Preschoolers sample from probability distributions

Stephanie Denison (smdeniso@berkeley.edu) **Elizabeth Baraff Bonawitz (liz_b@berkeley.edu)**
Alison Gopnik (gopnik@berkeley.edu) **Thomas L. Griffiths (tom_griffiths@berkeley.edu)**
Department of Psychology, University of California, Berkeley, Berkeley, CA 94720 USA

Abstract

Researchers in both educational and developmental psychology have suggested that children are not particularly adept hypothesis testers, and that their behavior can often appear irrational. However, a growing body of research also suggests that people do engage in rational inference on a variety of tasks. Recently researchers have begun testing the idea that reasoners may be sampling hypotheses from an internal probability distribution when making inferences. If children are reasoning in this way, this might help to explain some seemingly irrational behavior seen in previous experiments. Forty 4-year-olds were tested on a probabilistic inference task that required them to make repeated guesses about which of two types of blocks had been randomly sampled from a population. Results suggest that children can sample from a probability distribution as evidenced by the fact that, as a group, they engaged in probability matching and that the dependency between successive guesses decreased over time.

Keywords: Cognitive Development; Causal Learning; Approximate Bayesian Inference; Sampling Hypotheses

Introduction

Young children are faced with a variety of novel situations on a daily basis. They encounter countless episodes in which they must reason about why particular events unfold the way they do, what this means in terms of how related events might unfold in the future, and how this newly acquired information fits into the knowledge they already possess. Humans revise their beliefs throughout development, often beginning with relatively flawed beliefs and progressing towards an increasingly accurate portrayal of the world. However, no current theory provides a satisfactory explanation of how children decide which hypotheses to test. Somehow they must search through the potentially infinite number of hypotheses that exist at the beginning of the learning process. Here, we investigate this question by asking whether young children can make probabilistic inferences via a process of sampling hypotheses from probability distributions.

The question of whether children and adults are capable of using rational inference to search through a hypothesis space and revise their beliefs has drawn mixed empirical findings. To begin, Piaget noted that children tend not to reason systematically about hypotheses, at least until they reach the formal operational stage in late childhood (Piaget, 1983). Since Piaget, some researchers have found evidence to corroborate this claim, stating that children often appear to navigate randomly through a selection of predictions and explanations (Siegler & Chen, 1998). For example, researchers in educational psychology have revealed evidence suggesting that young children and even non-expert adults are not particularly skilled hypothesis testers (e.g., Kuhn, 1989; Klahr, Fay, & Dunbar, 1993). Furthermore, developmental psychologists have found that children often revise their beliefs surprisingly slowly, suggesting a struggle to efficiently update theories (e.g., Carey, 1991; Wellman, 1990).

On the other hand, at least two kinds of evidence exist to suggest that children might be capable of using rational inference to generate, search through and evaluate hypotheses. First, recent research in cognitive psychology suggests that people reason in ways that are consistent with optimal Bayesian models in a variety of tasks (e.g., Griffiths & Tenenbaum, 2005; Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Although most of this work examines adult reasoning, a growing body of evidence suggests that children can also reason in a way that is consistent with Bayesian inference (e.g., Gopnik et al., 2004; Kushnir & Gopnik, 2005; Schulz & Gopnik, 2004; Schulz, Bonawitz, & Griffiths, 2007; Goodman et al., 2006). For example, Xu and Tenenbaum (2007) found that preschoolers can systematically integrate prior knowledge regarding hierarchical information with evidence in order to apply the correct labels to a variety of objects in a word learning task and 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. Second, many researchers advocate the theory-theory of conceptual development, which states that children's knowledge is organized into abstract, coherent conceptual systems, similar to those found in science (Carey, 1985; Gopnik & Meltzoff, 1997; Murphy & Medin, 1985; Wellman & Gelman, 1992). This framework predicts that children will engage in hypothesis testing in ways similar to scientists during learning, and much evidence has accumulated in support of this view (e.g. see Karmiloff-Smith & Inhelder, 1974; Bonawitz, Lim, & Schulz, 2007; Legare, Gelman, & Wellman, in press). However, the theory-theory does not specify where the hypotheses are derived from in the first place or how children could be expected (albeit unconsciously) to compute full Bayesian inference over (often) infinite hypothesis spaces.

