed that the average of the guesses of a group of people about the weight of an ox was closer to its actual weight than any of the individual guesses and he dubbed this phenomenon the "wisdom of crowds". Recent work exploring the wisdom of crowds effect links some instances of the effect to the Sampling Hypothesis. Vul and Pashler (2008) asked individuals to make guesses about a list of real-world statistics such as the percentage of the world's airports that are in the United States. Participants were assigned to two conditions. In the immediate condition, participants were asked to make guesses about a variety of statistics and then asked the questions a second time directly afterwards. In a delayed condition, the questions were asked for the second time 2 weeks later. As a whole, the average of the responses of all of the participants was close to the true value of the statistic, consistent with the wisdom of crowds effect. Averaging responses within a single participant also produced a more accurate estimate, showing that the merits of a crowd can be produced within a single person. However, there was a greater benefit of averaging guesses in the delayed group than in the immediate group.

Viewed through the lens of the Sampling Hypothesis, the results of Vul and Pashler (2008) suggest that their adult participants were sampling guesses from an internal distribution rather than always providing an optimal guess. The dependency between those samples depended on the amount of time that had passed, with the delayed group producing something closer to independent samples than the immediate group. The different effects of averaging in the two groups reflects the fact that the value of taking multiple samples increases when those samples are independent. Vul and Pashler suggest that these results may indicate that adults are sampling hypotheses. However, we do not know whether young children would behave in the same way.

1.1.2. Probability matching

Probability matching refers to the empirical observation of a match between the frequency of different responses and the probability that those responses are correct. There is extensive evidence for probability matching in non-human animals in the context of reinforcement learning (see Myers, 1976 and Vulkan, 2000, for reviews). If non-human animals are given a task in which one behavior is reinforced 33% of the time and the other is reinforced 67% of the time, they will often adjust their behavior to produce the first behavior 33% of the time and the second 67% of the time (Neimark & Shuford, 1959). From a reinforcement learning perspective this behavior is puzzling. Of course if the agent aims to maximize reward, the better strategy is to always produce the behavior that results in a reward 67% of the time. However, it has been suggested that the probability matching shown by animals such as fish, birds and rats that is sub-optimal in the context of individual reinforcement experiments may result from the fact that probability matching can result in optimal rewards in competitive foraging settings (Seth, 2011). That is, in a patchy environment with one food source producing, for example, 70% of the reward and the other producing 30% of the reward, some types of animals will match probabilities by distributing themselves in a 70:30 split to each food source (Harper, 1982; Kamil & Roitblat, 1985; Lehr & Pavlik, 1970). This matching behavior maximizes reward for the entire group, and so might be an evolutionarily determined strategy specifically designed for foraging contexts. An alternative hypothesis, however, is that the agent's aim might be to learn about the environment rather than simply maximize reward. By continuing to test the low probability option some of the time, the agent can begin to estimate the distribution of rewards in the environment (Stephens & Krebs, 1986). This alternative would be more closely related to the Sampling Hypothesis, with the assumption that these responses are intended to act as tests of hypotheses rather than to produce rewards.

Probability matching has also been shown in children in similar reinforcement paradigms. For example, if there are two levers, one that generates a reward when depressed 70% of the time and another that generates the reward 30% of the time, young children learn (over a series of 100 trials) to favor the lever which generates the reward more frequently. However, young preschoolers (i.e., 3-year-olds) actually tend more towards maximization when making probabilistic inferences, while 4- and 5-year-olds, like non-human animals, show probability matching in reinforcement learning (e.g. Jones & Liverant, 1960).

There has been much less work exploring probability matching beyond simple reinforcement learning. Will children probability match when they are formulating hypotheses rather than simply learning reinforced responses? In language learning paradigms, when children are inferring more abstract linguistic hypotheses, they do not probability match but rather maximize, in fact they trend more towards maximizing than adults do (Hudson Kam & Newport, 2005, 2009). In the case of causal inference, there are some suggestive results in which the variability of children's guesses does seem to be related to the probability of different hypotheses (e.g., Bonawitz & Lombrozo, 2012; Kushnir & Gopnik, 2007; Kushnir, Wellman, & Gelman, 2008; Sobel et al., 2004). However, this possibility has not been systematically tested—these patterns of responding may reflect matching, or they may reflect a noisy maximization process.

The Sampling Hypothesis predicts that the variability in children's hypotheses should reflect the posterior probability of those hypotheses—more probable hypotheses will be produced more often, while less probable hypotheses only appear occasionally. This is a kind of probability matching—the distribution of responses should match the posterior distribution—but it implies a level of sophistication that goes beyond what is typically assumed when the term "probability matching" is used. Rather than simply matching the frequency of rewarded responses or the frequency of particular linguistic constructions, we expect children to match the posterior probabilities of different hypotheses. By constructing tasks where these posterior probabilities vary, and where the posterior probabilities differ from the overall frequency of possible responses, we can separate the Sampling Hypothesis from other strategies that might result in probability matching.

1.2. Testing the predictions of the Sampling Hypothesis

Our experiments test the predictions of the Sampling Hypothesis using a causal learning task that does not involve reinforcement. In particular, children in our task had to learn about the probability of different hypotheses by considering the distribution of different colored blocks in a bag. When a bag has twice as many red blocks in it as blue ones, it is twice as likely that a random block that falls out of the bag will be red rather than blue. Other studies show that even infants are sensitive to this sort of distributional information and can use it to make probability judgments (Teglas, Girotto, Gonzalez, & Bonatti, 2007; Teglas et al., 2011; Xu & Garcia, 2008). This technique also allows us to fine-tune the probability of different hypotheses quite precisely by manipulating the number of blocks in the bag, and it means that children are never differentially reinforced for their responses. Instead, the children had to use the distribution to inform their guesses about which block had fallen from the bag and caused an effect.

We use this paradigm as the basis for a series of experiments. Experiment 1 tests the basic prediction of probability matching in two ways and examines the pattern of dependencies in children's responses as a function of time, as in Vul and Pashler (2008). Experiments 2 and 3 provide a more fine-grained investigation of probability matching, varying the probabilities of different hypotheses and examining how this affects children's responses. Experiment 4 investigates the level of sophistication of children's probability matching, using a more complicated procedure to determine the probabilities of different hypotheses; this ensures that children were not using a simpler strategy of matching responses to the number of chips in the bag.

2. Experiment 1: Sampling and dependency

Experiment 1 examined whether children's behavior would match the basic prediction of probability matching in our causal learning task. In addition, we took the opportunity to explore any patterns of dependency that appear in children's judgments and to see how these are influenced by a delay. On each of three trials, children were asked to guess the color of an unseen block that activated a novel toy, taking into account the fact that the block fell out of a bag containing a 4:1 ratio of red to blue blocks. Children were split into two conditions: the short wait condition, where children saw the three trials immediately following one another in a single testing session, and the long wait condition, where children saw each trial 1 week apart. We test probability matching in two ways: We predict that across children, the distribution of the first guess will closely match the distribution of blocks in the bucket. We also predict that, when the dependency between guesses is minimized, the distribution of the children's three guesses will similarly reflect the posterior distribution. Following Vul and Pashler (2008), we expect that children in the long wait condition will show less dependency between guesses than children in the short wait condition. Thus, the distribution of guesses in the long wait condition should be closer to the posterior distribution than in the short wait condition.

2.1. Methods

2.1.1. Participants

Forty 4- and 5-year-olds were tested individually in quiet rooms at preschools located on the U.C. Berkeley campus. The children were randomly assigned to one of two conditions, each consisting of 20 children: the long wait condition (12 females; Mean age = 54.1 months; R = 48.4–62.8 months) and the short wait condition (9 females; Mean age = 53.5 months; R = 48.1–59.0 months). One additional child was tested and excluded due to failing a comprehension check. The children's ethnicities and socioeconomic status reflected the composition of the area.

2.1.2. Stimuli

A large box (12 in. (30.48 cm) × 12 in. (30.48 cm) × 18 in. (45.72 cm)) constructed out of cardboard and covered in yellow felt was used. A toy consisting of a transparent sphere connected to a cylindrical shaft was inserted in a hole in the top of the box on the front right corner such that only the sphere (which had a spinner and lights) was visible to the children. The toy was activated by pressing a button on the shaft, causing the sphere portion to light up and play music. An opaque activator bin, made of a plastic container and construction paper, was placed on the back left corner of the box. Additional stimuli included red, blue, and green domino sized wooden blocks; a rigid green bag; and a transparent container (see Fig. 1).

