Title
Inductive Inference in Infants and Young Children: The Role of Probabilistic Reasoning

Permalink
https://escholarship.org/uc/item/4ff9g022

Author
Denison, Stephanie Mia

Publication Date
2012

Peer reviewed|Thesis/dissertation
Inductive Inference in Infants and Young Children:
The Role of Probabilistic Reasoning

by

Stephanie Mia Denison

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Psychology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Prof. Fei Xu, Co-Chair
Prof. Alison Gopnik, Co-Chair
Prof. Thomas L. Griffiths
Prof. Michael A. Ranney

Fall 2012
Inductive Inference in Infants and Young Children: The Role of Probabilistic Reasoning

Copyright 2012
by
Stephanie Mia Denison
Abstract

Inductive Inference in Infants and Young Children: The Role of Probabilistic Reasoning

by

Stephanie Mia Denison
Doctor of Philosophy in Psychology
University of California, Berkeley

Prof. Fei Xu, Co-Chair
Prof. Alison Gopnik, Co-Chair

A recent surge of research in cognitive developmental psychology examines whether human learners, from infancy through adulthood, reason in ways consistent with Bayesian inference. However, when exploring this question an important first step is to identify the available inference mechanisms and computational machinery that might allow infants and young children to make inductive inferences.

A number of recent studies have asked if infants may be “intuitive statisticians,” making inferences about the relationship between samples and populations in both looking-time and choice tasks. In this dissertation, I present three sets of empirical experiments in support of infants’ ability to make probabilistic inferences. The first empirical chapter examines the age at which infants begin making probabilistic inferences. I present an experiment suggesting that 6- but not 4-month-old infants can make generalizations about the likely composition of a large population after observing the contents of a small sample drawn randomly from that population. The second empirical chapter presents seven experiments that compare infants’ and adults’ abilities to make more complex probabilistic inferences. These experiments suggest that infants can integrate both probabilistic and deterministic physical constraints in probabilistic inference, and that in some cases, they show greater competence in doing so than adults. The final chapter exploring infant probabilistic inference presents evidence from three experiments with 10- to 12-month-old infants. In these experiments, results suggest that infants can use single-event probability computations to make predictions about where to direct their search to locate a desired object. These experiments also suggest that infants use proportional reasoning and not a simple heuristic based on comparisons of absolute quantity to make probabilistic inferences.

In the fourth empirical chapter I present data from a series of four experiments conducted with preschool-aged children. My colleagues and I use a causal learning task that is broadly similar in structure to the series of infant experiments reported in Chapter 4: Children are asked to use single-event probability computations to make causal inferences. The key motivation of this chapter is to reconcile divergent findings in the literature suggesting that, on the one hand, children reason in ways that are consistent with rational inference and, on the other hand, children tend to produce responses that are quite variable in nature. This chapter outlines a proposal termed the “Sampling Hypothesis”, which suggests that the variability in young children’s responses may be part of a rational strategy for inductive inference. In the reported experiments, we find evidence to suggest that children sample responses from the distribution of
possible hypotheses that explain the observed data, weighting the different hypotheses according to their probability. This chapter provides an illustration of one of the ways in which a learner’s ability to engage in probabilistic inference, which comes online early in infancy, can provide the foundation for more complex inductive inferences later in development.

In the final chapter I discuss the implications of this work, point to a number of remaining open questions, and consider some future directions for this line of research. I conclude with the suggestion that infants and young children are much more sophisticated at making probabilistic inferences than was previously thought. The competences demonstrated by infants and young children in the reported experiments appear to draw on an intuitive probability notion that is early emerging and unavailable for conscious reflection. Moreover, I suggest that young learners are capable of making rapid inductive inferences by capitalizing on their ability to compute probabilities, in order to acquire knowledge in a variety of domains, including causal reasoning.
To my mom, who inspired this work through her love of infants and of me.
Contents

List of Figures. ................................................................. vi
List of Tables. ................................................................. vii

1. Introduction. ................................................................. 1
   1.1 An emerging theoretical perspective. .............................. 2
   1.2 Goals of the present dissertation. .................................. 2
      1.2.1 Goals of the infant probabilistic inference experiments. . 2
      1.2.2 Goals of the preschooler causal learning experiments. .... 3
   1.3 Background literature. ................................................. 5
      1.3.1 Probabilistic inference in young children ...................... 5
      1.3.2 Rudimentary probabilistic inference in infancy ............... 5
   1.4 Précis. ................................................................. 7

2. The Emergence of Probabilistic Reasoning in Very Young Infants .... 8
   2.1 Introduction ........................................................... 8
   2.2 Methods ............................................................... 11
      2.2.1 Participants ...................................................... 11
      2.2.2 Materials ......................................................... 12
      2.2.3 Apparatus ......................................................... 12
      2.2.4 Design and Procedure ......................................... 12
      2.2.5 Predictions ....................................................... 13
   2.3 Results. ............................................................... 13
   2.4 Discussion ........................................................... 15

3. Integrating Physical Constraints in Probabilistic Inference: A Comparison of
   Infant and Adult Cognition ................................................. 18
   3.1 Introduction ........................................................... 18
   3.2 Experiment 1a: Can Infants Integrate a Probabilistic Constraint? .. 19
      3.2.1 Methods ........................................................... 19
         3.2.1.1 Participants .................................................. 19
         3.2.1.2 Materials .................................................... 19
         3.2.1.3 Apparatus .................................................... 20
         3.2.1.4 Procedure .................................................... 20
3.2.1.5 Design ...................................................... 21
3.2.1.6 Predictions ............................................... 22
3.2.2 Results ....................................................... 23
3.2.3 Discussion ................................................. 23
3.3 Experiment 1b: Replication .................................... 24
3.3.1 Method ........................................................ 25
3.3.1.1 Participants .............................................. 25
3.3.1.2 Materials .................................................. 25
3.3.1.3 Procedure ............................................... 25
3.3.1.4 Design and Predictions ................................. 25
3.3.2 Results ........................................................ 25
3.3.3 Discussion ................................................... 26
3.4 Experiment 1c: Can Adults Integrate a Probabilistic Constraint? .... 26
3.4.1 Method ........................................................ 26
3.4.1.1 Participants .............................................. 26
3.4.1.2 Procedure ............................................... 26
3.4.2 Results ........................................................ 27
3.4.3 Discussion ................................................... 27
3.5 Experiment 2a: Can Infants Integrate a Deterministic Constraint? .... 27
3.5.1 Method ........................................................ 28
3.5.1.1 Participants .............................................. 28
3.5.1.2 Materials .................................................. 28
3.5.1.3 Procedure ............................................... 28
3.5.1.4 Design and Predictions ................................. 28
3.5.2 Results and Discussion ..................................... 29
3.5.2.1 Comparison of Experiments 1b and 2a .............. 29
3.6 Experiment 2b: Can Adults Integrate a Deterministic Constraint? .... 30
3.6.1 Methods ........................................................ 30
3.6.1.1 Participants .............................................. 30
3.6.1.2 Procedure ............................................... 30
3.6.2 Results ........................................................ 30
3.6.3 Discussion ................................................... 30
3.7 Experiment 3: Can Adults Integrate a Probabilistic Physical Constraint when Prior Knowledge is Mitigated? ................................................. 31
3.7.1 Methods ........................................................ 31
3.7.1.1 Participants .............................................. 31
3.7.1.2 Materials and Procedure ................................ 31
3.7.2 Results ........................................................ 31
3.7.2.1 Comparison of Experiment 1c and Experiment 3 .... 32
3.7.3 Discussion ................................................... 32
3.8 General Discussion .............................................. 33