The sampling hypothesis

Although Bayesian inference corresponds well to the theory-theory of conceptual development, researchers who advocate a rational approach to human inference do not suggest that adults and children 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 both from a formal and a practical standpoint, given the large number of hypotheses that would require consideration. One way to think about how the mind may be approximating Bayesian inference is to start with good engineering answers to this problem. Techniques for approximating Bayesian inference have

already been developed in the fields of machine learning and statistics and we can see whether humans are also using some version of these strategies.

One strategy for implementing Bayesian inference is sample-based approximation (Shi, Feldman, & Griffiths, 2008; Sanborn, Griffiths, & Navarro, 2006). This approach states that people might be approximating Bayesian inference by evaluating a small sample of the many possible hypotheses. This "sampling hypothesis" has been supported by additional empirical data that suggest people often base their decision on just a few samples (Goodman et al., 2008; Mozer, Pashler, & Homaei, 2008). Indeed, in many cases an optimal solution is to take only one sample (Vul, Goodman, Griffiths, & Tenenbaum, 2009). Sampling partly involves picking a hypothesis at random from the distribution. However, the process is not entirely random in that distributional information may be used to generate hypotheses that are highly likely more often than those that are less likely. This strategy allows the learner to entertain a variety of hypotheses, ensuring that they will spend more resources testing likely hypotheses but will not overlook a lower probability hypothesis that could turn out to be correct.

The sampling strategy predicts "probability matching": aggregating over numerous samples, generated by different individuals in a group, should return the original distribution; as the number of samples approaches infinity, the closer the result will be to the distribution. This benefit of averaging is called the "wisdom of crowds". If instead people generate a "best guess", then aggregating over numerous samples should result in an inaccurate reflection of the distribution, characterized by an overweighting of the most likely hypothesis. Sampling also depends on independence between guesses; the more independent the draws from the distribution, the more accurate the sample will be. However, we might expect that if a single individual is generating multiple guesses, then there may be dependence between guesses, but this dependence may decrease as time between guesses increases.

Recently, Vul and Pashler (2008) tested the sampling hypothesis in adults. They asked individuals to make guesses about a variety of real-world statistics such as: What percentage of the world's airports are in the United States? In an *Immediate* condition, participants were asked to make guesses about a variety of real-world statistics and then asked the questions a second time directly after. In a *Delayed* condition, the question was asked for the second time two weeks later. It was found that an individual's error was reduced when their guesses were averaged compared to each of their individual guesses in both the Immediate and Delayed conditions. There was also a greater benefit of averaging guesses in the Delayed group than in the Immediate group; the independence of guesses and, therefore, accuracy was greater after a time delay. This suggests that adults were most likely sampling guesses from an internal distribution rather than always providing an optimal guess.

The results from Vul and Pashler (2008) suggest that adults may be approximating rational solutions when making guesses about frequencies, in accordance with the sampling hypothesis. We turn to the question of whether or not children are drawing samples from probability distributions in a similar way. We explore two predictions of the sampling hypothesis. First, if children use a strategy of sampling hypotheses from a distribution, we should see that the probability with which they select hypotheses should match the distribution. This contrasts with a strategy of maximizing (always choosing the most likely answer) or guessing (randomly providing responses, independent of their probability), which make different predictions. We will refer to this as the *probability matching* prediction. Second, because sampling depends on independence, we can predict that increasing dependencies between guesses will decrease the degree to which responses accurately reflect the distribution. We will call this the *dependency* prediction.

While results of several studies seem to suggest that children do in fact probability match in numerous situations (e.g. see Kam & Newport, 2009; Kushnir, Wellman, & Gelman, 2008; Bonawitz, Chang, Clark, & Lombrozo, 2008; Sobel, Tenenbaum, & Gopnik, 2004), to our knowledge, no research has demonstrated the dependency prediction, or analyzed results in terms of the sampling hypothesis. While much research has demonstrated the sophisticated graded response of children on average, any particular child's response is often, paradoxically "non-optimal." That is, often developmental studies involve forced-choice responses, and so the predictions of any single child seem in conflict to rational models: Why wouldn't children simply always choose the most likely response, rather than some fraction of children choosing the likely response and some smaller fraction choosing the unlikely response? If children are in fact approximating rational inference by sampling hypotheses at least in some situations, this may provide an account of these data. More importantly, the sampling hypothesis may also provide an account of how children navigate through potentially infinite hypothesis spaces during learning: rather than computing full Bayesian inference over the whole hypothesis space, children sample a subset of hypotheses. We now turn to our experiment to explore this question.

