

# **Rational Randomness: The role of sampling in an algorithmic account of preschooler's causal learning**

Bonawitz, E., Gopnik, A., Denison, S., Griffiths, T. L.

University of California, Berkeley; Berkeley, CA 94720

## **Table of Contents**

- 1 Rational Randomness
- 2 The Algorithm Problem and Marr's levels of analysis
  - 2.1 Computational Level
  - 2.2 Algorithmic level
- 3 Approximating Bayesian Inference with Monte Carlo Methods
- 4 The Sampling Hypothesis and children's inferences
  - 4.1 Preschoolers producing responses consistent with Bayesian inference
  - 4.2 Alternatives to sampling
  - 4.3 Empirical support for children's sampling
- 5 Exploring specific sampling algorithms in children's causal inferences
  - 5.1 Win-stay, lose-shift algorithms
  - 5.2 Markov-Chain Monte Carlo algorithms
- 6 Discussion
  - 6.1 Open Questions
- 7 Conclusions
- 8 References

## **Abstract**

Probabilistic models of cognitive development indicate the ideal solutions to computational problems that children face as they try to make sense of their environment. Under this approach, children's beliefs change as the result of a single process: observing new data and drawing the appropriate conclusions from those data via Bayesian inference. However, such models typically leave open the question of what cognitive mechanisms might allow the finite minds of human children to perform the complex computations required by Bayesian inference. In this chapter we highlight one potential mechanism: sampling from probability distributions. We introduce the idea of approximating Bayesian inference via Monte Carlo methods, outline the key ideas behind such methods, and review the evidence that human children have the cognitive prerequisites for using these methods. As a result, we identify a second factor that should be taken into account in explaining human cognitive development -- the nature of the mechanisms that are used in belief revision.

## 1 Rational Randomness

Over the past ten years, probabilistic approaches to cognitive development have become increasingly prevalent and powerful. These approaches can be seen as a computational extension of the “theory theory” -- the idea that children’s learning is similar to learning in science. In both cognitive development and science, learners begin with beliefs about the world that are gradually, but rationally, revised in the light of new evidence. Probabilistic models provide a way of characterizing both these beliefs—as structured models of the world—and the process of belief revision.

In this chapter we’ll describe recent work that addresses two problems with the probabilistic approach. One is what we’ll call the “algorithm problem”. Probabilistic approaches to cognitive development, like rational models in general, began with a computational level analysis. Researchers have shown that, given particular patterns of evidence, children draw rationally normative conclusions. However, this raises the question of exactly what computations or algorithms children’s minds might perform to yield those answers. This problem is particularly important because some of the most obvious possible procedures, such as enumerating each possible hypothesis and checking it against the evidence, are clearly computationally intractable.

The other problem is what we’ll call the “variability problem”. When we ask a group of children a question, typically they will produce a variety of answers. When we say that four-year-olds get the rationally “right” answer, what we really mean is that more of them produce the correct answer than we would expect by

chance. Moreover, individual children characteristically will give different answers to the same question on different occasions. They show lots of variability in their individual behavior; their explanations often appear to randomly jump from one idea to the next rather than linearly converging on the correct beliefs (Piaget, 1983; Siegler, 1996). We can witness this same kind of apparently random variability in children's play and informal experimentation. Rather than systematically acting to test one hypothesis at a time, children appear to veer at random from one kind of test to another (Chen & Klahr, 1999; Inhelder & Piaget, 1958).

This variability was one of the factors that originally led Piaget to describe young children's behavior as irrational. Indeed, such findings have led some researchers to suggest that children's behavior is always intrinsically variable and context-dependent (e.g., Greeno, 1998; Lave & Wenger, 1991; Thelen & Smith, 1994). This would seem to make children's learning very different from the kind of systematic and rational hypothesis testing we expect from science.

In this chapter we will argue that the solutions to these two problems, the algorithm problem and the variability problem, are related. Sampling from a probability distribution, rather than exhaustively enumerating possibilities, is a common strategy in algorithms for Bayesian inference used in computer science and statistics. There are many different sampling algorithms, but all of them have the feature that only a few hypotheses are, or even a single randomly selected hypothesis is, tested at a time. It can be shown that in the long run, an

algorithmic process of this kind will approximate the ideal Bayesian solution to the search problem.

First, we will argue that the idea of sampling, in general, helps make sense of children's variability. We will argue that the way children act *is* consistent with a rational account of belief revision and only seems irrational because our intuitions about what an ideal learner should look like do not take into account the complexity of the inferences that children need to make and the algorithmic procedures they use to make them. In particular, systematic variability is a hallmark of sampling processes. By thinking about how children might use effective algorithmic strategies for making such inferences, we come to see these apparently irrational behaviors in a different light. We are starting to show empirically that children's variability is, in fact, systematic in just the way we would predict if they were using a sampling-based algorithm.

Second, we will describe two particular psychologically plausible sampling algorithms that can approximate ideal Bayesian inference—the Win-Stay, Lose-Shift procedure and a variant of the Markov Chain Monte Carlo algorithm. We will show empirically that, in different contexts, children may use something like these algorithms to make inferences about the causal structure of the world.

## **2 The Algorithm Problem and Marr's levels of analysis**

Marr (1982) identified three distinct levels at which an information-processing system can be analyzed: the *computational*, *algorithmic*, and *implementational* levels. We will focus on the computational and algorithmic

levels here. (The implementational level, which answers the question of how the system is physically realized—e.g. what neural structures and activities implement the learning processes described at the algorithmic level—warrants focus in future work.) We will first give a brief overview of the computational and algorithmic levels and then delve more deeply into the specific algorithms young learners may be using.

### *2.1 Computational Level*

Marr's computational level focuses on the computational problems that learners face and the ideal solutions to those problems. For example, Bayesian inference provides a computational-level account of the inferences people make when solving inductive problems, focusing on the form of the computational problem and its ideal solution. Bayesian models are useful because they provide a formal account of how a learner should combine prior beliefs and new evidence to change her beliefs.

In Bayesian inference, a learner considers how to update her beliefs (or hypotheses,  $h$ ) given some observed evidence (or data,  $d$ ). Assume that the learner has different degrees of belief in the truth of these hypotheses before observing the evidence, and that these degrees of belief are reflected in a probability distribution  $p(h)$ , known as the *prior*. Then, the degrees of belief the learner should assign to each hypothesis after observing data  $d$  are given by the *posterior* probability distribution  $p(h|d)$  obtained by applying Bayes' rule

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in \mathcal{H}} p(d|h')p(h')} \quad (1)$$

where  $p(d|h)$  indicates the probability of observing  $d$  if  $h$  were true, and is known as the *likelihood*.

An important feature of Bayesian inference is that it doesn't just yield a single deterministically correct hypothesis given the evidence. Instead, Bayesian inference provides an assessment of the probability of all the possible hypotheses. The "prior" distribution initially tells you the probability of all the possible hypotheses. Each possible hypothesis can then be assessed against the evidence using Bayes rule. This produces a new distribution of less likely and more likely hypotheses. Bayesian inference proceeds by adjusting the probabilities of all the hypotheses, the distribution, in the face of new data. It transforms the "prior" distribution you started with—your degrees of belief in all the possible hypotheses—into a new "posterior" distribution. So, in principle, Bayesian inference not only determines which hypothesis you think is most likely, it also changes your assessment of all the other less likely hypotheses.

The idea that inductive inference can be captured by Bayes' rule has been applied to a number of different aspects of cognition, demonstrating that people's inferences are consistent with Bayesian inference in a wide range of settings (e.g. Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Griffiths & Tenenbaum, 2009; Kording & Wolpert, 2004; Weiss, Simoncelli, & Adelson, 2002; Xu & Tenenbaum, 2007). People do seem to update their beliefs given evidence in the way that specific Bayesian models predict.