4. Single-Event Probabilistic Inference in Infancy .............................. 34
4.1 Introduction ................................................................ 34
4.2 Experiment 1: Quantity Control of Target Objects ...................... 35
4.2.1 Methods ........................................................ 35
4.2.1.1 Participants .................................................. 35
4.2.1.2 Materials ..................................................... 35
4.2.1.3 Procedure and Design ..................................... 36
4.2.2 Results and Discussion ..................................... 37

4.3 Experiment 2: Pitting Quantity Against Probability ..................... 37
4.3.1 Methods ....................................................... 38
4.3.1.1 Participants ................................................ 38
4.3.1.2 Materials .................................................. 38
4.3.1.3 Procedure, Design and Predictions ....................... 38
4.3.2 Results and Discussion ..................................... 38

4.4 Experiment 3: Controlling Quantity of All Objects ......................... 39
4.4.1 Methods ....................................................... 39
4.4.1.1 Participants ................................................ 39
4.4.1.2 Materials .................................................. 39
4.4.1.3 Procedure, Design and Predictions ....................... 39
4.4.2 Results and Discussion ..................................... 40
4.4.2.1 Results of Experiments 1 – 3 ............................. 40

4.5 General Discussion .............................................. 41

5. Rational Variability in Children’s Causal Inferences: The Sampling Hypothesis ................................................................. 43
5.1 Introduction ................................................................ 43
5.1.1 Belief Revision and Sampling ................................ 44
5.1.2 The wisdom of crowds ......................................... 45
5.1.3 Probability matching .......................................... 46
5.1.4 Testing the predictions of the Sampling Hypothesis ............... 47
5.2 Experiment 1: Sampling and Dependency ................................ 47
5.2.1 Methods ....................................................... 48
5.2.1.1 Participants ................................................ 48
5.2.1.2 Stimuli ...................................................... 48
5.2.1.3 Procedure .................................................. 48
5.2.2 Results .......................................................... 50
5.2.2.1 Probability matching on initial trial ....................... 50
5.2.2.2 Probability matching across all trials ..................... 50
5.2.2.3 Dependency Measures .................................... 51
5.2.3 Discussion ..................................................... 52
5.3 Experiment 2: Varying Proportions .................................... 53
5.3.1 Method .......................................................... 53
5.3.1.1 Participants ................................................ 53
5.3.1.2 Stimuli ...................................................... 53
5.3.1.3 Procedure .................................................. 53
5.3.2 Results .......................................................... 54
5.3.2.1 Probability matching ..................................... 54
5.3.2.2 Comparing probability matching and noisy-max ........... 54
5.3.3 Discussion ..................................................... 56
5.4 Experiment 3: Varying Proportions with Three Alternatives .......... 57
5.4.1 Method ................................................................. 57
  5.4.1.1 Participants ..................................................... 57
  5.4.1.2 Stimuli .......................................................... 57
  5.4.1.3 Procedure ...................................................... 58
5.4.2 Results .............................................................. 58
  5.4.2.1 Probability matching ........................................ 58
  5.4.2.2 Comparing probability matching and noisy-max ...... 58
5.4.3 Discussion ........................................................ 59
5.5 Experiment 4: Beyond Frequency Matching .................. 59
  5.5.1 Method ............................................................. 60
  5.5.1.1 Participants .................................................... 60
  5.5.1.2 Stimuli .......................................................... 60
  5.5.1.3 Procedure ...................................................... 61
  5.5.2 Results ............................................................. 62
  5.5.3 Discussion ........................................................ 62
5.6 General Discussion .................................................. 63
5.7 Conclusion ........................................................... 65

6. Conclusions .............................................................. 66
  6.1 Conclusions and implications of the empirical work ........ 66
  6.2 Remaining questions and future directions .................... 67
    6.2.1 Remaining questions and future directions from the infant
    work ................................................................. 67
    6.2.2 Remaining questions and future directions from the preschool
    experiments .......................................................... 69
  6.3 Conclusions ........................................................ 70

Bibliography ............................................................... 71
List of Figures

2.1 The sequence of a test trial in Experiment 1 (Xu & Garcia, 2008). The experimenter shakes the box, closes her eyes and draws out balls from the closed box. She then reveals the population. ................................................................. 9

2.2 The two possible outcomes in the current experiment. The population boxes displayed simultaneously ensure equal amounts of pink and yellow are displayed. .... 11

3.1 Schematic representation of test trials. .......................................................... 22

3.2 Illustration of predictions for Experiments 1 and 2. ........................................ 23

3.3 Comparison of adult performance in Experiment 1c versus Experiment 3. ........ 32

4.1 Schematic representation of a test trial. .......................................................... 37

5.1 Stimuli and procedure used for testing the Sampling Hypothesis in children. .... 39

5.2 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. ................................................................. 56
# List of Tables

2.1  Mean looking times in seconds to each box by age group. .......................... 15

4.1. Descriptive statistics and binomial probabilities collapsed across Experiments 1-3. . . . 41

5.1. Pattern of responses expected under independent sampling compared with frequencies in the long wait and short wait conditions. ....................................................... 51

5.2. Experiment 1 transition matrices in the two conditions. ................................. 52

5.3. Wald tests for the 95:5, 80:20 (Expt.1), and 75:25 conditions. .......................... 55
Acknowledgments

I thank my advisor, Fei Xu, for the countless hours she has spent providing insight into this work. Fei incited my interest in Developmental Psychology during my undergraduate degree, and has managed to keep my attention on the subject ever since. Her incredible ability to do high quality research on a wide range of topics has inspired me to continue in the field beyond graduate school. If I can someday become an expert in as large a number of topic areas as Fei, I will surely never lose interest in my career. I was also fortunate to have Alison Gopnik as an advisor. Alison has a true vision for the field, and makes a laudable investment in moving it forward. I came to Berkeley for the explicit purpose of immersing myself in a culture that is committed to addressing big picture questions in the field. Alison has provided the exact level of insight and breadth of education that I was hoping to receive here. I am extremely grateful to Liz Bonawitz and Tom Griffiths for their contributions to my education at Berkeley. Liz is possibly the most motivated human being I have ever encountered, and I have thoroughly enjoyed all of my collaborations with her. Tom’s influence on the way that I now approach research is much larger than should be possible given the relatively small amount of time we’ve spent working together. Liz and Tom heavily contributed to the contents of Chapter 5. Chapter 2 is based on a published paper and I thank Christie Reed for her contributions to that work.