Do children sample hypotheses?

We investigate the sampling hypothesis in preschool-aged children by testing their ability to use probability information to make guesses about which of two colored blocks was most likely to be sampled from a population (consisting of a 4:1 ratio) on a single random draw. This design allows us to investigate whether children demonstrate the first of two sampling signatures: probability matching. First, we predict that if individual children are sampling from a distribution of hypotheses, their responses will be closer to the correct distribution (i.e., 80% red blocks) than would be predicted by random guessing (50% red, blue guesses) or maximizing (100% red guesses). Second, the dependency prediction suggests that

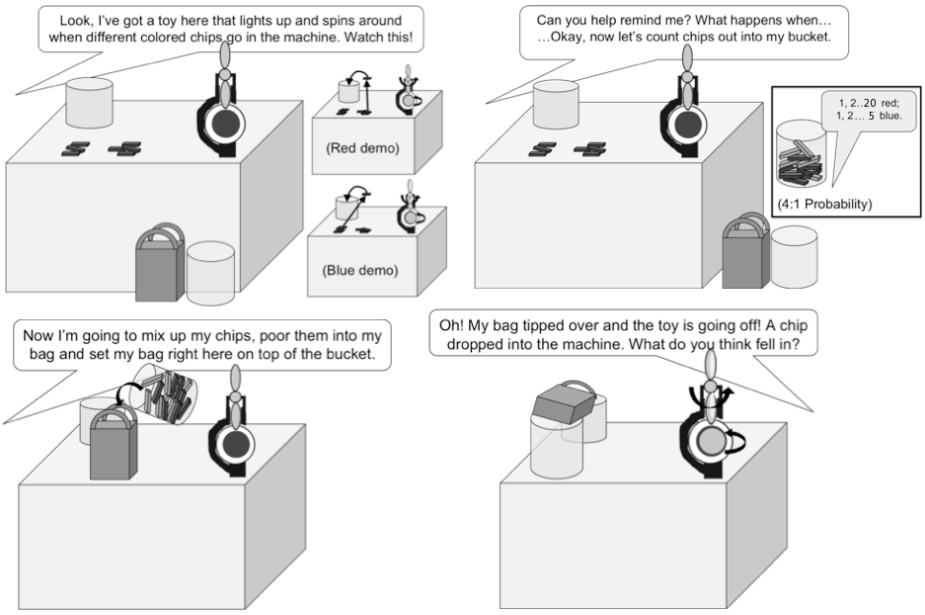

Figure 1: Stimuli and method used to test the sampling hypothesis in children.

children who are given a long delay between guesses will demonstrate greater independence of guesses across trials than children who provide guesses following only a very short delay and we can model these dependencies via a Markov process. As a result of differences in independence between guesses, the distribution over guesses in the *Long Wait* condition should be closer to the predicted distribution than the distribution over guesses in the *Short Wait* condition.

Methods

Participants Forty 4-year-olds were recruited from 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, which included 8 males and 12 females (M = 54 months; R = 48 mos – 62 mos) and the *Short Wait* condition, which included 11 males and 9 females (M = 53 months; R = 48 mos – 59 mos). On additional child was tested and excluded due to failure to pass an initial comprehension check (see procedure below). The children's ethnicities reflected the composition of the area.

Stimuli A large box (12in $\times$ 12in $\times$ 18in) constructed out of cardboard and covered in yellow felt previously used in Bonawitz et al. (2008) was used. All five surfaces excepting the back side of the box were intact and covered with felt. A hole was cut out of the top of the box in the front right corner where a toy with a transparent sphere with lights and a spinner inside connected to a cylindrical shaft was inserted such that only the sphere was visible to the children. The toy activated by pressing a button on the shaft, causing it to light up and play music. An opaque activator bin was placed on the back left corner of the box. Additional stimuli included red, blue, and green domino sized wooden blocks; one red, one yellow and one green paper cup; a rigid green bag; and a transparent container. (See Figure 1).