Bayesian models have proved especially helpful in understanding how children might develop intuitive theories of the world. We can think of intuitive theories as hypotheses about the structure of the world, particularly its causal structure. Causal graphical models (Pearl, 2000; Spirtes, Glymour, & Schienens, 1993) and more recently hierarchical Bayesian models (Tenenbaum, Griffiths, & Kemp, 2006) provide particularly perspicuous representations of such hypotheses. In particular, by making explicit and systematically relating the structure of hypotheses to probabilistic patterns of evidence, Bayesian causal models can establish the probability of particular patterns of data given particular hypotheses. This means that Bayesian inference can then be used to combine prior beliefs and the likelihood of newly observed evidence given various hypotheses, to update the probability of hypotheses – making some beliefs more and others less likely.

In fact, there is now extensive evidence that this computational approach provides a good explanatory account of how children infer hypotheses about causal structure from evidence. We can manipulate the evidence children see about a causal system, as well as their beliefs about the prior probability of various hypotheses about that structure, and see how this influences their inferences about that system. Quite typically children choose the hypotheses with the greatest posterior probability in Bayesian terms (Bonawitz et al, 2012; Bonawitz, Fischer, & Schulz, 2012; Goodman et al, 2008; Gopnik et al. 2001, 2004, Kushnir & Gopnik 2005, 2007, Schulz, Gopnik & Glymour 2007, Lucas, Gopnik, & Griffiths, 2010; Schulz, Bonawitz, & Griffiths, 2007; Sobel et al. 2004).



However, the finding that the average of children's responses looks like the posterior distributions predicted by these rational models does not necessarily imply that learners are actually carrying out the calculation instantiated in Bayes' rule at the *algorithmic* level. Indeed, given the computational complexity of exact Bayesian inference, this would be impossible. So it becomes interesting to ask *how* learners might be behaving in a way that is consistent with Bayesian inference.

## 2.2 Algorithmic level

Marr's algorithmic level asks how an information-processing system does what it does; for example, what cognitive processes do children use to propose, evaluate, and revise beliefs? The computational level has provided an important perspective on children's behavior, affording interesting and testable qualitative and quantitative predictions that have been borne out empirically. But it is just the starting point for exploring learning in early childhood. Indeed, considering other levels of analysis can help to address significant challenges for Bayesian models of cognitive development.

In particular most computational-level accounts do not address the problem of *search*. For most problems, the learner can't actually consider every possible hypothesis, as it would be extremely time-consuming to enumerate and test every hypothesis in succession. Researchers in AI and statistics have raised this concern, showing that given complex problems and the time constraints of real world inference, full Bayesian inference quickly becomes computationally

intractable (e.g., Russell & Norvig, 2003). Thus, rational models raise questions about how a learner might search through a (potentially infinite) space of hypotheses: If the learner simply maximized, picking out only the most likely hypothesis to test, she might miss out on hypotheses that are initially less likely but actually provide a better fit to the data. This problem might appear to be particularly challenging for young children who, in at least some respects, have more restricted memory and information-processing capacities than adults (German & Nicholas, 2003; Gerstadt, Hong, & Diamond, 1994).

Applications of Bayesian inference in computer science and statistics often try to solve the computational problem of enumeration and evaluation of the hypothesis space by sampling a few hypotheses rather than exhaustively considering all possibilities. These approximate probabilistic calculations use what are called “Monte Carlo” methods. A system that uses this sort of sampling will be variable—it will entertain different hypotheses apparently at random. But this variability will be *systematically related* to the probability of the hypotheses—more probable hypotheses will be sampled more frequently than less probable ones. The success of Monte Carlo algorithms for approximating Bayesian inference in computer science and statistics suggests an exciting hypothesis for cognitive development. The algorithms children use to perform inductive inference might be similarly based on sampling from the appropriate probability distributions. We explore this *Sampling Hypothesis* in detail in the remainder of the chapter.

The Sampling Hypothesis provides a way to reconcile rational reasoning with variable responding, and it has the potential to address both the algorithmic and search problem. It also establishes an empirical research program, in which we look for the signatures of sampling in general, and of specific sampling algorithms in particular, in children's behavior

### **3 Approximating Bayesian Inference with Monte Carlo Methods**

Monte Carlo algorithms include a large class of methods that share the same general pattern. They first define the distribution that samples will come from, then randomly generate the samples, and finally aggregate the results. The goal is usually to approximate an expectation of a function over a probability distribution (e.g., the mean of the distribution, or the probability that a sample from the distribution has a particular property); that is, an approximation of the distribution is given by summing over all the different individual samples generated during the MCMC process.

The simplest Monte Carlo methods directly generate samples from the probability distribution in question. For example, if you wanted to know the mean of the distribution that assigns equal probability to the numbers one through six, you could calculate the exact mean by averaging over each probability for each value (one through six). Monte Carlo methods provide an alternative to numerically computing the mean: instead, you could imagine rolling a fair die to generate samples from this distribution, tracking the results of each roll, and then averaging the results together. This process would let you uncover an important

fact about the distribution without having to numerically calculate the probability of each possible outcome individually. Although in this example, it would be relatively trivial to numerically compute the exact mean, Monte Carlo sampling provides an alternative approach that can be used when the distributions become harder to evaluate, such as considering the product of multiple dice rolls. When the probability distributions we want to sample from get even more complex, more sophisticated methods need to be used to generate samples. In particular, it can quickly become computationally intractable to take samples directly from the posterior distribution itself, so various Monte Carlo algorithms have been developed to best approximate these different kinds of complexity.

Monte Carlo algorithms for approximating Bayesian inference are thus methods for obtaining the equivalent of samples from the posterior distribution without computing the posterior distribution itself. One class of methods, based on a principle known as *importance sampling*, generates hypotheses from a distribution other than the posterior distribution, and then assigns weights to those samples (akin to increasing or decreasing their frequency) in order to correct for the bias produced by using a different distribution to generate hypotheses (see Neal, 1993, for details).

The strategy of sampling from other known distributions and then updating the sample to correct for bias can also be used to develop algorithms for probabilistically updating beliefs over time. For example, in a particle filter (see Doucet, de Freitas, & Gordon, 2001, for details), hypotheses are generated based on a learner's current beliefs, and then reweighted to reflect the evidence

provided by new observations. This provides a way to approximate Bayesian inference that unfolds gradually over time, with only a relatively small number of hypotheses being considered at any one instant.

Another class of Monte Carlo methods make use of the properties of Markov chains. These Markov Chain Monte Carlo algorithms, such as the Metropolis Hastings algorithm, explore a posterior probability distribution in a way that requires only a single hypothesis to be considered at a time (see Gilks, Richardson, & Spiegelhalter, 1996, for details). In these algorithms, a learner generates a hypothesis by sampling a variant on his or her current hypothesis from a “proposal” distribution. The proposed variant is compared to the current hypothesis, and the learner stochastically selects one of the two hypotheses. This process is then repeated, and the learner gradually explores the space of hypotheses in such a way that each hypothesis will be considered for an amount of time that is proportional to the posterior probability of that hypothesis.

Overall, Monte Carlo methods have met with much success in exploring posterior distributions that are otherwise too computationally demanding to evaluate (see Robert and Casella, 2004, for a review). Recent work by Griffiths and colleagues has explored how Monte Carlo methods can be used to develop psychological models that incorporate the cognitive-computational limitations that adult learners face. Some empirically generated psychological process models turn out to correspond to the application of Monte Carlo methods. For example, Shi, Feldman, and Griffiths (2008) showed that importance sampling corresponds to exemplar models, a traditional process-level model that has been applied in a

variety of domains. Sanborn, Griffiths, and Navarro (2010) used particle filters to approximate rational statistical inferences for categorization. Bonawitz and Griffiths (2010) show that importance sampling can be used as a framework for analyzing the contributions of generating and then evaluating hypotheses.