I cannot thank my labmates enough for all of their support – intellectual and otherwise. In particular, Caren Walker, Brian Waismeyer, Jane Hu, Sophie Bridgers, Elizabeth Seiver, and Anna Waismeyer. Caren deserves extra recognition for spending more evenings and weekends working with me at coffee shops than either of us will ever be proud to admit – fortunately (or unfortunately for our productivity) we enjoyed this time more than one might guess. The undergraduate research assistants and staff in the Xu and Gopnik Labs also deserve thanks for their help with data collection, recruiting participants, and reliability coding.

Aside from Fei, who really inspired this dissertation intellectually, the largest amount of credit for allowing me to complete this work certainly goes to my family: My husband Richard has more faith in my abilities than he probably should and he has never once complained about the enormous sacrifices he’s had to make since I decided to attend Berkeley. Rich is intelligent, driven, and selfless and I know how lucky I am to have found a partner with this rare combination of personal attributes. I am also very thankful to my dad; his resilience is an inspiration and I am forever empowered by his pride in me. I owe an equally large amount of gratitude to my sisters, who among countless other things spent hours on the phone assuring me that this was all going to be worth it in the end. There are no words to describe how thankful I am to my mother for her constant support. As a neonatal nurse, she inspired my interest in development and my genuine love of babies, which made the day-to-day aspects of conducting this research so enjoyable. I wish so much that she could have had the opportunity to see this final product, as I’m certain she’s the only person other than my committee who would have read it cover to cover.

Of course this process would have been MUCH more difficult without a Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada.
Chapter 1

Introduction

What is the nature of early learning in infants and young children? What kinds of learning mechanism are responsible for the rapidly developing knowledge – about objects, people, causality, and numbers – that we see in infancy and beyond? Are we rational learners and can rational computational models capture human behavior? These questions have been of great interest to psychologists for many decades. This dissertation is focused on a body of new empirical evidence from infants and young children aimed at answering some of these questions.

The ability to make accurate inductive inferences based on limited data has implications for the longstanding debates in cognitive developmental psychology concerning the initial state of human learners and the learning mechanisms that can support conceptual development (e.g., Carey, 2009; Elman et al., 1996; Hirschfeld & Gelman, 1994; Smith, 2001; Spelke, 1994). Throughout the history of developmental psychology, these debates have divided developmental theorists into two opposing groups. One group, the nativists, tends to grant the human infant a great deal of initial conceptual knowledge. This theoretical outlook is usually coupled with the assumption that limited learning mechanisms need to be posited to support conceptual development, and that many developmental changes may be accounted for by brain maturation. The other group, the empiricists, tends to posit that human infants start out with only perceptual primitives and that they lack initial conceptual knowledge. This theoretical commitment is typically paired with the assumption that associative learning mechanisms are responsible for the accumulation of conceptual knowledge and conceptual change. Furthermore, developmental psychologists who fall in the nativist camp, positing early conceptual knowledge and minimal learning, tend to advocate domain-specific learning mechanisms. Conversely, those who fall in the empiricist camp, positing early perceptual processing and a large role for learning, tend to advocate data-driven, domain-general learning mechanisms.

Traditionally, developmental psychologists have taken a strong stance on both of these debates, placing themselves clearly on one side of the theoretical fence or the other. However, as both sides have provided more and more evidence in favor of their respective viewpoints, it has become increasingly clear that some philosophies from both camps are likely to have merit, while other aspects continually lack explanatory power for certain phenomena. Take for example the domain of word learning, where a satisfying theory must account for a number of known phenomena. Empiricist accounts of word learning (e.g., Colunga & Smith, 2005; Regier, 2003, 2005) account reasonably well for the fact that children are capable of learning words at
multiple levels of taxonomic hierarchies (e.g., they learn words such as animal, dog and poodle). However, they have difficulty dealing with the fact that children are able to learn the meaning of new words after observing very small numbers of exemplars, a phenomenon called fast-mapping, as the learning mechanisms typically posited by these accounts require a large number of object and label pairings to acquire new words. Conversely, more nativist approaches to word learning, which posit a number of innate constraints, can account for fast-mapping but they have difficulty accounting for how children acquire words at multiple levels of a taxonomy (e.g., Markman, 1989; Siskind, 1996). Thus, both nativist and empiricist approaches appear to deal quite readily with some aspects of word learning, but others simply cannot be explained. Unfortunately, the problem with confessing that both camps get parts of the argument right, and parts of the argument wrong, is that this can be viewed as theoretical fence sitting, which is typically frowned upon in science. Nonetheless, these strong dichotomies are dissatisfying to many developmentalists. This is likely due to the fact that everyone believes that there is some innate knowledge, they just might disagree over how structured or advanced it is, and everyone believes there is some role for learning but, again, they might disagree over how central the learning mechanisms are to the story of development.

1.1 An emerging theoretical perspective

Recently, Fei Xu and colleagues have introduced a theoretical framework that offers a middle ground between nativism and empiricism, termed Rational Constructivism (Xu, 2007; Xu, Dewar, & Perfors, 2009; Xu & Griffiths, 2011). One central focus of Rational Constructivism is to explore the role of domain-general inductive inference mechanisms in development. Inductive inference mechanisms may provide a more satisfying account of explaining how it is that children make rapid inferences from sparse data. They do not require that the learner bring a great deal of initial conceptual knowledge to any given task, but they can account for the rapid and accurate inferences that young children make with limited and imperfect data. In fact it has been suggested that some of the early constraints that young learners use to guide their inferences in domains such as word learning may be acquired through these learning mechanisms (Dewar & Xu, 2010; Griffiths et al., 2010; Kemp, Perfors & Tenenbaum, 2007; Smith et al. 2002). The inductive inference mechanisms central to Rational Constructivism are Bayesian in nature, allowing a role for both prior knowledge and input data in driving conceptual growth, and thus bridging the nativist-empiricist divide to provide an explanation for learning and development in early childhood. Inductive inference mechanisms based on Bayesian principles assume that the learner makes educated guesses about the probabilities of a set of hypotheses, and these degrees of belief can be updated when additional pieces of evidence are acquired. They allow the learner to make inductive leaps based on minimal amounts of stochastic data – data similar in kind to the type of input that humans receive in the real world.