2.1.3. Procedure

Each testing session in all experiments was videotaped for data retrieval and a second experimenter recorded all responses online.

In both the long wait and short wait conditions, the experimental session began with the child and experimenter sitting across from one another at a table with the large yellow box in between them—the front side facing the child and the back side facing the experimenter. The experimenter introduced children to the large yellow box saying, "This is my big toy and I'm going to show you how it works." The experimenter then took two blocks of each color (red, blue, and green) and placed them on the table. One block at a time, the experimenter picked up a block of each of the three colors and dropped it into the activator bin. She showed the children that when a red block or a blue block is placed in the activator bin, the toy lights up and plays music, and when a green block is placed in the bin, the toy does not activate. In reality, the experimenter was surreptitiously activating the toy by pressing a button hidden from view.

Previous work using this causal scenario suggests that children (and even adults) find this manipulation compelling and that use of the ineffective green block helps to establish that the red and blue blocks cause the effect (Bonawitz & Lombrozo, 2012).

In a comprehension check, children were asked whether each of the three colors would make the machine go. The experimenter picked up a block and asked, "What will happen if I put a [red, blue, green] block into the machine?" In order to be included in analyses, children had to remember that red and blue blocks make the toy go and green blocks do not. Order of colors was randomized across children for the initial demonstrations and the comprehension check, except that the green block was never demonstrated first in the initial demonstrations.

On Test Trial 1, the experimenter and child counted out 20 red blocks and 5 blue blocks (i.e., an 80:20 distribution) one at a time and placed them into a transparent container. Which block color was counted first was counterbalanced across children. After counting the blocks, the experimenter asked, in the same order as she counted, "So how many red ones did we count? And how many blue ones?" and corrected the child if (s)he was incorrect. Then she shook the blocks in the container to mix them and poured them into the rigid opaque bag. She placed the container upside down in front of the activator bin on the yellow box and placed the bag on top of the container. She then 'accidentally' knocked the bag over toward the activator bin. Just after the bag fell over, the experimenter activated the toy and said, "Oh, I think one of the blocks must have fallen into the toy and made it go! Can you tell me which color it was?" Once the child answered the question, the experimenter pretended to remove the block while turning off the toy. Finally she asked, "And why do you think it was a [red, blue] block?" Occasionally children initially responded "both" when asked which color fell in. The experimenter would then prompt the child by saying, "The toy only works when just one block falls in. What color do you think it was?"

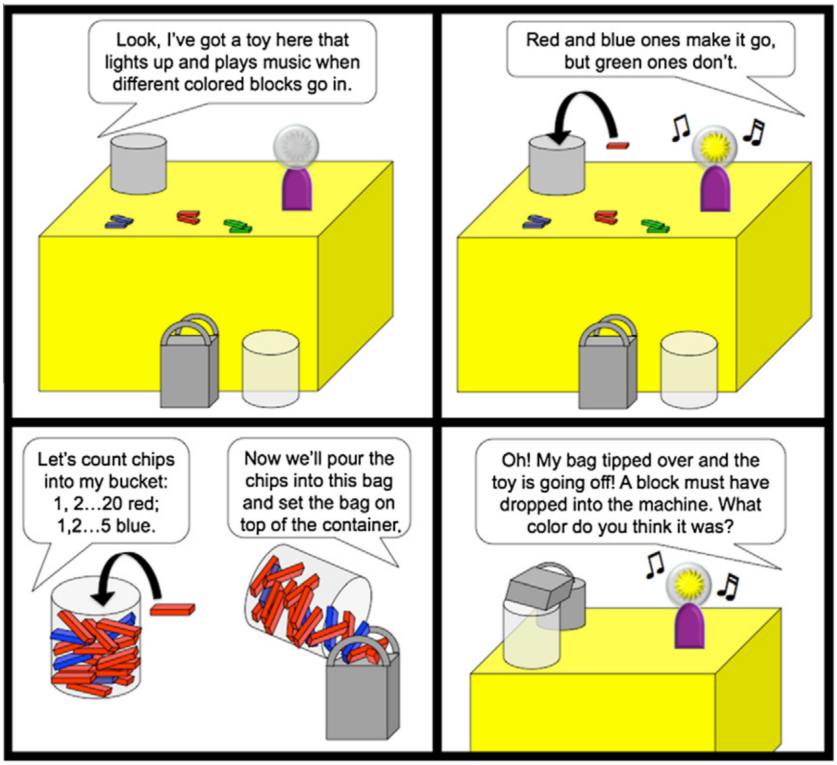
Look, I've got a toy here that lights up and plays music when different colored blocks go in.
Red and blue ones make it go, but green ones don't.
Let's count chips into my bucket: 1, 2…20 red; 1,2…5 blue.
Now we'll pour the chips into this bag and set the bag on top of the container.
Oh! My bag tipped over and the toy is going off! A block must have dropped into the machine. What color do you think it was?

Fig. 1. Stimuli and procedure used for testing the Sampling Hypothesis in children.

In the short wait condition, once children provided an answer for Trial 1, the experimenter began Trial 2 by saying, "That was kind of funny how I accidentally tipped the bag over and it made the toy go. Should I try to make that happen again? First we have to count our blocks again." The second and third trials progressed exactly the same as Trial 1, with 20 red and 5 blue blocks. The experimental session took approximately 9 min.

The long wait condition was identical to the short wait (20 red and 5 blue blocks on all trials) except that children completed Trial 1 in the first testing session, Trial 2 in a second testing session 1 week later, and Trial 3 in a third testing session 1 week after Trial 2. Children were reminded that the blue and red blocks make the machine go and green blocks do not at the beginning of each testing session. Each experimental session (i.e., each trial) took approximately 3 min.

2.2. Results

There were no age differences between groups ($t(38) = 0.11$, $p = .544$). Responses were coded by first author and reliability coded by a research assistant blind to experimental hypotheses for 75% of the trials. All responses uniquely and unambiguously were either "red" or "blue" and agreement was 100%. There was no effect of gender or which color was counted first in either of the two conditions; we collapsed across these variables for subsequent analyses.

2.2.1. Probability matching on initial trial

As should be expected, there were no differences between conditions for children's first predictions, χ^2 (1, $N = 40$) = 1.9, $p = .168$. To assess whether or not children probability matched, we averaged the first response of children in both the long wait and short wait conditions. Overall, children's responses reflected probability matching (28/40 trials, 70% providing the more probable chip response and 30% providing the less probable chip response). Though there was some noise not accounted for by probability matching, the children were not simply randomly guessing, as responses were significantly different from chance (binomial test, $p = .017$) but not significantly different from the predicted distribution of .8 (binomial test, $p = .175$). Similarly, children did not appear to "maximize" by always providing the most probable response (i.e. always choosing the red block), or responses would have approached ceiling.

2.2.2. Probability matching across all trials

The previous result suggested that there was probability matching across children – a kind of "wisdom of crowds" effect. Was there evidence of probability matching within individual children's responses, as in Vul and Pashler (2008)? We first computed the predictions of independent sampling; that is, given probability θ of sampling a particular block, what should the distribution of three responses look like? Because there are two possible responses (red (r) or blue (b)) and there are three trials, there are simply $2 \times 2 \times 2$ or 8 possible hypotheses (rrr, rrb, rbr, rbb,…,bbb). Thus, assuming independence between trials, the probability of any particular hypothesis (e.g., rrb) is simply the probability of sampling each block (i.e. $(.8) \times (.8) \times (.2)$). In this way, we can compute probabilities for all eight hypotheses. We compared this expected distribution to the observed distribution given by children in the short wait and long wait conditions (see Table 1). Both the long wait condition and short wait condition were significantly different from the expected distribution (long wait: χ^2 (7, $N = 20$) = 33.91, $p < .001$; short wait: χ^2 (7, $N = 20$) = 77.75, $p < .001$).[2] However, there was also a significant difference between children's responses in the short wait condition and the long wait condition, χ^2 (7, $N = 40$) = 22.3, $p = .002$, suggesting that the manipulation had an effect on children's pattern of responding.