Other research supports the idea that adults may be approximating rational solutions through a process of sampling. For example, adults, like children, often generate multiple answers to a question. If you ask adults how many beans are in a large jar they will provide a range of responses. A classic result shows that the averaged result of many such responses converges on the right answer, even though any individual guess may be very far from the correct mean, (“The Wisdom of Crowds”; Galton, 1907; Surowiecki, 2004). The same effect holds even when a single person makes multiple guesses, although if there are dependencies between an individual’s subsequent guesses, the average of the responses will not produce a correct approximation. This is because Monte Carlo methods require that responses be independently distributed. However, the overall “Wisdom of the Crowds” phenomena support the idea that (adult) individuals are not simply providing their best guess but rather are sampling from a subset of hypotheses when making inferences (Vul & Pashler, 2008). Related work suggests that people often base their decisions on just a few samples (Goodman, Tenenbaum, Feldman, & Griffiths, 2008; Mozer, Pashler, & Homaei, 2008) and in many cases an optimal solution is to take only one sample (Vul, Goodman, Griffiths, & Tenenbaum, 2009). These results suggest that adults may be approximating probabilistic inference through psychological processes that

are *equivalent* to sampling from the posterior. That is, the learner need not compute the full posterior distribution in order to sample responses that still lead to an approximation of the posterior. Here, we use “sampling from the posterior” to entail any sampling processes that produce equivalent samples, without necessarily requiring the learner to compute the full posterior.

These processes that approximate the full posterior are consistent with what we have termed the Sampling Hypothesis. The signature of sampling-based inferences is the fact that apparently random guesses actually reflect the probability of the hypotheses they embody. Each person may produce a different hypothesis about the outcome of two dice rolls on a different occasion, but hypotheses that are closer to correct – that is those that have a higher probability in the posterior distribution -- will be more likely to be produced than those that are less likely. If two fair dice are rolled, the most likely outcome is 7, however people generate a range of guesses with varying probability. Guesses of 6 or 8, will be five times more likely to be true than guesses of 2 or 12. The Sampling Hypothesis predicts that human beings are also five times more likely to produce those guesses; indeed, it predicts that the probability that an individual will guess any particular outcome will match the probability of generating that outcome under the true distribution.

#### **4 The Sampling Hypothesis and children’s inferences**

Might children’s inferences be consistent with the Sampling Hypothesis? The first step in exploring this claim is to see whether children produce

responses that are consistent with Bayesian inference in general. The second step is to see whether children produce behaviors that are consistent with sampling in particular. In order to demonstrate that children's responding is consistent with Bayesian inference, we must demonstrate that children are sensitive both to their prior beliefs and to the evidence they observe. As we described above, many studies show that children choose the probabilistically most likely hypothesis, but to truly test the prediction that children are sensitive to *posterior* distributions, then children's responses should change when both their prior beliefs and the probability of the evidence are independently manipulated. We highlight a few studies that suggest children produce responses consistent with Bayesian inference. We then briefly discuss alternatives to the sampling hypothesis. Finally, we turn to more detailed empirical evidence supporting the claim that children sample responses.

#### *4.1 Preschoolers producing responses consistent with Bayesian inference*

Before we are able to identify whether children sample responses in a way that approximates a posterior distribution, we must first demonstrate that children's responses are consistent with those distributions. Schulz, Bonawitz, and Griffiths (2007) presented preschoolers with stories pitting their existing theories against statistical evidence. Each child heard two stories in which two candidate causes co-occurred with an effect. Evidence was presented in the form:  $AB \rightarrow E$ ,  $AC \rightarrow E$ ,  $AD \rightarrow E$ , etc. In one story, all variables came from the same domain; in the other, the recurring candidate cause, A, came from a different



domain (A was a psychological cause of a biological effect). After receiving this statistical evidence, children were asked to identify the cause of the effect on a new trial. Consistent with the predictions of the Bayesian framework, both prior beliefs and evidence played a role in children's causal predictions. Four-year-olds were more likely to identify 'A' as the cause after observing the evidence than at Baseline. Results also revealed a role of theories in guiding children's predictions. All children were more likely to identify A as the cause within domains than across domains.

A particularly interesting empirical feature of this study is that because Schulz et al had a measure of children's prior beliefs at baseline, they could demonstrate that proportionally the children's responses (after having observed the evidence) were consistent with posterior distributions predicted by Bayesian models. That is, after observing the evidence, some children endorsed hypothesis "A" and others endorsed the other hypothesis. The proportion of children who favored "A" was the probability of "A" being the actual cause, given the prior beliefs of the children and the evidence observed. For example, when the posterior predicted 80% probability for hypothesis "A", then results revealed about 80% of the children choosing "A".

Other studies reveal children producing graded responses to evidence reflecting Bayesian posteriors (Bonawitz & Lombrozo, 2012; Kushnir & Gopnik 2005, 2007, Sobel et al 2004). For example, Bonawitz and Lombrozo (in press) investigated whether young children prefer explanations that are simple, where simplicity is quantified as the number of causes invoked in an explanation, and

how this preference is reconciled with probability information. Preschool-aged children were asked to explain an event that could be generated by one or two causes, where the probabilities of the causes varied across several conditions. Children preferred explanations involving one cause over two, but were also sensitive to the probability of competing explanations. That is, as evidence gradually increased favoring the complex explanation, the proportion of children favoring the complex explanation also increased. These data provide support for a more nuanced sensitivity to evidence. When evaluating competing causal explanations preschoolers are able to integrate evidence with their prior beliefs (in this case employing a principle of parsimony like Occam's razor as an inductive constraint). However, because children's prior beliefs were not independently evaluated, it is difficult to say whether the proportion of responses generated by children in these studies matched a true posterior distribution predicted by a Bayesian model.

#### *4.2 Alternatives to sampling*

The studies above suggest that children are sensitive to evidence in ways predicted by Bayesian inference; at least, a proportion of children select the hypothesis that is best supported by the evidence and their prior beliefs. There is also variability in children's responses –the proportion of times that they select a hypotheses increases as the hypothesis receives more support, but they will still sometimes produce an alternative hypothesis. But does that mean that children are sampling their responses from a posterior distribution? It might instead be

that the variability in the children's responses is simply the effect of noise—in fact this is the most common, if generally implicit, assumption in most developmental studies. This noise could be the result of cognitive load, context effects, or methodological flaws that lead children to stochastically produce errors. For example in the Schulz et al (2007) study, all the children might indeed be choosing the most likely hypothesis on each trial, always preferring the within-domain or cross-domain answer. But they might fail to express that hypothesis correctly because of memory or information-processing or communication problems. We call this alternative the *Noisy Maximizing* alternative to the sampling hypothesis.

Another alternative is that children's behavior does reflect the probability of hypotheses but does so through the result of a simpler process than hypothesis sampling. Consider a similar though simpler phenomenon that can be found in a much older literature. Children, like adults and even non-human animals, frequently produce a pattern called probability matching in reinforcement learning (e.g. Jones & Liverant, 1960). If children are rewarded 80% of the time for response "A" but 20% of the time for response "B", they are likely to produce "A" 80% of the time and "B" 20% of the time. If these responses reflect an implicit hypothesis about the causal power of the action ("A" will cause the reward), this probability matching looks a lot like the behavior in the more explicitly cognitive causal learning tasks. It might, however, simply be the result of a strategy we will call "Naïve Frequency Matching". Children using this

strategy would simply match the frequency of their responses to the frequency of the rewards.

The idea that the variability of children's responses could result from sampling is related to probability matching in that it predicts that the learner's responses on aggregate will match the posterior. Thus, probability matching following reinforcement is consistent with the Sampling Hypothesis, but it is also consistent with the Naïve Frequency Matching account. These two accounts differ in that sampling 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, sampling predicts that children's responding matches the posterior probabilities of different hypotheses.

Moreover, there is still something puzzling about probability matching from the point of view of simple reinforcement theory. Why would children use a naïve frequency matching strategy? Why don't all the children select the more probable hypothesis? This is the strategy, after all, that is most likely to yield a right answer and so enable the children to be rewarded. Why are 20% of children choosing the *less* likely hypothesis? If children are *sampling* responses from the posterior distribution, this could explain this result. On this view, the variability in children's responses may actually itself be rational, at least sometimes. In particular, it may reflect a strategy children use to select which hypotheses could explain the data they have observed.

There is little work exploring probability matching in children beyond simple reinforcement learning. What there is suggests that children do not, in fact,

probability match when they are considering more cognitive hypotheses—particularly linguistic hypotheses (Hudson Kam & Newport, 2005; 2009). In what follows, we present a set of studies that present the first test of the Sampling Hypothesis with children, distinguishing sampling from both the Naïve Frequency Matching and Noisy Maximizing alternatives.