1.2 Goals of the present dissertation
1.2.1 Goals of the infant probabilistic inference experiments

The purpose of Chapters 2 through 4 is to explore how inductive inference gets off the ground in early infancy. If Bayesian inference is a good candidate to drive learning in early childhood and throughout the lifespan, an important first step is to identify the computational machinery that might allow infants and young children to make inferences of this nature. Although young children appear to make inductive inferences in ways that are consistent with Bayesian inference, the question of when these mechanisms come online, and whether or not they are available in early infancy remain largely unknown. One central pre-requisite to computing Bayesian inference, or approximations to Bayesian inference, is the ability to reason about probabilities and probabilistic data. In these chapters, I present research that explores the developmental origins of probabilistic inference in human infants. Do untutored infants have intuitions about probability that could serve as a pre-requisite to later inductive inference?

I report findings from three series’ of experiments to investigate four key components of probabilistic inference in infancy that are central to answering the question of whether or not it is a good candidate for an early-emerging inductive inference mechanism. First I explore the age at which infants begin to engage in rudimentary probabilistic reasoning – making generalizations from samples to populations. The age at which probabilistic reasoning comes online in humans will be vital to answering questions about when infants can begin to use probabilistic inference to gain domain-specific knowledge.

Second, if probabilistic inference is central to a domain-general inductive learning mechanism that can support the accumulation of domain-specific knowledge, these computations should be integrated with infants other domain knowledge early on. Previous research has explored this in the psychological domain (Gweon, Schulz & Tenenbaum, 2010; Xu & Denison, 2009) and the physical domain (Denison & Xu, 2010a; Teglas et al., 2011). The experiments reported in Chapter 3 push infants’ abilities further than in previous research, requiring them to fully integrate probability and naïve physics in a way that previous experiments have not required.

The experiments in Chapter 3 also address a third important question: Do infants rely on the representativeness heuristic or on true analytical probability computations when making judgments in our experiments? Adults often rely on judgment heuristics when making inferences that require probability computations, thus we assess whether or not infants and adults use heuristics to make inferences in our task in Chapter 3. Finally, while infants learn a great deal through observing other people’s actions, they also use their own actions as a means for learning, through activities such as object manipulation, play, and search.

The fourth question, addressed in Chapter 4, is whether or not infants can use their probabilistic computations to guide prediction and action. If infants can make probabilistic inferences and use this information to guide their interactions with the world, then probability computations could be instrumental for active learning in the first year of life. This chapter also addresses a confound that is not controlled in any of the previous literature on infant probabilistic inference. The experiments examine whether infants’ probability computations are based on an ability to reason using proportions or absolute quantities, as proportional reasoning is the hallmark of true probabilistic inference.

1.2.2 Goals of the preschooler causal learning experiments
Following the presentation of evidence suggesting that even the youngest human learners can reason using probabilistic data, I report a series of four experiments in Chapter 5 examining causal learning in 4- and 5-year-old children. These experiments serve as a concrete example of how probabilistic inference can be involved in more complex inductive inference. This work is motivated by two factors: First, the experiments attempt to reconcile a tension that exists in the cognitive developmental literature. On one side of the debate, recent research suggests that that young children engage in rational inference, producing responses in a variety of experimental settings that are consistent with the predictions of Bayesian inference (Bonawitz, Fischer, & Schulz, 2012; Goodman et al., 2008; Gopnik, Sobel, Schulz & Glymour 2001; Gopnik et al., 2004; Kushner & Gopnik 2005, 2007; Schulz, Gopnik & Glymour 2007; Lucas, Gopnik, & Griffiths, 2010; Schulz, Bonawitz, & Griffiths, 2007). These tasks, which investigate causal inference in childhood, typically manipulate both the evidence that children observe, and their beliefs about the prior probability of various hypotheses. Generally speaking, the findings from these experiments show that children choose hypotheses with the greatest Bayesian posterior probability. In contrast to these findings, a wealth of evidence in the field also suggests that young children are irrational and variable when making inferences and engaging in belief revision. Observations of this kind began with the work of Piaget (1983), who argued that children do not reason systematically about hypotheses until they reach the formal operational stage in late childhood. More recently, Siegler and Chen (1998) have found that children often veer randomly through a selection of different predictions and explanations when making inferences, rather than systematically updating and converging on increasingly accurate beliefs and theories.

The second motivation for the series of experiments reported in the chapter is to address a conundrum raised by taking a Bayesian approach to human inductive inference. Most of the research that takes a Bayesian approach to cognitive development has focused on generating a computational level analysis of human behavior. Although this approach has produced a wealth of important experimental findings, it also introduces a new problem. Namely, these experiments typically do not reveal the exact computations that children engage in when making rational inferences; that is, they do not offer an algorithmic level analysis of children’s behavior. This predicament becomes further complicated by the fact that engaging in full Bayesian computations, requiring enumeration and calculation of all possible hypotheses, is computationally intractable for most problems.

Chapter 5 introduces the “Sampling Hypothesis”, a proposal aimed at addressing both of these problems. The hypothesis states that a learner might internally represent a probability distribution of potential hypotheses, and then sample these hypotheses from this distribution, thus making the computations less cumbersome, as only a small number of hypotheses require consideration. If a learner uses this kind of strategy, they will produce responses that are variable, but this variability should be systematically related to the posterior probability of the hypotheses. Thus the learner will produce the most likely hypotheses with higher probability, but will also produce some of the less likely hypotheses. This should result in a distribution that is centered approximately on the normatively correct response, but also allows the learner to try out a few hypotheses that, although initially less likely, could turn out to be correct.

The Sampling Hypothesis brings two specific predictions of children’s behavior. First, on aggregate children should produce a distribution of hypotheses in a given task that is in line with the actual distribution of hypotheses. Thus children should show probability matching to
the posterior distribution. Second, children should show evidence of dependencies between their responses, as this is a hallmark of sampling, and this dependence should decrease as a function of time between guesses. To test the predictions of the Sampling Hypothesis, children are presented with a causal inference task. In each experiment, children learn about a toy that lights up and plays music when chips of different colors are placed into a bin on the toy. The distribution of hypotheses that children should consider is given to them through a process of counting out various numbers of colored chips to be placed in a bag. A chip then falls out of the bag and into the activator, in such a way that the child cannot see the color of the chip. The children are simply asked to guess what color chip made the machine activate. Each experiment examines whether or not, on aggregate, children produce guesses that are consistent with the distribution of block colors. Additionally, the first experiment examines whether or not children show dependencies between their guesses and whether or not that dependence decreases as time between guesses increases. The final experiment in the series uses a more complicated process to randomly select a block, wherein the frequency of chip colors and the probability of the colors activating the toy are placed in conflict via a physical constraint. This experiment assesses whether children’s tendency to probability match in these tasks reflects sophisticated probabilistic inference, or is merely the result of naïve tabulations of frequencies.

1.3 Background literature

Below I review two sets of literature that are not substantially reviewed in the empirical chapters, but are highly relevant to the questions addressed, and paradigms used in this dissertation.