An examination of children's response patterns shows that two children in the long wait condition produced the "blue, blue, blue" response pattern, regardless of its extremely low predicted probability. When we exclude these two children from analyses, the pattern of the remaining 18 children's responses is only marginally different from the predicted distribution, χ^2 (6, $N = 18$) = 12.7, $p = .09$. This suggests that the responses from these two children heavily contributed to the initial difference found between the expected and empirically produced distributions. In contrast, the children's responses in the short wait condition were much further removed from the expected distribution. By far the most frequent response was for children to alternate responses across trials in spite of the relatively low probability of that hypothesis.

2.2.3. Dependency measures

We investigated the dependency of children's responses in two ways. A quick examination of Table 1 suggests that children in the short wait condition were alternating guesses, a strategy that demonstrates dependencies among those responses. To directly compare the two conditions, we coded children's responses in terms of whether they repeated a guess (e.g. "red" then "red" again) or alternated (e.g. "red" then "blue"), both patterns that would reflect dependencies among the responses. Comparing condition by repetition/alternation revealed significant differences both when we coded for repetition/alternation over all three responses, Fisher Exact ($N = 33$), $p < .001$, and when we coded for repetition/alternation over two responses, Fisher Exact ($N = 80$), $p < .001$.[3] Children were more likely to repeat or alternate guesses in the short wait than in the long wait condition.

Another way to think about dependency is to model children's responses as a Markov process and consider the transition matrix. We computed the empirical frequencies with which children moved from a "red block" response to a "blue block" response, and so forth (see

[2] Because the approximation to the χ^2 distribution is unreliable with small cell entries, we computed the null distribution numerically. We generated 10,000 contingency tables with these frequencies, computed χ^2 for each, and then computed p values by examining the quantile of the observed χ^2 value.

[3] We use Fisher Exact tests for consistency due to small sample sizes on some tests.

Table 1
Experiment 1: Pattern of responses expected under independent sampling compared with frequencies in the long wait and short wait conditions.

Responses	Expectation	Long wait (frequency)	Short wait (frequency)
Red, red, red	.512	.500 (10)	.050 (1)
Red, red, blue	.128	.050 (1)	.050 (1)
Red, blue, red	.128	.100 (2)	.500 (10)
Red, blue, blue	.032	.150 (3)	.000 (0)
Blue, red, red	.128	.000 (0)	.050 (1)
Blue, red, blue	.032	.050 (1)	.300 (6)
Blue, blue, red	.032	.050 (1)	.050 (1)
Blue, blue, blue	.008	.100 (2)	.000 (0)

Table 2). If children are producing independent samples, the probability of producing a particular response should be the same regardless of the previous response. However, this analysis revealed a strong dependency between responses in the short wait condition, Fisher Exact ($N = 20$), $p < .001$, and a much weaker dependency in the long wait condition, Fisher Exact ($N = 20$), $p = .029$. These results suggest that although children's pattern of responses in the long wait condition was close to the predicted distribution, there were still some dependencies between a single child's guesses. Indeed, this is particularly suggested by the anomalous frequency of the blue, blue, blue responses in the long wait condition, responses that might well have reflected a pattern of dependency even in the long wait condition; that is, these children may simply have repeated the response they made on the previous trial.

2.3. Discussion

This experiment examined whether the variability in children's hypotheses in a simple causal reasoning task reflected sampling from a probability distribution. The results provide evidence in support of the main prediction of the Sampling Hypothesis: children were probability matching. As a group, children provided a percentage of red and blue initial guesses that corresponded with the actual distribution of red and blue blocks in the population, rather than maximizing and choosing the red block on every guess or randomly guessing each color 50% of the time. Children in the long wait condition also generated a pattern of guesses that reflected probability matching within children across trials. The distribution of responses across trials reflected a sampling process more clearly in the long wait than in the short wait condition. The results thus suggest that this was due to the fact that the responses in the long wait condition were closer to a set of independent samples from the relevant distribution than were the responses in the short wait condition.

The Sampling Hypothesis suggests that in both short and long wait conditions children respond in a way that reflects sampling after each new query, and because responses are sampled close together, there are likely to be greater dependencies between guesses in the short wait condition. These dependencies could arise for a number of reasons; for example, recent research suggests that children's sensitivity to the knowledge and helpfulness of an interviewer can explain children's tendency to switch guesses on repeated questioning (Gonzalez, Shafto, Bonawitz, & Gopnik, 2012). Regardless of the specific factors that cause greater dependency when samples are generated over shorter intervals, the overall response pattern of the preschoolers is consistent with the results of Vul and Pashler (2008) with adults. There is some evidence for a "crowd within" effect and the effect is weaker when there is more dependency between responses.

Table 2
Experiment 1 transition matrices in the two conditions.

	Long wait		Short wait	
	Next *r*	Next *b*	Next *r*	Next *b*
Current *r*	21	7	4	17
Current *b*	4	8	18	1

While the results of this experiment seem consistent with the Sampling Hypothesis, they only provide preliminary evidence against alternative accounts of variability in children's responses. These children did not seem to be responding at chance or to be maximizing, but they might have been noisy maximizers. Children might have simply followed a strategy of choosing the more probable chip every time but sometimes failed to succeed because of memory or attention limitations, and this noise might have just happened to lead to a 70:30 distribution of guesses. We cannot know for certain that children were not maximizing without varying the proportion of blocks of different colors and examining the effect that this has on children's responses. This is what we did in Experiment 2.

3. Experiment 2: varying proportions

To determine whether children's responses truly reflect probability matching with some noise or instead reflect noisy maximization where all the variability is the result of noise, we manipulated the ratio of red to blue blocks in our causal learning scenario. Three groups of children were presented with different distributions of blocks with ratios of 95:5, 75:25, or 50:50. This design allows us to tease apart four possible strategies children might use in this task: (1) They may guess randomly, in which case children in all three conditions should choose each block on roughly 50% of trials. The probability matching results from Experiment 1 suggest this is not the case; however, additional data would provide further support for this claim. (2) They may use a maximization strategy and choose the majority-color block near ceiling in both the 95:5 condition and the 75:25 condition. Because children in Experiment 1 initially produced the more probable response 70% of the time (rather than 100%), we can also begin to rule out this account; however, additional data would also be useful. (3) They may maximize with noise—showing above chance predictions but no difference between the 95:5 and 75:5 condition (the noisy-max strategy). (4) They may match sampling from the distributions (as indicated by the Sampling Hypothesis) with a small amount of noise. In this case we should see a decreasing preference for the more probable block such that children in the 95:5 condition would guess that the majority-color block activated the machine most of the time, children in the 75:25 condition would choose the majority-colored block less often, and children in the 50:50 condition would randomly choose between the two colors. Thus, by manipulating the ratios, we can tease apart the noisy-max strategy from the predictions of the Sampling Hypothesis and reveal which strategy children actually use.

3.1. Method

3.1.1. Participants

Participants were 75 four- and five-year-old children who were either attending a U.C. Berkeley campus preschool and were tested in a quiet room in their school or were recruited and tested at a local museum. Children were split into three conditions: the 95:5 condition consisted of 25 children (12 females; Mean age = 58.9 months; *R* = 48.1–71.5 months); the 75:25 condition consisted of 25 children (8 females; Mean age = 58.3 months; *R* = 49.3–67.1 months); the 50:50 condition consisted of 25 children (15 females; Mean age = 61.8 months; *R* = 48.6–71.9 months). An additional 8 children were tested but not included in the final analyses. Children were excluded for interference from a sibling or parent (95:5 condition = 1 child; 50:50 condition = 1 child) or failing the comprehension check (95:5 condition = 1 child; 75:25 condition = 2 children; 50:50 condition = 3 children).

3.1.2. Stimuli

The stimuli were the same as in Experiment 1.

3.1.3. Procedure

The procedure was the same as Experiment 1, Trial 1 except that the distribution of red and blue blocks was manipulated across three conditions: In the 95:5 condition, the experimenter and child counted out 19 blocks of one color (either red or blue—counterbalanced across children) and 1 block of the other color. In the 75:25 condition, there were 15 blocks of one color and 5 blocks of the other color, and in the 50:50 condition, there were 10 blocks of each color. Which block color was counted first was counterbalanced. The experimental session lasted approximately 3 min, and children at the museums received a small gift for participating in the experiment.