#### *4.3 Empirical support for children's sampling*

In a first set of experiments, we explored the degree to which children match posterior probabilities in a causal inference task, set up as a game about a toy that activates when particular colored chips are placed inside a bin (Denison, Bonawitz, Gopnik, & Griffiths, 2010). The toy allowed us to precisely determine the probability of different hypotheses about which chip had fallen into the bin. Earlier studies have showed that even young infants are sensitive to probability in these contexts – 6-month-olds assume that the more frequent color chip will be more likely to be selected from the bag (Denison, Reed, & Xu, in press). In the first experiment, we tested two key predictions of the sampling hypothesis: probability matching and an effect of the dependency between responses. These are both known consequences of sampling behavior (Vul & Pashler, 2008).

Children were introduced to a toy – a large box with an activator bin and an attached smaller toy that could light up and play music. The experimenter demonstrated that placing red chips or blue chips into the activator bin caused the toy to activate. Then a distribution of 20 red and 5 blue chips were placed into a transparent container and transferred into a rigid, opaque bag. The

experimenter placed the bag on top of the toy and “accidentally” knocked it over away from the child and towards the activator bin, and the toy activated. The child was asked what color chip they thought fell into the bin to activate the toy. (See Figure 1.) Children in the *short wait* condition completed two additional trials immediately following this first trial, and children in the *Long Wait* condition completed two additional trials each one-week apart (all trials consisted of the same 80:20 distribution).

We manipulated time between guesses because (following results with adults from Vul and Pashler, 2008), we suspected that there would be greater dependence among guesses if they were spaced close together. As described previously, one of the requirements of producing a “good” approximation to the posterior is independence between samples (although, there are a few special cases in which some dependency between responses can still yield accurate approximations—a point we’ll return to later). In general, however, the sampling hypothesis predicts different patterns of response when there is more dependence between hypotheses. Thus, we predicted that the *long-wait* condition should have produced more independence between the hypotheses (e.g. the children may not have remembered what they had just said) and so produce a better approximation to the posterior.

Results indicated that, collapsed across conditions, children’s guesses on Trial 1 were in line with the signature of sampling: probability matching. Children guessed the red chip (i.e. the more probable chip) on 70% of trials and the blue chip on 30% of trials, not significantly different from the predicted distribution of

80% and 20%, respectively, but significantly different from chance (50%). Children's responses were also in agreement with the dependency results found with adults; children in both conditions showed some dependencies between guesses, but the dependencies were greater in the *short wait* than in the *long wait* condition. As the sampling hypothesis would predict, because there was less dependence, there was also a better fit to the actual probabilities in the *long wait* condition.

Although the results of this first experiment suggest that children's responses reflected probability matching, they are also congruent with the Noisy Maximizing prediction. That is children may have attempted to provide maximally accurate "best guesses" but simply failed to do so at ceiling levels due to factors such as task demands or cognitive load. In a second experiment, we tested the probability matching prediction more directly by systematically manipulating the distributions of chips children saw across three conditions. In a 95:5 condition children counted 19 red and 1 blue chips, in a 75:25 condition they counted 15 red and 5 blue chips, and in a 50:50 condition they counted 10 red and 10 blue chips. As predicted, children's responses reflected probability matching. Children's tendency to guess the red chip increased linearly as the proportion of red to blue chips increased from 50:50 to 75:25 to 95:5 (see Figure 2). This result is congruent with probability matching but not noisy maximizing, as noisy maximizing would have resulted in similar performance between the 75:25 and 95:5 conditions.

In a third 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 in two conditions were given distributions that included three different colors of chips: red, blue and green. The procedure unfolded as it did in the first two experiments. As in the second experiment, the distributions were systematically manipulated across conditions. Children in the *82:9:9 Condition* saw distributions of 18 red, 2 blue and 2 green chips, and children in the *64:18:18 Condition* saw 14 red, 4 blue and 4 green chips. In this experiment, children in both conditions guessed that the red chip had activated the machine more often than would be expected by chance but again they also did not choose the red chip at ceiling levels. Children's responses reflected probability matching in that children in the *82:9:9 Condition* guessed the red chip 72% of the time, significantly more often than children in the *64:18:18 Condition*, who guessed the red chip 53% of the time. The proportion of children choosing the red chip in the *82:9:9 Condition* was not different from the predicted distribution of 82% and the proportion of children who did so in the *64:18:18 Condition* was not different from the predicted distribution of 64%.

Children in the second and third experiments produced guesses that are consistent with the probability matching prediction of the sampling hypothesis. However, as we mentioned previously, children in a variety of reinforcement learning paradigms have demonstrated probability matching to the frequencies of



reinforced responses. The current studies did not involve any reinforcement, and children responded to the number of chips in the container rather than to frequency of effects, so they could not simply be explained in terms of reinforcement learning. However, these results are still consistent with a variation of the Naïve Frequency Matching account. Although responses were not reinforced, children in these experiments may have matched their responses to the overall frequency of the chips – they said “red” more often simply because they saw more red chips. We conducted a fourth experiment to test the prediction that children’s responses will match the *posterior distribution* of hypotheses and not simply match the frequencies of the different colored chips encountered in the experimental session.

The frequencies of chips can be separated from the probability of selecting each color by introducing a constraint on the generative process. In a variant of the procedure used in the first three experiments, children counted two separate distributions of chips with the experimenter: one distribution of 14 red and 6 blue chips and a second distribution of 0 red and 2 blue chips. The experimenter placed the separate distributions into two identical bags, mixed the placement of the bags around out of the child’s view, and then randomly chose one of the bags to place on the machine and knock over. In this case, if children are solely concerned with the frequencies of each color of chip, they should guess a red chip on 64% of the trials and a blue chip on 36% of the trials. If they are instead producing guesses based on the probability of either color chip falling from the randomly chosen bag, they should guess the red chip 35% of the time

and the blue chip 65% of the time [ $P(\text{blue chip}) = (1/2 \times 6/20 + 1/2 \times 2/2)$ ]. Children guessed the red chip on 32% of the trials, different from chance (50%) and the frequency matching prediction (64%) but not different from the posterior probability matching prediction (35%).

In sum, results from these four experiments suggest that children's responses in a simple causal inference task were in agreement with the sampling hypothesis. First, children showed dependencies between guesses on three consecutive trials and this dependency decreased as a function of time between guesses, and more independence led to greater probability matching. Second, children provided responses that, on aggregate, reflected the posterior distribution of hypotheses when making guesses involving either two possible hypotheses or three possible hypotheses, ruling out the possibility that children were noisily maximizing. Finally, children's guesses matched the posterior distribution of hypotheses rather than the simple frequencies of observed colors of chips. They rationally integrated the probability of randomly selecting one of the two distributions with the frequency of the chips within the distributions.

## **5 Exploring specific sampling algorithms in children's causal inferences**

The experiments discussed in the previous section provide preliminary support for the Sampling Hypothesis, suggesting that children are doing something that looks like sampling as opposed to noisily maximizing, and that children are going beyond making simple frequency tabulations in causal learning tasks. While these results suggest that learners sample responses from posterior distributions,

these studies were not designed to explore specific algorithms a learner might be using to select a hypothesis as she encounters new data, and they do not propose a specific mechanism for search through a hypothesis space.

There are myriad ways in which a learner could move through the space of hypotheses consistent with sampling algorithms. Learners may resample a best guess from the full posterior every time a new piece of data is observed. Learners may sample a hypothesis and stick with it until there is impetus to re-evaluate (e.g. maybe data reaches some threshold of “unlikeliness” to have been generated by the current hypothesis). The way in which a learner chooses to re-evaluate hypotheses may also differ: she may make subtle changes to the hypothesis she’s currently entertaining; she may go back and resample completely from the full posterior distribution; or she may choose a best guess from some surrogate distribution (an approximation to the posterior distribution). Learners could sample and simultaneously consider a few hypotheses or just one. These ideas about specific sampling and search algorithms have analogs in computer science and machine learning. We present two different studies designed to test whether children might be using variants of two types of search algorithms – a Win-Stay, Lose-Shift algorithm (WSLS) and a Markov Chain Monte Carlo (MCMC) algorithm.