Procedure *Short Wait Condition.* Each testing session was videotaped for data retrieval and a second experimenter recorded all responses online. The experiment began with the child and experimenter sitting across from one another 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. 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. Previous work suggests that children (and even adults) find this manipulation compelling (Bonawitz et al., 2008). A comprehension check was then performed to ensure children remembered that the blue and red blocks make the toy activate and that green blocks did not. Next, the experimenter began Trial 1 by having the child count 20 red blocks and 5 blue blocks one at a time and placing them into a transparent container. The order of block color was counterbalanced. After counting the blocks, the experimenter shook the blocks in the container to mix them and poured them into the rigid bag. She then placed the container upside down in front of the activator bin on the yellow box and placed it on top of the container. She then accidentally knocked it 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! 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] chip?" Once children provided an answer, 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 off. 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. At the end of Trial 3, the majority of children were asked three follow-up questions: the experimenter asked which color they guessed fell into the activator on the first, second, and third trials.

Long Wait Condition. The *Long Wait* condition was identical to the *Short Wait* except that children completed Trial 1 in the first testing session, Trial 2 in a second testing session one week later, and Trial 3 in a third testing sessions one week after Trial 2.

Results

There were no age differences between groups ($t(38) = 0.11, p = ns$). Responses were coded by first author and reliability coded by a research assistant blind to experimental hypotheses. All responses uniquely and unambiguously were either "red" or "blue" and agreement was 100%.

Probability Matching As expected, looking only at the first responses, there were no differences between conditions, $\chi^2(1, n = 40) = 1.9, p = ns$. To assess whether or not children probability matched, we averaged the first response of children in both the *Long Wait* and *Short Wait* condition. Overall, children's responses reflected probability matching (70% providing the more probable chip response). That is, results suggest that children were not simply randomly guessing, as responses were significantly different from chance ($p < .05$; binomial test), but not significantly different from the predicted distribution of .8 (p =*ns*, binomial test). Similarly, children were not "maximizing" by always providing the most probable response (i.e. always choosing the red chip), or responses would have approached ceiling.

Dependency Measures To assess whether children's responses were independent from one another across trials, we first assessed what the independent sampling assumption would predict. That is, given probability θ of sampling a particular chip, 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 2^3 or eight 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 chip (i.e. $(.8) * (.8) * (.2)$). In this way, we can compute probabilities for all eight hypotheses. We compared the expectation to the observed distribution of children in the *Short Wait* and *Long Wait* conditions (see Table 1).

Table 1: Pattern of responses expected under independent sampling compared with frequencies in the *Long Wait* and *Short Wait* conditions.

Responses	Expectation	*Long Wait*	*Short Wait*
red,red,red	.512	10	1
red,red,blue	.128	1	1
red,blue,red	.128	2	10
red,blue,blue	.032	3	0
blue,red,red	.128	0	1
blue,red,blue	.032	1	6
blue,blue,red	.032	1	1
blue,blue,blue	.008	2	0

Chi-squared tests revealed a significant difference between children's responses in the *Short Wait* condition to both the *Long Wait* condition, $\chi^2(7, N = 40) = 22.3, p < .05$, and to the expected distribution, $\chi^2(7, N = 20) = 18.6, p < .05$.[1] However, the difference between the *Long Wait* condition and the expected distribution was not statistically significant, $\chi^2(7, N = 20) = 6.57, p = ns$. This suggests that while children in the *Long Wait* condition were providing responses that followed the predictions of independent samples, children in the *Short Wait* condition were doing something else. Indeed, a quick examination of Table 1 suggests that children in the *Short Wait* condition were simply alternating responses. To directly compare the two conditions, we coded children's responses in terms of whether they repeated a response (e.g. "red" then "red" again) or alternated (e.g. "red" then "blue"). Comparing condition by repetition/alternation revealed significant differences both when we coded for repetition/alternation over all three responses, Fisher Exact $(N = 33), p < .0001$, and when we coded for repetition/alternation over two responses, $\chi^2(1, N = 80) = 29.5, p < .0001$.