1.3.1 Probabilistic inference in young children

The goals of this dissertation revolve around infants’ and young children’s abilities to reason using probability computations. There exists a long history of investigating probabilistic reasoning in young children, beginning with Piaget, and extending to contemporary researchers in both psychology and education. In Piaget’s original experiments, he asked whether or not young children had intuitions about quantitative proportions and randomization by assessing their ability to make an inference about the likely contents of a sample, based on the composition of a population (Piaget & Inhelder, 1951/1975). He showed 5- to 12-year-old children various distributions of two colors of tokens being placed in an opaque bag. He then simply asked the children which color they would be “most likely” to draw if they reached into the bag and pulled out just one token. In this task, children below the age of 7 did not systematically rely on the proportions of colored tokens to make their guesses and they did not appear to understand the concept of randomness. Children instead provided responses based on inappropriate elements such as idiosyncratic properties of the two colors; for example, “I’ll get red because I like red.” or an incorrect intuition of randomness, for example, “I’ll get red because red went first.”

Piaget’s conclusions have not gone unchallenged over the past 60 years. Yost, Siegel and Andrews revisited Piaget’s question in their 1962 paper, and found that 5-year-old children can succeed at analogous problems when the task is made more appropriate for a young age group. In particular, they found that children will provide correct answers at higher than chance levels.
when the verbal demands of the task are reduced. Convergent findings have been reported in subsequent experiments examining simple relationships between samples and populations with 4- and 5-year-old children (Acredolo, O’Connor, Banks & Horobin, 1989; Goldberg, 1966; Reyna & Brainerd, 1994). In addition to the evidence suggesting that preschool-aged children can make rudimentary probabilistic computations, more recent experiments have revealed that slightly older children, beginning at around ages 6 and 7, can make even very sophisticated inferences about the likely outcomes of uncertain events based on probability in a variety of contexts. For example, children make accurate probabilistic inferences in tasks involving complex judgments of expected values (Schlottmann & Anderson, 1994), and in tasks that require judgments based on the integration of prior probabilities and additional information (Girotto & Gonzalez, 2008; Gonzalez & Girotto, 2011).

1.3.2 Rudimentary probabilistic inference in infancy

Recently, researchers have begun investigating probabilistic inference in infancy, and the work in this dissertation is inspired by these paradigms. Two experiments, both employing the violation of expectation (VOE) looking-time paradigm, have investigated rudimentary probabilistic inference in infancy during the first year of life. The VOE looking-time paradigm capitalizes on the fact that infants look longer at events that they find unusual or surprising – events that violate their expectations. In a typical VOE experiment using visual displays, infants begin by looking at displays to become familiarized or habituated to the stimuli that will be used in the experimental session. Infants are then shown test trials, during which events that are more or less probable are shown, and infants’ looking times are recorded. The basic intuition is that infants should look longer at an outcome event that is less probable than one that is more probable.

Teglas, Gonzalez, Girroto & Bonatti (2007) used the VOE looking-time paradigm to ask whether 12-month-old infants could engage in probabilistic inference. In particular, they were interested in whether or not infants can make inferences about single-event probability, which in their paradigm requires the individual to observe a population of objects and reason about the likely outcome when just a single, random object is removed. Infants were familiarized to a lottery machine device displayed on a computer screen, containing four objects, in a 3 yellow to 1 blue ratio. On test trials, the machine was briefly occluded and infants were shown two alternating outcomes: either one of the yellow objects appeared to have exited from a chute in the bottom of the machine or the blue object exited from the chute. As a group, infants looked longer at the events in which the blue object exited the chute, suggesting that they found this outcome less probable. This suggests that 12-month-old infants realize that when a single object is randomly drawn from a 3:1 distribution, the object drawn is more likely to be of the majority kind.

Other studies have explored probabilistic inference in infancy by examining whether or not infants can reason about the relationship between large populations of objects and multi-object samples. Xu & Garcia (2008) tested 8-month-old infants’ abilities to make generalizations from small samples to larger populations using a VOE looking-time experiment with a live experimenter and live objects. The experiment began with infants viewing familiarization trials, during which an experimenter placed a single, covered box on the stage, shook it back and forth and removed the cover to show the infants the contents of the box – a
large collection of Ping-Pong balls. On alternating trials, the population of balls in the box was either in a 9:1 red to white balls ratio or a 9:1 white to red balls ratio, but these large boxes appeared identical when the covers were closed. This familiarization was followed by test trials, during which the experimenter again took out a large covered box and placed it on the stage, shook it back and forth a couple of times, closed her eyes, reached into the box and removed a sample. After removing the sample, the experimenter removed the cover on the box to reveal the population, always consisting of mostly red balls for one group of infants or mostly white balls for the other group. On alternating trials, the sample removed from the box consisted of either 4 red and 1 white Ping-Pong balls or 4 white and 1 red Ping-Pong balls (see Figure 2.1). Infants looked longer at the 4:1 white to red balls sample if they were in the group that was shown the mostly red box and they looked longer at the 4:1 red to white balls sample if they were in the group that was shown the mostly white box. Thus, infants in this experiment were able to assess what a population is likely to consist of, based on a small random sample. In a second experiment, Xu & Garcia (2008) demonstrated that infants could also make this inference in the reverse direction. That is, when the experimenter began by showing infants the open box filled with, for example, mostly red Ping-Pong balls and then removed the 4:1 and 1:4 samples on alternating trials, infants looked longer at the mostly white sample.

The two sets of experiments reviewed here suggest that 12-month-old infants can make single-event probabilistic inferences and that 8-month-old infants can make inferences from samples to populations and vice versa.
1.4 Précis

The remainder of this document is divided into five chapters. The first four chapters report experimental findings related to probabilistic inference in infancy and early childhood, using multiple methods and multiple age groups. In the sixth and final chapter, I recapitulate the results and implications of the experimental findings; discuss the relationship between the various results and suggest a number of future directions for the work.

Chapters 2 through 4 examine probabilistic inference in infancy. Chapter 2 investigates the age at which infants begin to engage in rudimentary probabilistic reasoning – making generalizations from samples to populations. Chapter 3 addresses a surprising conflict in the literature between experiments that examine adult versus infant probabilistic reasoning. Specifically, human adults have often been shown to use a set of heuristics in reasoning and decision-making, violating normative statistical rules, while human infants have been shown to be good intuitive statisticians. To begin to resolve this conflict, we conducted experiments with preverbal infants and adults using the same displays, which pit probability computations against the representativeness heuristic. Chapter 4 uses a novel experimental methodology directed at the question of whether or not infants can use single-event probability computations to make predictions that guide their actions.