In order to explore the WSLS and MCMC algorithms, we presented children with more complex causal learning tasks that unfolded over time. Children received new evidence at several stages, and at each stage we asked them to provide a new guess about what was going on. The pattern of responses

that children produced and particularly the dependencies among those responses, helped allow us to discriminate which specific algorithms they employed.

### *5.1 Win-stay, lose-shift algorithms*

To test the idea that children use the Win-Stay, Lose-Shift algorithm, we designed both deterministic and probabilistic causal tasks. In the deterministic task, data necessarily “rule out” a set of possible hypotheses; in the probabilistic task, the data are consistent with all hypotheses, but statistically may favor certain hypotheses over others. The task proceeds as follows: We let children take an initial guess, before seeing any evidence; we then show children some evidence and ask them about their hypotheses after the evidence; we then show children more evidence and ask them again about their hypotheses, and so forth and so on. Thus, children observed a sequence of data, and we could use the responses of an individual child as he or she moved through the hypothesis space following each piece of evidence, to help us tease apart different specific algorithms.

In particular, we were interested in algorithms based on the Win-Stay, Lose-Shift (WSLS) principle. These algorithms entertain a single hypothesis at a time, staying with that hypothesis as long as it adequately explains the observed data and shifting to a new hypothesis when that is no longer the case. The WSLS principle has a long history in computer science, where it is used in reinforcement learning and game theory (Robbins, 1952; Nowak & Sigmund, 1993), and

psychology, where it has served as an account of human concept learning (Restle, 1962). Bonawitz, Denison, Chen, Gopnik, & Griffiths (2011) provided a mathematical proof that demonstrates how specific algorithms using the WSLS principle can be used to sample from posterior distributions. The result is a set of surprisingly simple sequential algorithms for performing Bayesian inference.

The deterministic case of Win-Stay, Lose-Shift means that data necessarily rule out a set of hypotheses. The algorithm simply involves a process where the learner stays with a hypothesis when data are consistent, and shifts to a new hypothesis when data are inconsistent. The probabilistic case presents a more interesting, and ecologically plausible test of WSLS, so we focus on the probabilistic studies here.

In the probabilistic causal task, we introduced children to a machine that could be activated with different kinds of blocks. An experimenter demonstrated that three kinds of blocks activate the machine with different probabilities: red blocks activated the machine on five out of six trials, the green blocks on three out of six trials, and the blue blocks just once out of six trials. We then introduced children to a new block that had “lost its color” and told children we needed their help guessing what color the block should be: red, blue, or green. We then asked children what happened each time the mystery block was placed on the machine (either the machine activated or did not); after each observation we asked children what color they thought the block was now.

One specific WSLS algorithm proceeds on the problem of inferring the identity of the mystery block given probabilistic data as follows. The learner starts

out by sampling a hypothesis from the prior distribution, before seeing any data about the mystery block. Let's say that she happens to choose red by sampling it randomly from her prior; all that means is that she rolls a weighted die such that the weights of the colors on the die are proportional to her beliefs about how likely the block is to be each color, before seeing the evidence. In this case, for example, the prior evidence provides an equal probability for each block initially, so she might be equally likely to guess red, blue or green. In another case she might have reason to think that red blocks were more common, so that she would weight the internal throw of the die more heavily towards red, though blue or green might also turn up. Let's say this learner happens to roll red. Then the mystery block is set on the machine and it turns out that it activates the toy. The learner now must decide whether to stay with red or shift to another hypothesis. In this simple WSL algorithm, the decision to stay or shift is made based purely on the likelihood for the observed data. As seen in the demonstration phase, the red block activates the machine five out of six times and so the likelihood of seeing the toy light if the block really is red is simply  $5/6$ . So to make the choice to stay, we can imagine a coin with  $5/6$  probability of landing on stay and  $1/6$  probability landing shift. That is, although the evidence is consistent with the red block hypothesis, there is still a (small) chance that the coin will come up shift, and the learner will return to the updated posterior (which includes all the evidence observed so far) to sample their next guess. Each time the learner observes a new piece of data, she makes the choice whether to stay or shift, in this way, based only on the most recent data she has observed.

When applying this WSL algorithm, an individual learner may look like she is randomly veering from one hypothesis to the next, sometimes abandoning a likely hypothesis or sampling an unlikely one, sometimes being too strongly influenced by a piece of data and sometimes ignoring data that is unlikely under her current hypothesis. However, looking across a population of learners reveals a surprising property of this algorithm: Aggregating all the hypotheses selected by all of the learners returns the Bayesian predicted posterior distribution (or at least a sample-based approximation thereof). Thus the WSL algorithm provides a more efficient way to do Bayesian inference. The learner can maintain just a single hypothesis in her working memory and need only re-compute and resample from the posterior on occasion. Nevertheless, the responding of participants on aggregate still acts like a sample from the posterior distribution.

We can contrast this WSL algorithm with *independent* sampling, in which a learner simply samples a new hypothesis from the posterior distribution each time a response is required. In other words, on each trial the learner will choose red, blue, or green in proportion to the probability that the block is that particular color given the accumulated evidence. The WSL algorithm shares with independent sampling the property that responses on aggregate will match the posterior probability, but the algorithms differ in terms of the dependencies between responses. Because the learner resamples a hypothesis after each new observation of data in the independent sampling algorithm, there is no dependency between an individual's successive guesses. In contrast, the WSL algorithm predicts dependencies between responses: if the data are consistent

with the current hypothesis, then the learner is likely to retain that hypothesis. This specific instantiation of WSLS is thus one of the special cases where the algorithm approximates the correct distribution even though there are dependencies between subsequent guesses. This establishes some clear empirical predictions: Both algorithms will produce a pattern of responses consistent with Bayesian inference on any given trial, but they differ in the predictions that they make about the relationship between responses on successive trials.

The first thing we can examine is simply whether children's responses approximate the posterior distribution produced by Bayesian inference in aggregate; indeed children's predictions on aggregate correlated highly with Bayesian posteriors (Figure 3). Next, we can examine the dependencies between responses for the individual learners to investigate whether WSLS or independent sampling provide a better fit to children's responses. To compare children's responses to the WSLS and independent sampling algorithms, we first calculated the "shift probabilities" under each model. Calculating shift probabilities for independent sampling is relatively easy: because each sample is independently drawn from the posterior, the shift probability is simply calculated from the posterior probability of each hypothesis after observing each piece of evidence. Shift probabilities for WSLS were calculated such that resampling is based only on the likelihood associated with the current observation, given the current hypothesis. That is, with probability equal to this likelihood, the learner resamples from the full posterior. We also computed the log-likelihood scores for



both models—the probability that we would observe the pattern of responses from the children given each model. Children’s responses highly correlated with and had higher log-likelihood scores from the Win-Stay, Lose-Shift algorithm. This suggests that the pattern of dependencies between children's responses are better captured by the win-stay, lose-shift algorithm than by an algorithm such as independent sampling.

### *5.2 Markov-Chain Monte Carlo algorithms*

The results of the WSLS experiments suggest one algorithm that learners might use to sample and evaluate hypotheses. In the experiments we’ve considered so far, the space of possible hypotheses was relatively limited. Children only had to consider whether a red blue or green chip or block activated the machine. However, the question of how a learner *searches* through a space of hypotheses remains an important issue for cases when the space of hypotheses is much larger. Constructing an intuitive theory based on observing the world often confronts learners with a more complex space of possibilities.

In particular, in the examples we discussed so far, the causal categories the children saw (red, blue and green blocks) and the causal laws (chips activate the machine) were both well-defined -- they didn’t have to be learned. In other cases children have been shown to use probabilistic inference to uncover even relatively complex and abstract causal laws (e.g. the difference between a causal chain and a common cause structure, Schulz, Gopnik, & Glymour, 2007, or between a disjunctive or conjunctive causal principle, Lucas et al., 2010). Schulz

et al. (2008) and Seiver, Gopnik, and Goodman (in press) also showed that children can uncover new causal categories and concepts. In more realistic cases of theory change, learners, however, might face the “chicken-and-egg” problem: the laws can only be expressed in terms of the theory’s core concepts, but these concepts are only meaningful in terms of the role they play in the theory’s laws. How is a learner to discover the appropriate concepts and laws simultaneously, knowing neither to begin with? How could a learner search through this potentially infinite space of possibilities?