3.2. Results

All children generated a unique and unambiguous response of either "red" or "blue;" an assistant blind to condition and hypotheses coded 40% of the trials in each condition and agreement was 100%. There was a marginally significant difference in the ages of children across conditions ($F(2,72) = 2.73$, $p = .07$). This difference was largely due to the children in the 50:50 condition being slightly older than children in the other two conditions, as children in the 95:5 and the 75:25 conditions did not differ reliably in age, $t(48) = .03$, $p = .863$. There were no effects of gender, which color block (red or blue) was used as the majority color, or which color was counted first in any of the three conditions; we collapsed across these variables for subsequent analyses.

3.2.1. Probability matching

Children in the 95:5 condition guessed the majority-color block on 21/25 (84%) trials, which was significantly different from chance (binomial test, $p < .001$) and not significantly different from the expected (95%) distribution (binomial test, $p = .07$). Children in the 75:25 condition guessed the majority color block on 15/25 (60%) trials; this was not significantly different from either chance or the expected frequency of .75 (binomial tests, $p = .42$; $p = .14$, respectively). As predicted, children in the 50:50 condition chose each block roughly equally—the red block on 14/25 trials and the blue block on 11/25 trials, which did not differ from chance (binomial test, $p = .689$). A comparison of children's responses in the 95:5 condition to children's responses in the 75:25 condition reveals a marginally significant difference in choosing the majority color block between these conditions ($p = .06$, one tailed). These results thus provide some additional support for the hypothesis that the children were probability matching.

3.2.2. Comparing the probability matching and the noisy-max model

To directly compare the probability matching and the noisy-max strategy, we performed three additional analyses. Recall that the sampling prediction is that the proportion of blocks of a particular color in the sample would have a linear effect on the children's responses—as the proportion of blue blocks goes up, children should be proportionately more likely to guess that the causal block was blue. In contrast, noisy-max predicts no difference between those groups. We performed a logistic regression to test whether or not assignment to a particular condition significantly increased the log odds ratio of guessing the majority color block. Because the method for the initial predictions (i.e., Trial 1) in the 80:20 condition of Experiment 1 are identical to the 95:5 and 75:25 conditions here, we included these data in our analyses, providing yet another distribution to test. We dummy coded children's responses into 1's and 0's—children received a 1 for guessing the majority color block and a 0 for guessing the minority color block (the 50:50 condition was arbitrarily coded such that the color block children saw first when the distributions were counted was given a score of 1). We entered the data from all four conditions into the model and found that the odds ratio for choosing the majority color block was significant in the 95:5 and 80:20 conditions but not in the 75:25 condition (see Table 3 for significance tests for all conditions).

In our second analysis, we conducted a logistic regression with distribution of blocks in the bag (i.e., condition) as an ordered predictor variable (95:5; 80:20; 75:25; 50:50). We scaled the Condition variable to more accurately reflect the magnitude of the differences between conditions: the 95:5 Condition was scaled to log(19); the 80:20 Condition = log(4); the 75:25 Condition = log(3); and the 50:50 Condition = log(1). The regression found evidence for a linear increase in the proportion of choices of the more numerous block based on condition (Wald test: df = 3; $z = 2.99$, $SE = .199$, $p = .003$). This is consistent with the hypothesis that the children's responses are sensitive to the distribution of blocks, with a linear relationship being what we should expect if children are probability matching. The analysis also confirms that this probability matching is imperfect, as the coefficient for the linear model is .595, 95% Confidence Interval = (0.205, 0.985).

Finally, we compared the likelihoods of observing the data under a model of probability matching and a model of the noisy-max account. Both models predict random responding for the 50:50 condition, so we did not include responses for this condition in either model. We did include the responses in the 80:20 condition in Experiment 1. Both models are a mixture of chance responding and a variable θ, mediated by a free parameter α. The probability matching model is: $\alpha * \text{chance} + (1 - \alpha) * \theta$, where chance given two chips is .5 and θ reflects the probability of that block by condition (i.e. $\theta = .75$ in the 75:25 condition). Because the noisy max model predicts always selecting the maximally likely hypothesis, θ is 1 for all conditions such that the noisy max model is simply: $\alpha * \text{chance} + (1 - \alpha)$. The single free parameter α can be thought of as the parameter that varies how many children are chance responding and how many children perfectly match to θ. Thus, when $\alpha = 1$ chance responding is predicted, and when $\alpha = 0$ all responses are driven by θ. We selected the α that best fit the data for each model. The α that best accounted for the noisy-max model was .58, indicating that the best fit for this model assumes that more than half the children guess at chance and predictions should always fall at around 71% of the more probable block. The α that best fit the probability matching model was .3, indicating that the majority of children probability match but a few children guess at chance and draw response distributions towards .5. The probability matching model was a better fit to the data (log-likelihood = −46.7) than the noisy-max model (log-likelihood = −48.0); see Fig. 2.

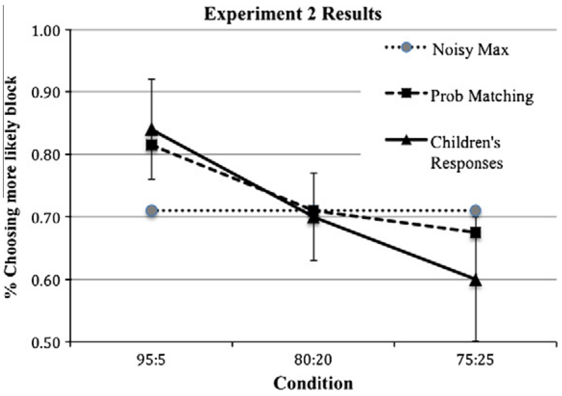

Fig. 2. Results of children's predictions in Experiment 2 and the 80:20 first predictions from Experiment 1, as compared to predictions of the noisy-max and the probability matching models.

3.3. Discussion

Children's tendency to guess the majority-color block decreased as the proportions in the distribution became less extreme—from 95:5 to 80:20 to 75:25 to 50:50. In addition, children's behavior did not deviate from the performance predicted by the Sampling Hypothesis in any of the conditions. There was, however, evidence against the maximizing hypothesis—children guessed the majority-color block more often in the 95:5 condition than in the 75:25 condition, and the probability matching model outperformed the noisy-max model. Children in the 95:5 condition also chose the majority color block at greater than chance levels; combining these results with those from Experiment 1 rules out the possibility that children were confused by the data and were simply responding at chance and suggests that a noisy-max model that requires a high level (58%) of chance responding to fit the data is unlikely.

Table 3
Experiment 2: Wald tests for the 95:5; 80:20 (Experiment 1); and 75:25 conditions.

Condition	Odds estimate	Std. error	*z*-Value	*p* Value
95:5	1.899	.678	2.801	.005
80:20	1.089	.531	2.052	.04
75:25	0.647	.574	1.127	.26

There are however, other possible formulations of the noisy max model that might better fit the data. It is possible that children's memories of the distributions were affected by noise and children maximized on the remembered counts, rather than the true counts. A noisy max model that does not predict a constant amount of noise for each distribution, but rather a larger amount of noise as the distributions approach chance could provide a better fit than the mixture model that we presented. Such a model would require additional free parameters to account for whether noise was contingent on the ratio between the majority color chip and all other chips of any color, or between the majority color chip and only the next most populous chip, as we will explore in Experiment 3. Although our current experiment cannot determine whether or not the remembered counts were affected by a model that takes into account this kind of noise, it seems somewhat unlikely due to the fact that children were asked about and reminded of the number of each color block after counting, and the sampling event took place less than 1 min later. Nonetheless, although we may be able to explain the data with a potentially less parsimonious process model, the overall pattern of a linear decrease in choosing the majority color block as the distributions become less extreme is consistent with probability matching.