Chapter 5 places the role of probabilistic inference in a different context – causal reasoning in preschool-aged children. This chapter explores the Sampling Hypothesis, which suggests that the variable nature of children’s responses in previous experiments may be part of a rational strategy for inductive inference. In particular, the chapter addresses the possibility that young learners may randomly sample from the set of possible hypotheses that explain the observed data, producing different hypotheses with frequencies that reflect their subjective probability.
Chapter 2

The emergence of probabilistic reasoning in very young infants

2.1 Introduction

Human learners are capable of making large inductive leaps in the face of data that are limited and often stochastic. It is an important and ubiquitous ability. For example, imagine a person from our hunter-gatherer ancestry trying to determine which types of trees produce berries that are good for eating. Let’s say they sample approximately five berries from a couple of trees and find that one tree produces four good tasting berries and the other only produces one or two. They may make the inference that berries from the former tree tend to be edible and that the latter tree type should be avoided. Or, picture a toddler attempting to learn her first words. She hears the word “doggie” in the presence of her family dog a few times, and she quickly generalizes the word to other dogs but not a stray cat or her pet hamster.

What are the cognitive mechanisms that allow human learners to make such rapid and often highly accurate inductive inferences? Recent research in cognitive development has focused on the origins of probabilistic inference in infancy as a possible starting point. First, 12-month-old infants can make inferences from populations to samples when reasoning about single-event probability (Denison & Xu, 2010b; Teglas et al., 2007; Teglas et al., 2011). Second, 8-month-old infants are capable of making inferences from small samples to large populations and vice versa (Xu & Garcia, 2008). Third, infants as young as 11 months take into account the implications of sampling conditions (e.g., random vs. non-random sampling) and object properties (e.g., solidity and cohesion) when making these inferences (Denison & Xu, 2010a; Gweon et al., 2010; Teglas et al. 2007; Xu & Denison, 2009).

In the current experiment, we explore the age at which infants begin to make inferences from samples to populations, looking for the first time at infants younger than 8 months. We ask whether very young infants can make basic probabilistic inferences using a variant of the paradigm first introduced by Xu and Garcia (2008). In their experiments, a looking-time paradigm was employed to reveal whether 8-month-old infants have an intuitive ability to make generalizations from samples to populations. In Experiment 1, infants were shown samples being drawn from a large covered box and, on alternating trials, the experimenter either removed 4 red and 1 white balls or 4 white and 1 red balls. Then the experimenter revealed the population of balls – a 9:1 ratio of red to white balls. Infants looked longer at the 4 white and 1 red ball sample (the improbable outcome) than the 4 red and 1 white ball sample (the probable outcome; see Figure 2.1).
Figure 2.1. The sequence of a test trial in Expt. 1 (Xu & Garcia, 2008). The experimenter shakes the box, closes her eyes and draws out balls from the closed box. She then reveals the population.

Although this might suggest that infants have a rudimentary ability to reason about probability, the authors note two possible interpretations of this looking pattern: The first, which we will call the “probabilistic account”, suggests that infants looked longer at the 4 white and 1 red balls sample because they understand the predictive relationship between samples and populations and thus they considered it to be a relatively improbable sample. The second, termed here the “perceptual mismatch account”, suggests that infants simply prefer to look at displays wherein the population box and sample container contrast in perceptual appearance. That is, infants simply looked longer at trials displaying the less probable sample because it created a perceptual mismatch between the two displays present on stage (see the outcomes in Figure 2.1). This account represents a lower-level interpretation of infant performance, as it predicts an identical looking pattern as the probabilistic account but does not require an understanding of the relationship between the sample and population.\(^1\)

To distinguish between these two interpretations, Xu and Garcia (2008, Expt. 3) designed a control experiment in which the 4:1 sample was no longer drawn from the population. Eight-month-olds participated in a procedure that was equivalent to the one just described except that

\(^1\) Adults viewed the Expt. 1 displays and rated the improbable outcome as “unexpected” and the probable outcome as “expected.” They did not note perceptual mismatches or probability in their explanations. This suggests that computations of probabilities may generally be largely implicit and inaccessible to conscious thought.
the relationship between the sample and population was eliminated: the samples were drawn from the experimenter’s pocket rather than from the box. This resulted in identical test trial displays to those in Experiment 1 but in this case, infants had no reason to expect a relationship between the contents of the small container and the population box. Eight-month-old infants looked about equally when the mostly red box was displayed with the 4 red and 1 white balls sample (the perceptual match) and the 4 white and 1 red balls sample (the perceptual mismatch). Therefore, when the relationship between the box and container was eliminated, neither display violated infants’ expectations. This provides evidence in favor of the probabilistic account of infants’ performance in Experiment 1, i.e., that infants were reacting to the relative improbability of the sample and not the perceptual mismatch.

After obtaining evidence that 8-month-olds can reason about samples and populations, Xu and Garcia began to explore whether even younger infants possess similar intuitions. It seems plausible that younger infants could succeed at a version of this task given evidence revealing sensitivity to statistical input from newborns to 6-month-olds in domains such as phoneme discrimination and visual pattern learning (e.g., Bulf, Johnson, & Valenza, 2011; Kirkham, Slemmer & Johnson, 2002; Maye, Werker, & Gerken, 2002). Thus, they tested a group of 6-month-old infants using the same procedure as Xu and Garcia (2008). The findings were inconclusive. Infants performed as expected in the replication of Experiment 1, looking longer at trials in which the experimenter sampled 4 white and 1 red balls than 4 red and 1 white balls from the mostly red population. However, infants continued to follow this looking pattern in the control experiment during which the experimenter drew from her pocket (Xu & Garcia, unpublished data).

Although this pattern of findings does not support the probabilistic account, it also cannot definitively rule it out. It is possible that younger infants appreciate the relationship between samples and populations but also look longer at the perceptual mismatch in the control task because they continue to react to the perceptual features of the mismatch when the sample was drawn from the experimenter’s pocket. Unfortunately the experimental design cannot tease apart these two interpretations.

In the current experiment, we use a design appropriate for testing probabilistic inference in younger infants wherein the perceptual mismatch is eliminated but the displays remain easy to process. We equated the overall quantity of each ball color present in the population boxes during test trials by keeping two complementary boxes on display throughout (see Figure 2.2). Each test trial began with the two covered population boxes and a small transparent container to hold a sample on stage. The experimenter drew the infants’ attention to each box and drew a sample of, e.g., 4 pink and 1 yellow balls from the box on the right and placed it in the container. Then the experimenter revealed that the box from which the sample was drawn had a 4:1 ratio of pink to yellow balls, and the other box had the opposite ratio. The trials alternated between a 4 pink and 1 yellow sample (the more probable sample) and a 4 yellow and 1 pink sample (the less probable sample). If infants are only sensitive to perceptual mismatches and not sampling, they should look equally at all test trials, as the large boxes on display have equal amounts of each color and the sample therefore creates a slight but equal mismatch across every trial. If, on the other hand, infants are sensitive to the relationship between the sample and population, they will look longer on trials where the less probable sample is drawn from the relevant population box.
Figure 2.2. The two possible outcomes in the current experiment. The population boxes displayed simultaneously ensure equal amounts of pink and yellow are displayed.