Recent, ongoing work by Bonawitz, Ullman, Gopnik, and Tenenbaum, has the goal of studying empirically how children's beliefs evolve through such a process of theory discovery, and understanding computationally how learners can converge quickly on a novel but veridical system of concepts and causal laws. Goodman et al. (2008) and Ullman et al. (2010) describe a sampling method using a grammar-based Metropolis-Hastings MCMC algorithm. The grammar generates the prior probabilities for the theories and the MCMC algorithm can be used to evaluate these theories given evidence. Specifically, the grammar is a broad language for defining theories, which is able to build a potentially infinite space of possibilities (see also Ullman, Goodman, & Tenenbaum 2010). This grammar produces the space of possible hypotheses, and even provides a measure of the probability of each hypothesis: this prior probability of each hypothesis is the probability that each hypothesis is generated by the grammar. The algorithm begins with a specific theory,  $t$ , and then uses the grammar to propose random changes to the currently held theory. This new

proposed theory is probabilistically accepted or rejected, depending on how well it explains the data compared to the current theory, as well as how much simpler or more complex it is.

Ullman et al. (2010) suggested that this method can explain how human learners, including young children, can rationally approximate an ideal Bayesian analysis. This method allows a practical learner to search over a potentially infinite space of theories, holding on to one theory at a time and discarding it probabilistically as new, potentially better alternatives are considered.

Bonawitz and colleagues have begun to explore how well this MCMC approach captures children's inferences about magnetic objects. Magnetism provides an interesting domain in which to conduct this investigation, because the space of possible kinds of causal interactions, the number of possible groups of objects, and the specific sorting of objects into those groups is very large. In particular, we can consider the search problem at multiple stages. First, given no evidence, we can consider which theories are a priori most likely. Second, given informative but still ambiguous data we can see how the probability of various theories will change. Third, given disambiguating data we can see if the system converges on the correct answer.

Observing unlabeled but potentially magnetic objects, like unlabeled blocks, interacting with two labeled instances of objects from causally meaningful categories (ie. blocks that are labeled with North and South polarities) provides a particularly interesting test. No matter how many observations are provided

between the unlabeled and labeled blocks, ambiguity remains: an ideal learner would not be able to infer whether the actual law is that like attracts like and opposites repel or whether the law is that likes repel and opposites attract. Bonawitz, Ullman, Gopnik, and Tenenbaum implemented a grammar-based Metropolis Hastings algorithm of magnetism discovery following this ambiguous evidence. Their model discovered these two possible alternative theories and these two theories scored highest in the search. Given that both of these theories were consistent with the observed data and were intuitively simple, this shows that stochastic search is an algorithm that can indeed be used to find reasonable theories. After providing disambiguating evidence, the model was also able to pick out the single, most likely theory.

These modeling results and those of Ullman, Goodman, and Tenenbaum (2010), which inspired this investigation, demonstrate that in practice, the MCMC algorithm can use relatively minimal data to effectively search through an infinite space of possibilities, discovering likely candidate theories and sorting of objects into classes.

Bonawitz, Ullman, Gopnik, and Tenenbaum are also empirically examining children's reasoning about magnets, to see whether children search through and evaluate hypotheses in a way consistent with the model predictions. In their ongoing studies, children are asked about their beliefs at different phases of the experiment: before they observe any evidence, after they observe some ambiguous (but still informative) evidence, and after they observe disambiguating evidence. They have found that prior to observing the evidence, children

entertain a broad space of possible causal theories about the possible groupings and interactions between the magnets. These hypotheses reflect the prior probabilities over theories generated by the grammar. Following the ambiguous evidence, children rationally respond by favoring the two “best” theories (that likes attract and opposites repel, and that likes repel and opposites attract), as predicted by the results of the search algorithm. When the children see a single, disambiguating intervention (e.g. when two objects sorted into the same group interact and either attract or repel), they converge on the correct theory—even when this means abandoning the theory they just held.

Strikingly, neither the initially ambiguous evidence, nor the single disambiguating trial are sufficient to infer the correct theory. Nevertheless, children are able to make an inductive leap during the experiment. They simultaneously integrate the partially informative (but still ambiguous) evidence given by the initial observed interactions with the final disambiguating trial between two unlabeled blocks. Thus, even in the course of a short experiment, preschool-aged children are able to solve a simple version of the chicken-and-egg problem in a basically rational way. They search through a space of possible hypotheses and integrate multiple pieces of evidence across different phases of the experiment.

Markov chain Monte Carlo algorithms provide an account of how a learner could move through a potentially infinite space of possible hypotheses, and still produce behavior consistent with exact Bayesian inference. There are several directions in which this line of research can be extended. One important step is

to understand and characterize the “building blocks” for intuitive theories. Following from this, it will be interesting to investigate how it might be computationally plausible for a system to learn to use simple algorithms to construct complex theories from these building blocks (Kemp, Goodman, & Tenenbaum, 2010). A second extension is to apply these models to “common sense” domains such as physics, psychology, and biology, and to the “real world” theories that children actually learn. Developmental learning mechanisms for this kind of abstract knowledge are currently poorly understood

## **6 Discussion**

We began this chapter with two problems for the idea that probabilistic models can capture how children learn intuitive theories – the algorithm problem and the variability problem. We’ve suggested that the Sampling Hypothesis may provide an answer to both these problems. In the first experiment we showed that children’s causal inferences have some of the key signatures of sampling – particularly a pattern of probability matching that goes beyond naïve frequency matching.

We then introduced specific sampling algorithms that approximate Bayesian inference. First we found that preschoolers’ responses on a causal learning task were better captured by a Win-Stay, Lose-Shift algorithm than by independent sampling. An attractive property of the Win-Stay, Lose-Shift algorithm is that it does not require the learner to compute and resample from the full posterior distribution after each observation. These results suggest that even

responses that sometimes appear non-optimal may in fact represent an approximation to a rational process, and provide an account of how Bayesian inference could be approximated by learners with limited cognitive capacity.

We also presented an account of how a learner might search through a potentially infinite hypothesis space, inspired by computational models which include Markov Chain Monte Carlo searches over logical grammars (Ullman, Goodman, & Tenenbaum, 2010). These searches include randomly proposed changes to a currently held theory, which are probabilistically accepted, dependent on the degree to which the new theory better accounts for the data. These same search and inference capacities may help to drive theory change in the normal course of children's cognitive development. At the least, Bonawitz et al.'s current experiments suggest that preschool-aged children are able to discover a correct theory from a space of many possible theories. Children search through a large space of possible hypotheses and are able to integrate multiple pieces of evidence across different phases of the experiment to evaluate the best theory.

### *6.1 Open Questions*

We have suggested that a learner could search through a hypothesis space in a number of ways, dependent perhaps on the task demands, developmental change, or even individual preference. Which algorithm a learner uses may also depend on the efficiency of the algorithm. However, how we define efficiency may depend on how difficult it is to compute posterior

probabilities, and then how difficult it is to generate one or a few samples from the posterior. Efficiency may require considering how many samples must be observed before the correct posterior is approximated and the cost of each observation. Thus, which algorithms are most efficient may depend on the nature of the problem being solved and on the capacities of the learner. So, we don't have good answers to when specific algorithms may be favored over others and in which contexts, but it's an important line for future research.

We can also ask whether sampling behavior is rational. A casual answer is "yes"—because we show how a "rational" or "computational" level analysis can be approximated at the algorithmic level. However, again assessing rationality depends on the goals of the learner. In some circumstances, a learner may want to quickly converge on the most likely answer. In other circumstances, however, the learner may want to explore more of the possibilities. These "exploit" or "explore" strategies might lead a learner to use different kinds of algorithms. Sampling and searching through a space of hypotheses may be a particularly useful learning mechanism for exploratory learning. It allows a learner the possibility of discovering an unlikely hypothesis that may prove correct later (after observing additional data). Were a learner to simply maximize, always choosing the most likely hypothesis, he might miss out on such a discovery.