There was a linear decrease in the number of guesses indicating the majority color block from the 95:5 to the

80:20 (Experiment 1), to the 75:25 conditions; however we did not find the predicted statistically significant difference between children's responses in the 95:5 and 75:25 conditions, but only a trend towards such a difference. Moreover, children's responses in the 75:25 condition were not significantly different from chance responding of 50%. This may be due to the fact that our experimental design, which allows for just one data point from each participant, and where chance is 50%, lacked enough statistical power to uncover these differences. To further investigate whether or not children are producing responses consistent with the Sampling Hypothesis, we assessed probability matching in a third experiment. This additional experiment introduces three potential hypotheses from which children can choose. This moves chance responding from 50% to 33% and allows us to test whether or not children produce responses consistent with probability matching when three potential responses are possible, rather than just two.

4. Experiment 3: varying proportions with three alternatives

In this experiment, we tested the probability matching prediction with a different, more complex set of hypotheses. Do children continue to produce guesses that reflect probability matching when more than two alternative hypotheses are available? In this experiment, children were given distributions that included three different colors of objects, all of which made the toy activate. The design was similar to Experiment 2, as the distributions were systematically manipulated across two conditions: an 82:9:9 condition and a 64:18:18 condition.

4.1. Method

4.1.1. Participants

Participants were 100 four- and five-year-old children who were either attending a U.C. Berkeley campus preschool or were recruited and tested at a local museum. Children were split into two conditions: the 82:9:9 condition consisted of 50 children (28 females; Mean age = 58.6 months; R = 50–70 months); the 64:18:18 condition consisted of 50 children (23 females; Mean age = 58.7 months; R = 48–71 months). An additional 9 children were tested but not included in the final analyses. Children were excluded for interference from a sibling or another child (82, 9, 9 condition = 1 child; 64, 18, 18 condition = 3 children); walking around to the back of the machine and discovering the way the machine truly worked (64, 18, 18 condition = 1 children); experimenter error (64, 18, 18 condition = 3 children); or refusing to agree that any blocks of any color made the machine work (82, 9, 9 condition = 1 child).

4.1.2. Stimuli

We used the stimuli from Experiments 1 and 2 and a second analogous set of materials. This second set of materials consisted of a box made of cardboard and multicolored construction paper (mostly black and orange), and it had an airplane toy that spun and lit up when the button was pressed. The objects for counting were poker chips covered in black, white, and yellow electrical tape and the bag was yellow.

4.1.3. Procedure

The procedure unfolded as in Experiment 2, except that children were shown that objects of all colors make the machine work. The comprehension check simply consisted of the experimenter asking the children what colors the blocks were and if they made the machine work. Children in the 82:9:9 condition counted, for example, 18 red, 2 blue and 2 green blocks or 18 white, 2 yellow and 2 black chips. Children in the 64:18:18 condition counted, for example, 14 red, 4 blue and 4 green blocks or 14 white, 4 yellow and 4 black chips. The majority color block or chip was counterbalanced across children in both conditions.

4.2. Results

All children generated a unique and unambiguous response of one of the six colors, and an assistant blind to condition and hypotheses coded 50% of the trials in each condition and agreement was 100%. There were no differences in the ages of children across conditions ($F(1,98) = 0.01$, $p = .921$). There were no effects of gender, which toy was used ($N_{New\ Toy(82:9:9)} = 8$; $N_{New\ Toy(64:18:18)} = 5$), which color block was used as the majority color, or which color was counted first in either condition; we collapsed across these variables for subsequent analyses.

4.2.1. Probability matching

Children in the 82:9:9 condition guessed the majority-color block on 36/50 (72%) trials, which is significantly different from chance of 33% (binomial test, $p < .001$) and not significantly different from the expected frequency of 82% (binomial test, $p = .11$). Children in the 64:18:18 condition guessed the majority color block on 24/50 (48%) trials; significantly different from chance of 33% (binomial test, $p = .04$) and but also significantly different from the expected frequency of .64 (binomial test, $p = .03$), consistent with the idea that while there is a pattern of probability matching, there is also some noise in children's responding. Importantly, and consistent with the probability matching model, children in the 82:9:9 condition chose the majority color more often than children in the 64:18:18 condition, χ^2 (1, $N = 100$) = 5.06; $p = .025$).

4.2.2. Comparing the probability matching and the noisy-max model

As with Experiment 2, we compared the likelihoods of observing the data under a model of probability matching and a model of the noisy-max and compared the majority color chip choices to the combined minority color chip choices. Whether using the α that best fit the data for each model in Experiment 2 or choosing a new α for each model that best fit only the data from Experiment 3, the probability matching model was a better fit to the data (log-likelihood = −65.7 when α set from Experiment 2 ($\alpha = .30$); log-likelihood = −65.4 when α fit to data ($\alpha = .43$)) than the noisy-max model (log-likelihood = −70.1 when

α set from Experiment 2; log-likelihood = −67.3 when α fit to data (α = .80)). Note also that while the best value for α in the noisy max model varies greatly for the results of Experiments 2 and 3 (from .58 in Experiment 2 to .80 in Experiment 3) the best α in the probability matching model is relatively constant across the two experiments (.30 and .43 respectively) indicating that the probability matching model is a more robust model.

4.3. Discussion

The data from the experiments we have described thus far support the probability matching prediction of the Sampling Hypothesis. However, the factors that are influencing children's hypotheses in these tasks may be more or less sophisticated. For example, children's attention may have simply been more strongly drawn toward the majority-colored block in the 95:5 condition (Experiment 2) and the 82:9:9 condition (Experiment 3) because there were more of these blocks shown overall compared to the minority-color blocks. Although using a low-level, naïve frequency matching strategy to make inferences on these tasks would produce probability matching behavior, ideally, we would like to confirm that children are instead reasoning in a more sophisticated way about the probability that each type of block fell out of the bag. One way of showing that children are using a more advanced strategy than simple frequency matching is to test whether they can consider the process by which the data were generated, effectively integrating prior probabilities into their judgments.

5. Experiment 4: beyond frequency matching

In Experiment 4, we designed a task that directly pits naïve frequency matching against a more sophisticated sampling strategy. The design consists of a set of events in which the more numerous color block was actually less likely to have made the machine go than the less numerous color block. For example, there might be more red blocks overall, but it is more likely that a blue block fell into the machine. We asked whether children correctly reasoned about how a sample could be generated by integrating the distributional information overall with information about the physical separation of the population of objects into two distinct distributions. Previous research using looking-time with infants suggests that they can compute probabilities in situations where overall numerosity and probability conflict, based on a physical constraint on the sampling process (Denison & Xu, 2010; Teglas et al., 2011).

To disentangle probability from numerosity, we split the blocks into two separate containers. The experimenter counted 14 red blocks and 6 blue blocks into Container 1 and 2 blue blocks into Container 2. Hence, there were many more red blocks than blue blocks overall. In what we will call the separate distributions condition, the blocks were transferred from each transparent container into corresponding separate opaque bags, then a single bag was selected at random and this bag was knocked over, causing the machine to activate. Correct predictions for children in this condition require the integration of multiple sources of information: First, children must realize that the population of objects is now physically separated so that the objects in each container cannot transfer from one distribution to another or simply be summed over. Second, if children assume that the sampled bag was randomly selected, then they must combine the 50% probability of choosing either distribution (bag) with the probability of sampling a particular object color within each distribution. Thus, the probability of a blue block falling out is: the probability that the first bag was selected (50%) times the probability of a blue block being selected given that bag (6/20), plus the probability of the second bag being selected (50%) times the probability of a blue block falling from that bag, given selection (100%). This equals a sum total 65% probability that a blue block activated the machine, in spite of the fact that only 36% of the blocks were blue. If, on the other hand, children are engaging in a simpler strategy of naïve frequency matching, they should probability match across the entire population. That would mean choosing the more numerous red blocks rather than the more probable blue blocks: 64% of the blocks overall are red, but given the causal situation, there is only a 35% probability that a red block activated the machine.

Children in a second control group, called the merged distributions condition, saw the blocks being separated into two transparent containers in the same proportions as described above. However, these children then saw all of the blocks being poured into a single opaque bag so that the distributions were no longer separated for the remainder of the procedure. We expect that children in this condition, like those in Experiments 1 and 2, will probability match across the entire population, favoring the more numerous red blocks in their guesses.