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 chip" response to a "blue chip" response, and so forth (see 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 $(N = 20), p < .0001$, and a much weaker dependency in the *Long Wait* condition, Fisher Exact $(N = 20), p = .03$. These results suggest that although children's pattern of responses in the *Long Wait* condition were close to the predicted distribution, there were still dependencies between a single child's guesses.

[1] 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.

Table 2: 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

We conducted one final analysis to rule out the hypothesis that children in the *Short Wait* condition showed more dependency in responses than children in the *Long Wait* condition purely because children in short wait were more likely to remember their guesses. If children in the *Long Wait* condition had simply forgotten their previous responses more often than children in the *Short Wait* condition, they would be much less likely to show dependencies between guesses simply due to memory differences between trials. Recall that at the conclusion of the experiment children were asked which color block they had said fell in on each of the three previous trials. Looking at whether children answered all questions correctly, we found no difference in memory between conditions, $\chi^2(1, N = 32) = 3.14, p = .08$. However, because there was arguably a marginal difference between conditions, we also gave the children a memory score from 0-3 depending on how many memory questions children answered correctly; comparing memory scores also revealed no significant differences, $t(30) = -1.52, p = ns$.

Discussion

Our experiment examined whether children's responses in a simple causal reasoning task could be accounted for in terms of sampling from a probability distribution. The results of the experiment provide evidence in support of the sampling hypothesis in children. First, children's behavior reflected probability matching. That is, as a group, children provided a percentage of red and blue 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 50% of each color. Second, children's responses reflected the predicted patterns of independence and dependence across conditions. After delays of one week, children showed a greater amount of independence between guesses than did children who did not experience a delay. Furthermore, in contrast to results from the *Long Wait* condition, analyses of the *Short Wait* condition revealed that individual children showed strong dependence between their three guesses; thus these children were not randomly sampling from the distribution.

One might ask whether the findings suggesting that children are probability matching in our experiment were an artifact of our particular design. If children were aware that they would be asked the same question multiple times, they might not have been motivated to provide an optimal response, knowing that they would have two more chances to provide guesses. However, children were not aware that they would be playing the game multiple times in either the *Long Wait* or the *Short Wait* conditions. Furthermore, it is unlikely that such young children would be capable of engaging in such sophisticated planning. Moreover, at the conclusion of the three trials we asked children whether they remembered the guesses they had provided on each trial. Across both conditions children's memories were fuzzy, with the majority of them only being able to accurately report one or two of their initial responses in both the *Long Wait* and *Short Wait* conditions and there was no difference between conditions on the memory check.

Although our findings were consistent with the sampling hypothesis in that children both probability matched and displayed greater independence of guesses given a time delay, we did not find evidence for a "wisdom of crowds" effect. The wisdom of crowds predicts that when guesses are aggregated across individuals, this should provide a score that is closer to the actual distribution than the individual guesses alone. Instead, we found no differences between children's first guess and the majority of three responses, $\chi^2(1, N = 40) = .23, p = ns$. Given that Vul and Pashler (2008) found this benefit with adults, we might have expected to find a similar increased advantage of aggregation with children. However, we elected to use a forced choice paradigm due to the young age of our participants, and this may have reduced the sensitivity of our measure such that we were unable to detect the effect. In future work we may explore this further by designing a task that would allow children to make more fine-grained responses.

Future Work

Future work will continue to evaluate the sampling hypothesis in children to investigate the role of evidence in children's hypothesis generation and sampling. For example, we are looking at whether young children are capable of rapidly updating hypotheses based on evidence during a causal learning task. The prediction following the Sampling Hypothesis is that children will update their hypothesis space following either confirming or disconfirming evidence and will adjust their predictions accordingly, and should sample their next hypothesis from the remaining possible hypotheses.