One of the most promising implications of examining learning at the algorithmic level is that other aspects of development (e.g. memory limitations, changes in inhibition, changes in executive function) can be connected more explicitly to rational models of inference. For example, a particle filter



approximates the probability distribution over hypotheses at each point in time with a set of samples (or "particles"), and provides a scheme for updating this set to reflect the information provided by new evidence. The behavior of the algorithm depends on the number of particles. With a very large number of particles, each particle is similar to a sample from the posterior. With a small number of particles, there can be strong sequential dependencies in the representation of the posterior distribution. Developmental changes in cognitive capacity might correspond to changes in the number of particles, with consequences that are empirically testable.

Finally, we suggested that moving forward also involves connecting the algorithms that children might be using to carry out learning with ways in which the algorithms could be implemented in the brain. Ma et al (2006) suggest that cortical circuits may carry out sample-based approximations, reflecting the variability in the environment. Probabilistic sampling algorithms can also capture ways in which inputs should be combined (e.g. across time, sensory modalities, etc) taking the reliability of the input into account, and recent research on neural variability demonstrates this in the brain (Fetsch, Pouget, DeAngelis, & Angelaki, in press; Beck, et al., 2008). Other work may examine the implication of how growing dense connections between brain regions connect to particular algorithms and how those algorithms are affected as regions are pruned (as in later adolescence.)

## **7 Conclusions**

In the course of development, children change their beliefs, moving from a less to more accurate picture of the world. How do they do this given the vast space of possible beliefs? And how can we reconcile children's cognitive progress with the apparent irrationality of many of their explanations and predictions? The solution we have proposed is that children may form their beliefs by randomly sampling from a probability distribution. This Sampling Hypothesis suggests a way of efficiently searching a space of possibilities in a way that is consistent with probabilistic inference, and it leads to predictions about cognitive development. The studies presented here suggest that preschoolers are approximating a rational solution to the problem of probabilistic inference via a process that can be analyzed as sampling, and that the samples that children generate are affected by evidence. By thinking about the computational problems that children face and the algorithms they might use to solve those problems, we can approach the variability of children's responses in a new way. Children may not just be effective learners despite the variability and randomness of their behavior. That variability, instead, may itself contribute to children's extraordinary learning abilities.

## 8 References

- Beck, J., Ma, W. J., Kiani, R., Hanks, T. D., Churchland, A.K., Roitman, J. D., Shadlen, M. N., Latham, P. E., & Pouget, A. (2008). Bayesian decision making with probabilistic population codes. *Neuron*, *60*, 1142-1152.
- Bonawitz, E., Denison, S., Chen, A., Gopnik, A., & Griffiths, T. L. (2011). A simple sequential algorithm for approximating bayesian inference. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*.
- Bonawitz, E.B., Fischer, A., & Schulz, L. (in press). Teaching three-and-a-half year olds to reason about ambiguous evidence. *Journal of Cognition and Development*.
- Bonawitz, E. B., & Griffiths, T. L. (2010). Deconfounding hypothesis generation and evaluation in Bayesian models. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.
- Bonawitz, E.B, & Lombrozo, T. (in press) Occam's Rattle: Children's use of simplicity and probability to constrain inference. *Developmental Psychology*.
- Bonawitz, E.B., van Schijndel, T., Friel, D., & Schulz, L. (2012) Balancing theories and evidence in children's exploration, explanations, and learning. *Cognitive Psychology*, *64*, 215-234.
- Chen, Z. & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development*, *70*(5), 1098-1120.

- Denison, S., Bonawitz, E. B., Gopnik, A., & Griffiths, T. L. (2010). Preschoolers sample from probability distributions. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.
- Denison, S., Reed, C. & Xu, F. (in press) The emergence of probabilistic reasoning in very young infants: evidence from 4.5- and 6-month-old infants. *Developmental Psychology*.
- Doucet, A., de Freitas, N., & Gordon, N. J. (Eds.) (2001). *Sequential Monte Carlo methods in practice*. Berlin: Springer-Verlag.
- Fetsch, C. R., A. Pouget, A., DeAngelis, G. C., & Angelaki, D. E. (in press). Neural correlates of reliability-based cue weighting during multisensory integration. *Nature Neuroscience*.
- Galton, F. (1907) Vox Populi. *Nature* 75, 450-45.
- German, T.P., & Nichols, S. (2003). Children's inferences about long and short causal chains. *Developmental Science*, 6, 514-523.
- Gerstadt, C.L., Hong, Y. J., & Diamond, A. (1994) The relationship between cognition and action: performance of children 3 1/2–7 years old on a Stroop-like day–night test. *Cognition* 53, 129–153.
- Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. Boca Raton, FL: Chapman and Hall/CRC.
- Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, 116, 661-716

- Gopnik, A., Glymour, C., Sobel, D., Schulz, L., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 1-31.
- Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5), 620-629.
- Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*, 108-154.
- Greeno, J. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53*, 5–26.
- 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.
- Inhelder, B. & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.
- Jones, M. H., & Liverant, S. (1960). Effects of age differences on choice behavior. *Child Development, 31*(4), 673-680.
- Kemp, C., Goodman, N. & Tenenbaum, J. (2010). Learning to learn causal models. *Cognitive Science, 34*(7), 1185-1243.

- Kording, K., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244-247.
- Kushnir, T. & Gopnik, A., (2005). Children infer causal strength from probabilities and interventions. *Psychological Science*, *16*, 678-683.
- 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*, *44*, 186-196.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press.
- Lucas, C. G., Gopnik, A., & Griffiths, T. L. (2010). Developmental differences in learning the forms of causal relationships. *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.
- Ma, W. J., J. M. Beck, Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*, 1432-1438.
- Marr, D. (1982). *Vision*. Freeman Publishers.
- Mozer, M., Pashler, H., & Homaei, H. (2008). Optimal predictions in everyday cognition: The wisdom of individuals or crowds? *Cognitive Science*, *32*, 1133-1147.
- Neal, R. M. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto.