5.1. Method

5.1.1. Participants

Participants were 33 four- and five-year-old children who were either attending a U.C. Berkeley campus preschool or were recruited and tested at a local museum. The children were randomly assigned to two conditions: the separate distributions condition (20 children; 10 females; Mean age = 56.4 months; R = 49.3–62.3 months) and the merged distributions condition (13 children; 8 females; Mean age = 57.8 months; R = 50.0–70.7 months). In the separate distributions condition, no additional children were tested and excluded, but because there were three trials in this task, two children had one of the three trials excluded for failing a comprehension check. In the merged distributions condition, two additional children were tested but excluded from final analyses, one because of experimenter error and another for failing the comprehension checks on every trial. Three children had a single trial excluded for failing to pass the comprehension check for that particular trial.

5.1.2. Stimuli

Identical stimuli were used for both conditions. Because we know that children show dependence between responses when asked a question on the same toy (Experiment 1), in this experiment, we introduced a completely novel toy for each trial, with novel activation rules and novel activator objects. This allowed us to ensure that a child's response on the first trial would not influence their responding on subsequent trials. For Trial 1, identical stimuli to Experiment 1 were used with the following additions: two transparent containers were used rather than one, two identical blue rigid bags were used rather than the one green bag, and two cards mounted on black construction paper with color-printed pictures depicting the separate distributions of blocks contained in the transparent containers were used.

For Trial 2, a different large box made of cardboard and decorated with multi-colored (mostly purple, green, and yellow) construction paper and a toy fan that functioned similarly to the sphere and cylinder toy were used. The blocks used for Trial 2 were approximately 1 in (footnote 3). Lego pieces covered in orange, purple, and brown electrical tape. The two identical bags were yellow and green, and there were two pictures depicting the two separate distributions of the Lego blocks in the transparent containers.

For Trial 3, the new box and poker chips used for some of the children in Experiment 3 were used (yellow, black and white chips). The two identical bags were yellow with flowers, and there were two pictures depicting the distributions of the poker chips in the transparent containers.

5.1.3. Procedure

Trial 1 proceeded as in Experiment 1 until the end of the comprehension check. The experimenter then brought out two transparent buckets and placed them in front of the child about a foot apart on the table. The experimenter said, "Look at these two buckets. Let's count 14 red blocks into this bucket here (pointing to the bucket on her left)." The experimenter then did the same with 6 blue blocks, placing them in the same bucket and mixed the blocks around in the bucket. She asked the child how many red blocks and how many blue blocks were in the bucket. Then she pointed to the other container and said, "Can you help me count two blue ones into this one here?" After placing them in the bucket, she said, "How many blue ones are in here? And are there any red ones?" Next she told the child they would play a fun matching game. She showed the child two pictures, each displaying the contents of one bucket, and the child was asked to indicate by pointing which picture looked like which bucket.

In the separate distributions condition, the experimenter then brought out the two identical blue bags and said, "Look at my two bags, they look the same! I'm going to take all of these blocks here (picking up the container on her left) and pour them into this bag. There they go! Now I'm going to take this other bag over here, and I'm going to pour all of these ones (picking up the container on her right) into here." Next the experimenter told the child they were going to play a switching game and started trading the places of the bags in a circular fashion so that the child could not tell which bag was which. Then she brought the bags back up and said, "Now I'm gonna choose a bag…hmm, which bag? I know; I'll play eenie, meenie, miney, moe", and chose the bag apparently at random. In the merged distributions condition, the experimenter instead poured all the objects into one bag. The trial then continued as in the separate distributions condition, excluding any parts that made reference to separate distributions or multiple bags.

The two conditions were then identical: the experimenter took the bag and said, "I'm just going to put the bag on my toy for a second." As she placed the bag on the large toy, she 'accidentally' tipped it over, just as in Experiment 1, exclaiming, "Oh, a block fell out and made the machine go" as the toy activated. She asked the child what color they thought fell in to cause the toy to activate and why. After this, she brought out the two pictures again and asked the child to point to the picture of the distribution they thought was in the bag that was knocked over.

Trials 2 and 3 were identical to Trial 1 except the other sets of toys, blocks, bags, and pictures were used. For Trial 2, children saw that purple and orange blocks activated the fan and brown blocks were inert. The distributions were 14 purple and 6 orange blocks in one bucket and 2 orange blocks in the other bucket. For Trial 3, children saw that white and black poker chips activated the fan and yellow poker chips were inert. The distributions were 14 black and 6 white poker chips in one bucket and 2 white poker chips in the other bucket. This made purple and black the more probable objects for the merged distribution condition and orange and white the more probable objects for the separate distribution condition for Trials 2 and 3 respectively. Each experimental session took approximately 13 min. In the separate distributions condition, 10 of the 20 children were only given a single trial with the blocks; 10 children completed all three trials.[4] In the merged distribution condition all children completed all three trials. The order of counting blocks and chips into the buckets (14 red then 6 blue into a single bucket, 6 blue then 14 red into a single bucket, or 2 blue into a single bucket) was counterbalanced. The bag chosen for placement on the toy was counterbalanced. Each experimental session took approximately 13 min.

5.2. Results

Responses fell unambiguously in one of the two color categories. An assistant blind to hypotheses coded 48% of the trials with 100% agreement. There were no differences in performance based on gender or which color objects were counted first in either condition. In the separate distributions condition, there were no differences in performance between children who completed just the first trial ($N = 10$) or all three trials ($N = 10$), $z = 0$. We collapsed across these variables for the remainder of the analyses.

In the separate distributions condition, children chose the correct color (blue, orange, or white—i.e., the overall less numerous color) on 26/38 (68%) of trials. This was not different from the predicted distribution of 65% for

[4] Children completed either 1 or 3 trials because we developed the multi-toy testing method part way into data collection; we did not feel it was appropriate to discard data from the first 10 children using the identical, but single response method.

the rational sampling strategy predicted by the Sampling Hypothesis (binomial test, $p = .798$), and it is higher than chance (50%) performance (binomial test, $p = .034$) and also different from the naïve frequency matching prediction of 36% (binomial test, $p < .001$). This suggests that children were in fact able to combine the 50% probability of choosing a particular distribution with the 30% and 100% probability of obtaining the correct colored object within each of these containers (dual color vs. uniform color). In the merged distributions condition, children chose the overall more numerous object color (red, purple, or black) on 24/36 (67%) of trials. This is not different from the predicted distribution of 64% (binomial test, $p = .886$) and is marginally different from chance (50%) (binomial test, $p = .065$). It is also significantly different from children's choices in the separate distributions condition (24/36 trials vs. 12/38 trials), $t(72) = 3.18$, $p = .002$.

5.3. Discussion

The results of Experiment 4 suggest that children are using a sophisticated sampling strategy. Preschoolers in the two conditions provided different patterns of responses based on the distributional information and how data were generated. In the separate distributions condition, children integrated their prior knowledge about how the blocks were selected with their knowledge about the frequencies of different colors. In the merged distributions condition, children guessed the more numerous color at a rate equivalent to the expected distribution when summed across the entire population, as in Experiments 1 through 3.

The results from the merged distributions condition control for other possibly simpler explanations of the children's behavior in the separate distributions task. For example, one might wonder if children simply chose the more probable object because it appeared in both bags. Such arguments become less likely given the findings in the nearly identical merged distributions condition. Thus, the results from Experiment 4 suggest that children are using a more sophisticated sampling strategy than simply naïve frequency matching. Children appear to be reasoning about how a sample could be generated by integrating the distributional information overall with information about the physical separation of the population of objects into two distinct distributions.

The results from Experiment 4 not only support the Sampling Hypothesis but they also suggest that preschoolers are strikingly sophisticated in making probabilistic inferences in general. Previous research on probabilistic inference in preschoolers has rarely gone beyond asking children to make predictions about the likelihood of a sample from a single population or the likelihood of obtaining a particular object from two populations with different proportions of the target object. Indeed preschoolers are generally assumed to have difficulty with more complex probabilistic inferences. However, in a recent experiment, Girotto and Gonzalez (2008) asked children to make more complex probabilistic inferences, testing their ability to combine prior probability with subsequent information. In one of their experiments, children were shown a distribution containing, for example, four black circles, three white squares and one black square and were asked what color the experimenter was likely to pull out on a random draw. Then the experimenter drew an object blindly from the distribution and said that he could feel it was, for example, a square. School-aged, but not preschool-aged children correctly inferred that, initially a black object was more likely to be drawn, but after receiving the updating information, a white object was more likely. Our task, in the separate distributions condition, requires children to engage in a slightly different computation. We did not provide disambiguating information about which distribution was selected, thus children had to combine the 50% probability of either bag being chosen with the 0:2 and 6:14 distributions of items.