Another future direction will involve investigating the sampling hypothesis in even younger children and current research suggests some possible appropriate methods. In an experiment examining single-event probability, Denison and Xu (in press) used a crawling procedure to show that 13-month-old infants can make predictions about single-event probability. They used two trials, one to establish which of two object-types individual infants preferred and another to test probabilistic inference. They showed infants two large populations of objects, one with a 4:1 ratio of desirable: not-desirable objects and the other with the opposite ratio. The experimenter removed a single item from each of the two populations one at a time and placed them into separate opaque containers. The infant was then encouraged to crawl to the container of their choice. Findings suggested that infants could predict which of the two populations would most likely yield a single-item

sample of their preferred object.

Finally, although other work suggests that children do demonstrate graded sensitivity to probabilities with similar designs (Bonawitz et al., 2008) and we chose a sample probability that maximized the difference between chance response and a strategy of maximizing, further conditions could strengthen our findings here by demonstrating that children's responses match probabilities across an array of values. For example, ongoing studies in our lab suggest that preschool-aged children's first responses do also match to samples where the probabilities are 19:1, 15:5, 12:6, and 10:10. Furthermore, we can demonstrate that children can sample from probability distributions in a more complex hierarchical sampling task. We have adapted the current procedure to show children an overall population of blocks that is physically separated into two sub-populations with different distributions. This design allows assessment of children's ability to make valid probabilistic inferences when they must take into account the condition that the block is being sampled from only one of the two sub-populations.

Conclusions

The current experiment provides a first step in examining the sampling hypothesis in children. Children in our experiment engaged in probability matching and demonstrated increased independence of guesses when given a time delay, suggesting that they may have engaged in a process of sampling from probability distributions. This sampling behavior may begin to explain how children navigate through the potentially infinite number of hypotheses they face at the outset of a learning process. More generally, the sampling hypothesis may also begin to explain how it is that children's behavior can appear irrational when examined individually but may actually reflect a rational strategy overall.

Acknowledgments. Thanks to participating daycares and families, as well as Tiffany Tsai, Madeline Hanson, Beth McCarthy and Jennifer Ng for help with data collection. This research was supported in part by the James S. McDonnell Causal Learning collaborative and grant IIS-0845410 from the National Science Foundation.

References

Bonawitz, E., Chang, I., Clark, C., & Lombrozo, T. (2008). Ockham's razor as inductive bias in preschoolers causal explanations. In *Proceedings of the 7th international conference of development and learning.*

Bonawitz, E., Lim, S., & Schulz, L. (2007). Weighing the evidence: Children's theories of balance affect play. In *Proceedings of the 29th annual conference of the cognitive science society.*

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (1991). Knowledge acquisition: Enrichment or conceptual change? In S. Carey & S. Gelman (Eds.), *Epigenesis of mind: Essays on biology and cognition.* Hillsdale, NJ: Erlbaum.

Denison, S., & Xu, F. (in press). Twelve- to 14-month-old infants can predict single-event probability with large set sizes. *Developmental Science*.

Goodman, N., Baker, C., Bonawitz, E., Mansinghka, V., 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.*

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

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., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Cambridge, MA: MIT Press.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

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

Karmiloff-Smith, A., & Inhelder, B. (1974). If you want to get ahead, get a theory. *Cognition*, *43:3*, 195-212.

Klahr, D., Fay, A., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology*, *25*, 111-146.

Kuhn, D. (1989). Children and adults as intuitive scientists. *Psychological Review*, *96*, 674-689.

Kushnir, T., & Gopnik, A. (2005). Young children infer causal strength from probabilities and interventions. *Psychological Science*, *16:9*, 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.

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

Legare, C., Gelman, S., & Wellman, H. (in press). Inconsistency with prior knowledge triggers children's causal explanatory reasoning. *Child Development*.

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

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*, 289-316.

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

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Schulz, L., Bonawitz, E., & Griffiths, T. (2007). Can being scared give you a tummy ache? naive theories, ambiguous evidence and preschoolers causal inferences. *Developmental Psychology*, *43*, 1124-1139.

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

Shi, L., Feldman, N., & Griffiths, T. L. (2008). Performing Bayesian inference with exemplar models. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society.*

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

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

Vul, E., Goodman, N. D., Griffiths, T. L., & Tenenbaum, J. B. (2009). One and done? optimal decisions from very few samples. In *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.

Wellman, H. M. (1990). *The child's theory of mind*. Cambridge, MA: MIT Press.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337-375.

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