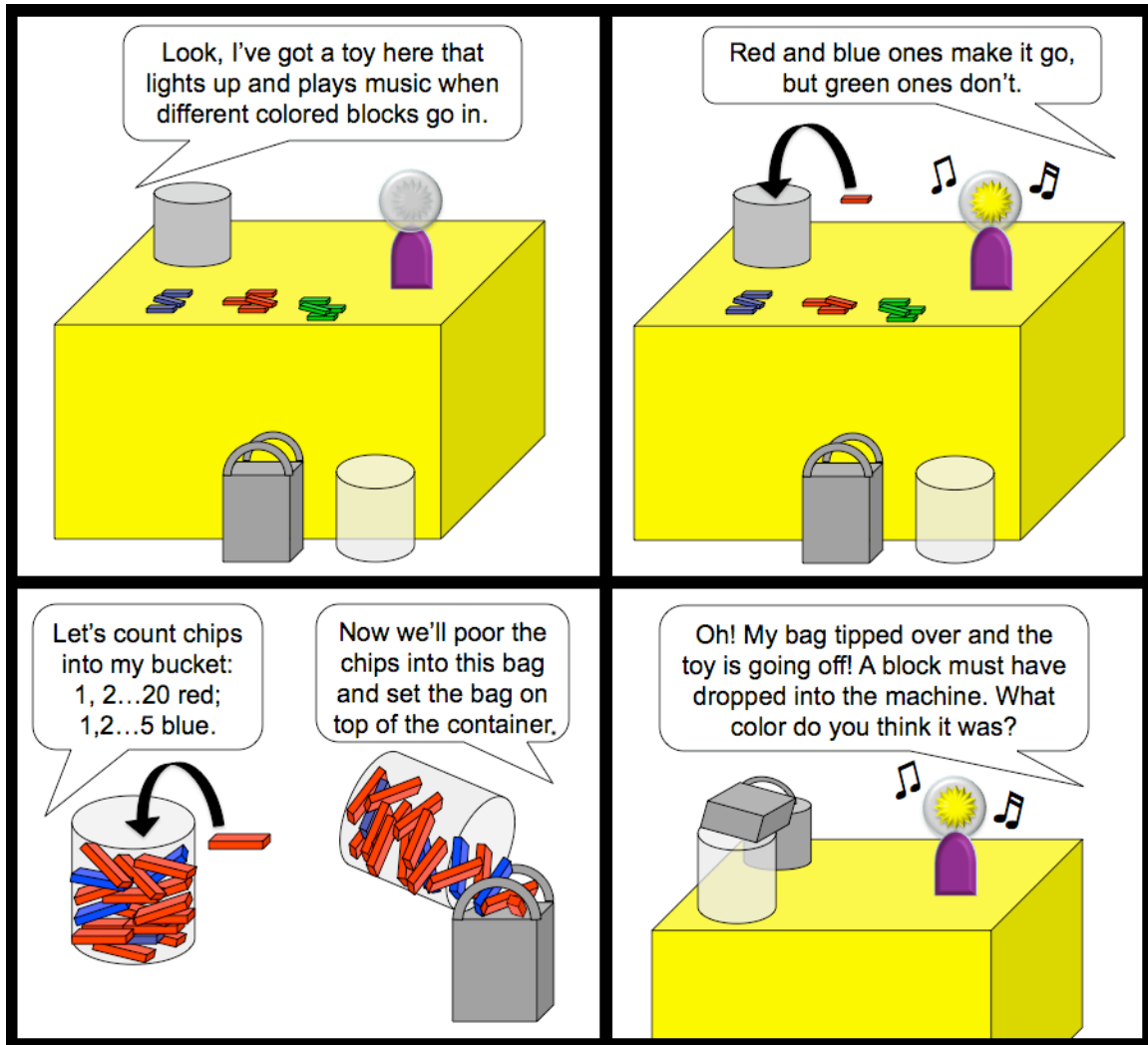
- Nowak, M., & Sigmund, K. (1993). A strategy of win-stay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature*, 364, 56-58.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, UK: Cambridge University Press.
- Piaget, J. (1983). Piaget's theory. In P. Mussen (Ed.), *Handbook of Child Psychology* (4th ed., Vol. 1). New York: Wiley.
- Restle, F. (1962). The selection of strategies in cue learning. *Psychological Review*, 69, 329-343.
- Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58, 527-535.
- Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York.
- Russell, S. J., & Norvig, P. (2003), *Artificial Intelligence: A Modern Approach* (2nd ed.), Upper Saddle River, NJ: Prentice Hall.
- 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.
- Schulz, L.E., Bonawitz, E. B., & Griffiths, T. L. (2007). Can being scared make your tummy ache? Naive theories, ambiguous evidence and preschoolers' causal inferences. *Developmental Psychology*, 43, 1124-1139.
- Schulz, L.E., Goodman, N.D., Tenenbaum, J.B., & Jenkins, C.A. (2008). Going beyond the evidence: Abstract laws and preschoolers' responses to anomalous data. *Cognition*, 109(2), 211-223.

- Schulz, L.E., Gopnik, A., & Glymore, C. (2007) Preschool children learn about causal structure from conditional interventions. *Developmental Science* 10(3), 322-332
- Shi, L., Feldman, N. H., & Griffiths, T. L. (2008). Performing Bayesian inference with exemplar models. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.
- Seiver, E., Gooman, N.D., Gopnik, A. (in press) Did she jump because she was the big sister or because the trampoline was safe? Causal inference and the development of social attribution. *Child Development*.
- Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.
- 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.
- Spirtes, P., Glymour, C., & Schienens, R. (1993). *Causation prediction and search*. New York: Springer-Verlag.
- 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*. Little, Brown.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, 10, 309-318.

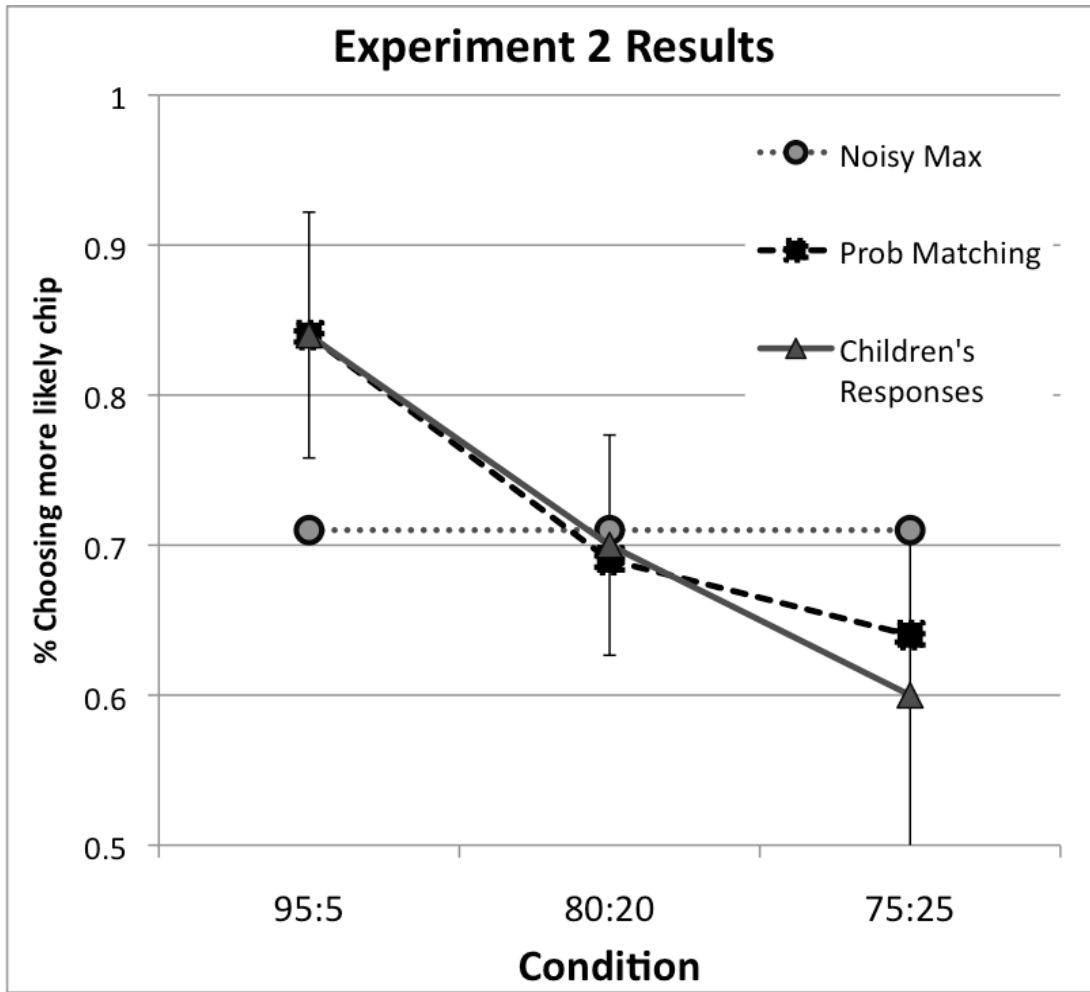


- Thelen, E. & Smith, L. B. (1994) *A dynamics systems approach to the development of perception and action*. MIT Press.
- Ullman, T., Goodman, N., & Tenenbaum, J. (2010). Theory acquisition as stochastic search. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the 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. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.
- Vul, E., & Pashler, H. (2008). Measuring the Crowd Within: Probabilistic Representations Within Individuals. *Psychological Science*, 19 (7), 645-647.
- Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, 5(6), 598-604.
- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245-272.

**Figure 1:** Stimuli and procedure used for testing the Sampling Hypothesis in children.



**Figure 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.



**Figure 3:** Method for WSLS and Bayesian posterior probability and children's data from WSLS experiment for each block, red (R), green (B), and blue (B) after observing each new instance of evidence, using parameters estimated from fitting the Bayesian model to the data.