We tested both 4.5- and 6-month-old infants in this design because we did not have strong a priori predictions about the age at which this mechanism comes online. Although there is ample empirical evidence of statistical learning and probabilistic reasoning in the second half of the first year, experiments conducted on infants younger than 8 months are relatively scant. The most relevant findings with young infants are from research on visual statistical learning and conditional probability computations in early infancy. Evidence of visual statistical learning of transitional probabilities in infancy has been found in 2-month-old infants, and was recently extended to newborn infants (Bulf et al., 2011; Kirkham et al., 2002). On the other hand, investigations of conditional probability computations with young infants have revealed that this comes online much later, some time between 5 and 8 months (Sobel & Kirkham, 2006, 2007). Due to the differences in findings in these related abilities with young infants, 4.5- and 6-month-olds seem appropriate age groups with which to begin an investigation of rudimentary probabilistic reasoning.

2.2 Methods

2.2.1 Participants

Participants were 32 infants: Sixteen 6-month-olds (6 males; $M = 6.1$ months, $R = 5.5$ to 6.6 months) and sixteen 4–month–olds (12 males; $M = 4.5$ months; $R = 4.0$ to 5.0 months). Ten infants (six 6-month-olds; four 4-month-olds) were tested but excluded due to fussiness (4), inattention during sampling (2), providing looking times over 3.5 standard deviations above
average (1) or parental interference (3). Infants were recruited from the San Francisco Bay Area. Socioeconomic status and ethnicities reflected the general distribution of the area and infants were required to be exposed to English a minimum of 50% of the time. Infants received a small gift for their participation.

2.2.2 Materials

Ping-Pong Balls. A total of 166 (83 yellow and 83 pink) ping-pong balls were used.

Boxes and Containers. A small, transparent Plexiglas container with an open top (20 cm x 4.5 cm x 4.5 cm) was used to display the samples during test trials.

Two large (31 cm x 23.5 cm x 23.5 cm) boxes were used to display the populations during the familiarization and test trials. The boxes were rectangular cubes with Plexiglas windows to show the populations of ping-pong balls and a hidden center compartment to hold the samples that were removed from the box during test trials. From the infants’ perspectives, the box appeared as one single unit, filled completely with ping-pong balls. The Plexiglas display windows were covered with fabric curtains to ensure that the boxes appeared identical when the curtains were lowered. The “mostly pink” box contained 60 pink and 15 yellow balls (pink to yellow = 4:1); the “mostly yellow” box contained the opposite (pink to yellow = 1:4).

2.2.3 Apparatus

Testing occurred in a room divided in half by curtains spanning its width and height. The curtains had a cut-out above a puppet stage that measured 94 cm x 55 cm (width x height). The experimenter sat behind the stage with her upper body and head visible. A curtain could be lowered to conceal the experimenter between trials. A camcorder filmed the infant through a small hole in the curtain below the stage; it was connected to a TV monitor that an observer used to code looking times online using JHAB (R. Casstevens, 2007). A second camcorder recorded the experimenter’s behavior.

Infants sat in a high chair approximately 70 cm from the center of the stage. The parent sat next to the infant facing the opposite direction and was instructed to avoid looking at the stage.

2.2.4 Design and Procedure

Calibration. To calibrate each infant’s looking window, a squeaky toy was used to direct the infant’s attention to the outside parameters of the stage.

Free Play Phase. After calibration, the infant was shown a small open box with 3 pink and 3 yellow ping-pong balls. He/she was encouraged to play with the balls for approximately 30 seconds. This was done to demonstrate to infants that the balls were discrete objects.

Familiarization trials (4 trials). The experimenter placed the two large boxes on the stage 30 cm apart with the front curtains down. She shook the box on the right side of the stage, saying, “What’s in this box?” She then shook the box on the left saying, “What’s in this box?” She lifted the front covers of both boxes simultaneously, revealing the separate populations of mostly pink and mostly yellow balls, and said “Look, [baby’s name], look!” She then put her
head down and directed her gaze to the floor. The observer began timing upon hearing the second, “look”. Trials ended when the infant looked away for 2 consecutive seconds.

The boxes were presented in the same locations for all 4 familiarization trials within a single experimental session and positioning was counterbalanced across infants. Between familiarizations the boxes were removed from the stage and a back curtain was lowered. These trials were included to familiarize infants to the materials and to the general experimental procedure. Additionally, exposing infants to two contrasting populations might cue them to attend to ratios. The familiarizations lasted approximately 3 minutes.

**Test trials (6 trials).** On each test trial, the experimenter placed the two large boxes on the stage (keeping them in the same locations on all 6 trials) with the front curtains lowered. The experimenter always sampled from the box on her right. She shook each box one at a time while saying, “What’s in this box?” She then closed her eyes, turned her head away, and reached into the box on her right. She pulled out 3 ping-pong balls and placed them into the small Plexiglas container in the middle of the stage one at a time. She then closed her eyes, turned her head away, and reached into the box on her left (not pulling out any ping-pong balls) and placed her hand on top of the small Plexiglas container in the middle of the stage to mimic the sampling motions made with the box on the right. She then repeated these actions, pulling out 2 more ping-pong balls from the right hand box and mimicking this action with the left hand box. On alternating trials the sample removed from the population box was either 4 pink and 1 yellow balls or 4 yellow and 1 pink balls. Then the experimenter lifted the front covers of both boxes on the stage simultaneously and said “Look, [baby’s name], look!” She put her head down and directed her gaze toward the floor. The observer began timing upon hearing the second, “look”, and ended the trial after 2 consecutive seconds of looking away. Between trials, the stage was cleared. The test trials lasted approximately 6 minutes.

**Design.** The side that the population boxes (mostly pink or mostly yellow) were on and whether the infant saw the 4 pink and 1 yellow ball sample first were fully counterbalanced across infants.

**2.2.5 Predictions**

If infants are sensitive to the relationship between samples and populations (i.e., assuming random sampling, the composition of a sample is likely to reflect the overall composition of a population) they should look longer at test trials displaying outcomes that violate this expectation than outcomes that are in line with this expectation. Therefore infants who saw the experimenter sampling from the mostly pink population should look longer at trials in which 4 yellow and 1 pink balls were sampled than trials in which 4 pink and 1 yellow balls were sampled. Conversely, infants who saw the experimenter sampling from the mostly yellow population should show the opposite looking pattern.

**2.3 Results**

A second observer, blind to trial order, coded 50% of the infants offline. Interscorer
reliability averaged 92%. Preliminary analyses found no effects of gender, test trial order (probable-outcome vs. improbable-outcome first), or the population box sampled from (mostly pink or mostly yellow) for both age groups. There was also no difference in duration of looking on familiarization trials between the two age groups. Subsequent analyses collapsed over these variables.