We cannot say with certainty why 4- and 5-year-olds in our experiment were able to combine probability in this sophisticated way. One possibility is that the physical separation of the distributions into two sets assisted young children in making accurate inferences in our task. Girotto and Gonzalez used a single distribution, separable only on the basis of object features or categories (e.g., shape), and this may have made their task more challenging for very young children. A second possibility is that use of a causal inference task helped children reveal earlier competence. Evidence from previous experiments suggests that adults are better at making judgments that require probability computations when causal variables are made clearer, as they are less likely to engage in base-rate neglect in these circumstances (Krynski & Tenenbaum, 2007). Additionally, evidence suggests that children perform better in probability tasks when they are encouraged to use intuitive estimation strategies, rather than reason explicitly about numbers or likelihood (Ahl, Moore, & Dixon, 1992; Bonawitz & Lombrozo, in press; Boyer, Levine, & Huttenlocher, 2008; Jeong, Levine, & Huttenlocher, 2007). Our paradigm may lead children to think less explicitly about the proportions of objects in the distributions and rely on a more intuitive probability sense by asking them to make a causal inference about an accidentally falling block, rather than asking what item was "mostly likely" to be drawn from a distribution.

6. Analysis and discussion of children's explanations

Finally, we measured and analyzed one additional aspect of children's performance in all four of the experiments that warrants discussion. In every experiment, the experimenter ended each trial by asking children: "Why do you think it was a [child's produced response] one that fell in?" All responses fell into one of five bins: (1) No explanation (e.g., "Just because"); (2) an appeal to the activation of the toys ("because it makes it spin") (3) a random response (e.g., "Because red is the color of hot lava"; (4) an appeal to the order of counting (e.g., "the red went in first"); or (5) an appeal to the distribution of the chips or how they were sampled ("Because blue was the most").

Of 220 children, on 337 trials (excluding children in the Experiment 2, 50:50 condition and children that were not asked, due to experimenter error, $N = 3$), only 19 total explanations were produced that appealed to the distribution (6%), with only 16 individual children producing a response of this nature on at least one trial (7%). The most common responses fell into Category 1 (40% of all responses), followed sequentially by 2 (34%), 3 (15%), and 4 (6%). This was the case even though the "distribution" explanations were counted very liberally, as any explanation that appealed to anything about the number of blocks in the distribution was counted (e.g., "Because there were lots of red ones" or "There were 19 blues"). No children appealed to the proportions of the objects or made any attempts at describing the random nature of the process.

Despite this inability to coherently explain why they guessed a particular color, children in all experiments guessed the more probable object color reliably more often than would be expected by chance. This finding is in agreement with other experiments examining probabilistic inference in early childhood (Denison, Konopczynski, Garcia, & Xu, 2006). Our results are also consistent with other research that demonstrates a gap between children's success on implicit measures of evidential reasoning (e.g. Schulz & Bonawitz, 2007) and their failures to explicitly demonstrate understanding of ambiguous evidence (e.g. Bindra, Clarke, & Schultz, 1980). Future experiments could explore why children struggle with explanations in these tasks where evidence is indeterminate, in contrast to findings suggesting that children as young as 3-years of age can produce sensible explanations in tasks examining non-probabilistic concepts (Bartsch & Wellman, 1989). Do children struggle specifically with probabilistic explanations? Or, do children more generally struggle to explain how they come to know particular facts from observed evidence? In any event, the explanation data from the present experiments suggests that children's probabilistic inferences are intuitive and unavailable for conscious reflection.

7. General discussion

We have suggested that rationality and variability might be reconciled by the Sampling Hypothesis. Some variability in children's responding may0020indeed be caused by random guessing or by factors such as cognitive load, methodological problems, or context effects. However, our results suggest that, at least in contexts like causal inference, a rational strategy of sampling responses from a distribution could also account for variability. The Sampling Hypothesis can be distinguished from a random guessing or noisy-max strategy by its hallmark: Response variability should be determined by the posterior distribution over hypotheses—learners should select hypotheses with probabilities that match the posterior.

Children in Experiment 1 showed probability matching in their initial guesses as well as in the distribution across three responses in the long wait condition. Children in Experiments 2 and 3 provided additional evidence of probability matching as the distribution of block colors was systematically manipulated across conditions. Finally, children in Experiment 4 demonstrated a capacity to go beyond naïve frequency matching. These children integrated the 50% probability of obtaining one of two sets of objects with the distributions contained in the two sets.

We also observed a consequence of sampling—dependency in responses decreases as the time between generating samples increases, and decreased dependency leads to a closer fit to the underlying distribution. We observed this signature dependence relationship between successive guesses in the causal inference task of Experiment 1. Specifically, children who provided three guesses in close temporal proximity showed more dependence than children who experienced a long delay between guesses, and those children's guesses were also less likely to reflect a sampling pattern. In general, children's guesses matched the distribution of the blocks in the bag more closely as the responses became more independent.

These results also suggest that children are demonstrating a level of sophistication that goes beyond what is traditionally referred to as "probability matching." In particular, the results of Experiment 4 suggest that children are not simply making guesses based on the number of blocks in the bucket; they use information about how the samples are generated to formulate a hypothesis about which block fell in the machine. These results also extend beyond traditional reinforcement learning tasks, which often show frequency matching. Children were never reinforced in our tasks; in fact, they received no feedback at all. They simply observed the evidence and then expressed a hypothesis about the contents of the machine.

Although we predicted probability matching in our experiments, one might question whether probability matching is, in fact, a sign of optimal inference. Much of the literature in economics and psychology highlights irrational cases of probability matching in decision-making. For example, consider a game in which an experimenter presents a person with multiple trials, and their task is to predict which of two options is most likely to occur to gain rewards (one has say a 67% probability of occurring). Rather than optimizing and always choosing the more probable outcome in these games, adults and school-age children often match the probabilities, thus decreasing their returns. Researchers studying this type of phenomenon often posit that probability matching arises from an incorrect belief that one can "outsmart" a game of chance (see Vulkan, 2000, for a review).

Although this is undoubtedly a poor strategy in some tasks, recent research suggests that probability matching can actually arise from more rational strategies. As we discussed earlier the rational choice changes if an agent is motivated to learn about the world, in general, rather than merely to maximize the gain of a particular choice. Choosing an option that leads to a particular outcome not only gives you the utility of that outcome, it also can provide you with information about other options and outcomes. In fact, people who are more committed to finding patterns in data are more likely to probability match; moreover, they are also more likely to discover patterns if they exist even at the cost of failing to maximize gains on particular trials (Gaissmaier & Schooler, 2008). Children are particularly likely to be motivated to discover new information rather than to achieve a particular goal, and it is possible that they might probability match on a variety of tasks for this reason. It is sensible to assume that children should be "riskier" in their hypothesis testing than adults both because they are less sure overall of how things in the world work, and because their protected immaturity means that they are more sheltered from the consequences of their decisions. If children simply maximized at all times, they might miss out on hypotheses that, although initially low in probability, actually turn out to be correct.

Additionally, sampling may be involved in rational processes for approximating Bayesian inference, and so lead to probability matching behavior, rather than being a strategy for inference or action itself. Sampling from the distribution is a rational strategy because people are typically unable to test all competing hypotheses, and so need a process to choose which hypotheses to evaluate. Sampling is involved in most of the effective machine learning algorithms that solve this problem. The nature of selecting samples from a distribution requires that when the individual samples are aggregated over many sampling events, the distribution will be returned. Thus probability matching behavior might be an epiphenomenon of a more generally useful and internal processing algorithm.

What might such an algorithm be like in detail? Though we have found evidence that suggests children are sampling from a distribution, we have not proposed *how* these samples are generated for a learner. That is, we can ask: How are children representing a distribution initially, and what are the specific algorithms they might be using to generate samples? How do those algorithms operate as new evidence is gathered? One important direction for future work is to investigate the role of evidence in children's hypothesis generation and sampling. By examining how children's pattern of responses change following newly observed evidence, we can begin to identify the specific strategies, consistent with the Sampling Hypothesis, that children may be using to initially generate and then evaluate hypotheses. For example, current work with adults suggests that the learner may use a specific algorithm (the Win-Stay, Lose-Shift strategy) that requires only occasionally resampling a hypothesis from the full posterior distribution. This algorithm may therefore be more computationally tractable than resampling after each new observation (Bonawitz, Denison, Chen, Gopnik, & Griffiths, 2011). Other more complex algorithms, such as particle filters (Levy et al., 2009; Sanborn, Griffiths, & Navarro, 2006) and Markov chain Monte Carlo (Ullman, Goodman, & Tenenbaum, 2010), can also be used to draw samples from a posterior distribution and may play a role in explaining how children are capable of making probabilistic inferences with limited computational resources.

8. Conclusions

We proposed that the Sampling Hypothesis might help to explain some competing findings on children's hypothesis testing and theory building strategies. If children are, in fact, approximating rational inference by sampling hypotheses as our results suggest, this provides an account of the variability that is often observed in patterns of responding and connects that variability to computational level accounts. More generally, the Sampling Hypothesis also suggests that while children's responses can appear irrational when examined individually, they may actually reflect a rational strategy overall.

Acknowledgments

Thanks to participating daycares and families, as well as Justine Hoch, Ariel Chang, Tiffany Tsai, Madeline Hanson, Sonia Spindt, Beth McCarthy, Jennifer Ng, Jan Iyer, and Christy Tadros for help with data collection. Thanks also to Sophie Bridgers and two anonymous reviewers for feedback on the manuscript and Paul Meinz for consultation on data analyses for Experiment 2. This research was supported in part by the James S. McDonnell Causal Learning collaborative and a Grant IIS-0845410 from the National Science Foundation.

References

Ahl, V. A., Moore, C. F., & Dixon, J. A. (1992). Development of intuitive and numerical proportional reasoning. *Cognitive Development, 7*(1), 81–108.

Bartsch, K., & Wellman, H. M. (1989). Young children's attribution of action to beliefs and desires. *Child Development, 60*(4), 946–964.

Bindra, D., Clarke, K., & Schultz, T. (1980). Understanding predictive relations of necessity and sufficiency in formally equivalent "causal" and "logical" problems. *Journal of Experimental Psychology: General, 109*(4), 422–443.

Bonawitz, E. B., Denison, S., Chen, A., Gopnik, A., & Griffiths, T. L. (2011). A simple sequential algorithm for approximating Bayesian inference. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 2463–2468). Austin, TX: Cognitive Science Society.

Bonawitz, E. B., & Lombrozo, T. (2012). Occam's rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*. http://dx.doi.org/10.1037/a0026471.

Boyer, T. W., Levine, S. C., & Huttenlocher, J. (2008). Development of proportional reasoning: Where young children go wrong. *Developmental Psychology, 33*(5), 1478–1490.

Denison, S., Konopczynski, K., Garcia, V., & Xu, F. (2006). Probabilistic reasoning in preschoolers: Random sampling and base rate. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 1216–1221). Mahwah, NJ: Erlbaum.

Denison, S., & Xu, F. (2010). Integrating physical constraints in statistical inference by 11-month-old infants. *Cognitive Science, 34*, 885–908.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review, 57*(2), 94–107.

Estes, W. K., & Suppes, P. (1959). Foundations of linear models. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory* (pp. 137–179). Stanford, CA: Stanford University Press.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition, 109*(3), 416–422.

Galton, F. (1907). Vox Populi. *Nature, 75*(1949), 450–451.

Girotto, V., & Gonzalez, M. (2008). Children's understanding of posterior probability. *Cognition, 106*(1), 325–344.

Gonzalez, A., Shafto, P., Bonawitz, E., & Gopnik, A. (2012). Is that your final answer? The effects of neutral queries on children's choices. In *Proceedings of the thirty-fourth cognitive science society*.

Goodman, N. D., Baker, C. L., Bonawitz, E. B., Mansinghka, V. K., Gopnik, A., Wellman, H., et al. (2006). Intuitive theories of mind: A rational approach to false belief. In *Proceedings of the 28th annual conference of the cognitive science society* (pp. 1382–1387).

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*(1), 108–154.

Gopnik, A. (2012). Scientific thinking in young children: Theoretical advances, empirical research and policy implications. *Science, 337*, 1623–1627.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*(1), 3–32.

Harper, D. G. C. (1982). Competitive foraging in mallards: "Ideal free" ducks. *Animal Behavior, 30*, 575–584.

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development, 1*(2), 151–195.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology, 59*(1), 30–66.

Jeong, Y., Levine, S. C., & Huttenlocher, J. (2007). The development of proportional reasoning: Effect of continuous versus discrete quantities. *Journal of Cognition and Development, 8*(2), 237–256.

Jones, M. H., & Liverant, S. (1960). Effects of age differences on choice behavior. *Child Development, 31*(4), 673–680.

Kamil, A. C., & Roitblat, H. L. (1985). The ecology of foraging behavior: Implications for animal learning and memory. *Annual Review of Psychology, 36*, 141–169.

Krynski, T., & Tenenbaum, J. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General, 136*(3), 430–450.

Kushnir, T., & Gopnik, A. (2007). Conditional probability versus spatial contiguity in causal learning: Preschoolers use new contingency evidence to overcome prior spatial assumptions. *Developmental Psychology, 43*(1), 186–196.

Kushnir, T., Wellman, H. M., & Gelman, S. A. (2008). The role of preschoolers' social understanding in evaluating the informativeness of causal interventions. *Cognition, 107*(3), 1084–1092.

Lehr, R., & Pavlik, W. B. (1970). Within-subjects procedural variations in two-choice probability learning. *Psychological Reports, 26*, 651–657.

Levy, R., Reali, F., & Griffiths, T. L. (2009). Modeling the effects of memory on human online sentence processing with particle filters. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.). *Advances in neural information processing systems* (Vol. 21, pp. 937–944). Cambridge, MA: MIT Press.

Mozer, M. C., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science, 32*(7), 1133–1147.

Myers, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to human learning and motivation* (pp. 171–205). Hillsdale, NJ: Erlbaum.

Neimark, E. D., & Shuford, E. H. (1959). Comparison of predictions and estimates in a probability learning situation. *Journal of Experimental Psychology, 57*(5), 294–298.

Piaget, J. (1983). Piaget's theory. In P. Mussen (Ed.). *Handbook of child psychology* (4th ed.) (Vol. 1). New York: Wiley.

Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer-Verlag.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117*(4), 1144–1167.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th annual conference of the cognitive science society* (pp. 726–731). Mahwah, NJ: Erlbaum.

Schulz, L. E., & Bonawitz, E. B. (2007). Serious fun: Preschoolers engage in more exploratory play when evidence is confounded. *Developmental Psychology, 43*, 1045–1050.

Schulz, L. E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared cause tummy aches? Naive theories, ambiguous evidence, and preschoolers' causal inferences. *Developmental Psychology, 43*(5), 1124–1139.

Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*(2), 162–176.

Seth, A. K. (2011). Optimal agent-based models of action selection. In A. K. Seth, T. J. Prescott, & J. J. Bryson (Eds.), *Modeling natural action selection*. Cambridge University Press.

Shi, L., Feldman, N. H., & Griffiths, T. L. (2008). Performing Bayesian inference with exemplar models. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 745–750). Austin, TX: Cognitive Science Society.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology, 36*(3), 273–310.

Sobel, D. M., Tenenbaum, J. M., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science, 28*(3), 303–333.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. New York, NY: Doubleday.

Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 19156–19159.

Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332*, 1054–1058.

Ullman, T., Goodman, N., & Tenenbaum, J. (2010). Theory acquisition as stochastic search. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 2840–2845). Austin, TX: Cognitive Science Society.

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? Optimal decisions from very few samples. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 66–72). Austin, TX: Cognitive Science Society.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*(7), 645–647.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14*(1), 101–118.

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

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*(2), 245–272.