Looking times for test trial outcomes were analyzed using a 2 x 2 ANOVA with outcome (probable vs. improbable) as the within-subjects factor and age (4.5-month-olds vs. 6-month-olds) as the between-subjects factor. A significant interaction between Outcome and Age was found, \( F(1, 30) = 7.03, p = .013 \), effect size \( (\eta^2) = .190 \). There were no other significant main effects or interactions.

To break down the interaction, we conducted follow-up t-tests exploring the effect of test trial outcome (probable vs. improbable) for each age group separately (see Table 1 for mean looking times). Six-month-old infants looked reliably longer at the improbable outcome \( (M=8.63s, SD=5.05) \) than the probable outcome \( (M=5.96s, SD=2.81) \), \( t(15) = 2.67, p = .011 \), effect size \( (d) = 0.679 \). Twelve of sixteen infants looked longer at the improbable outcome, more infants than would be expected by chance, binomial test, \( p = .038 \), 1-tailed. In contrast, 4.5-month-olds looked about equally at the improbable outcome \( (M=6.05s, SD=3.14) \) and the probable outcome \( (M=7.45s, SD=5.18) \), \( t(15) = 1.19, p = .250 \), effect size \( (d) = 0.321 \). Seven of sixteen infants looked longer at the improbable outcome, not different from chance, binomial test, \( p = 0.408 \), 1-tailed.

We also coded infants’ scanning behavior offline (eight 6-month-olds and fourteen 4.5-month-olds) to obtain more fine-grained information about whether infants of each age group attended to both boxes during test trials\(^3\). We calculated the average duration of looking to the sampled vs. the unsampled boxes during the sampling phase of the test trials, starting with the experimenter shaking the first box at the start of the trial until she finished sampling and revealed the populations for the online coder to begin timing. We performed this analysis to examine whether infants of each age group attended to the same parts of the stage during sampling. The 6-month-old infants looked approximately equally to both the sampled and unsampled boxes, \( F(1,7) = 9.22, p = .306 \), effect size = 0.148. The 4.5-month-olds did not show this pattern; they spent significantly more time attending to the unsampled box than the sampled box, \( F(1,13) = 19.21, p = .001 \), effect size = 0.596 (See Table 1 for means).

We also coded infants’ average duration of looking to each box during the display phase of the test trials, commencing when the populations were revealed (i.e., when the experimenter said, “look”) and ending when the online coder stopped timing (when the infant looked away for 2 consecutive seconds). This analysis addresses the potential concern that infants may have looked significantly longer at the sampled box than the unsampled box when the populations and samples were visible, and thus were reacting to a perceptual mismatch between the sampled box and the sample, rather than estimating probability. We ran a repeated-measures ANOVA with duration of looking towards the sampled vs. the unsampled box as a within-subjects factor and age group as a between-subjects factor. There were no main effects or interactions (all \( p \)-values > .5). Thus, when all of the perceptual information was in sight, (i.e., the populations and the

---

\(^2\) Reliability was calculated as the proportion of total time both observers agreed that infants were looking at the displays (see Kellman & Spelke, 1983). Thus, percentage agreement = 1-(Absolute difference in time between original and second coder/original coder). We then obtained an average across all 160 trials.

\(^3\) We coded half of the sample of 6-month-olds (8 randomly chosen infants) because the scanning behavior of these randomly selected infants showed no signs of potential differences. We coded our full sample of 4.5-month-olds with the exception of 2 infants whose data were unavailable because their videos were damaged.
sample were visible) infants of both age groups looked equally at both boxes (see Table 1 for means).

### Table 2.1. Mean looking times in seconds to each box by age group

<table>
<thead>
<tr>
<th>Data Source</th>
<th>Sampled Box (SE)</th>
<th>Unsampled Box (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Display Phase</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-Month-Olds (N=14)</td>
<td>2.35 (1.57)</td>
<td>2.02 (1.65)</td>
</tr>
<tr>
<td>6-Month-Olds (N=8)</td>
<td>4.92 (2.08)</td>
<td>4.93 (2.19)</td>
</tr>
<tr>
<td><strong>Sampling Phase</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-Month-Olds (N=14)</td>
<td>1.96 (.50)</td>
<td>5.13 (0.88)</td>
</tr>
<tr>
<td>6-Month-Olds (N=8)</td>
<td>3.41 (0.66)</td>
<td>4.49 (1.17)</td>
</tr>
</tbody>
</table>

### 2.4 Discussion

Our results suggest that 6-month-old infants can make generalizations from samples to populations. When perceptual features are equated and infants cannot react to displays based solely on perceptual mismatches, 6-month-old infants look longer at a less probable sample of, for example, 4 yellow and 1 pink balls drawn from a mostly pink box than at a more probable sample of 4 pink and 1 yellow balls. Four-month-olds did not show this pattern; they looked roughly equally at both samples. This suggests that the ability to make generalizations from samples to populations emerges at around 6 months of age.

In addition, further analyses were conducted to address the potential concern that, despite efforts to draw attention to both boxes on stage, infants only attended to the sampled box during the timed portion of the test trials, and then simply reacted to the perceptual features between that box and the sample. This was not the case. These analyses argue against the interpretation that 6-month-olds were simply reacting to perceptual mismatches: When the displays were revealed, both age groups attended to each box equally. This weakens the perceptual mismatch account of the 6-month-olds’ data, as infants of both ages attended to the same perceptual features during test trials but only the 6-month-olds demonstrated increased looking on trials when improbable samples were drawn.

It is still possible that infants used a reasoning bias known as the representativeness heuristic (i.e., samples and populations should be similar in appearance, e.g., Tversky & Kahneman, 1974) to make judgments in our experiment. This is a different concern than the methodological issue raised throughout regarding perceptual features in the displays, as this assumes that infants reasoned correctly about random sampling, but questions whether they used a reasoning shortcut rather than probability computations. Some evidence from older infants

---

4 We ran a number of additional analyses, none of which returned statistically significant results. We ran an ANOVA to determine if infants who were “correct” (i.e., looked longer overall at improbable samples) showed different scanning behavior than infants who were “incorrect” (i.e., looked longer at probable samples). There were no interactions between age, whether or not infants’ looking times were “correct”, and their duration of looks to each box location ($p’$s > .05). In addition, there were no effects of whether or not infants’ looking times were correct and the number of times they scanned between the two population boxes and the sample.
suggests that this is an unlikely alternative interpretation. At 8- and 11-months, infants are able to make probabilistic inferences when samples and populations do not match in appearance (Denison & Xu, 2010a; Denison & Xu, 2011). In these experiments, populations with three sets of balls were used and infants were shown that all green balls – 50% of the balls in the population – were stuck and unmovable, and they were required to compute on the remaining sets of red and yellow balls in the population. On the test trials, both the probable and improbable samples (which only contained red and yellow balls) looked quite different from the population, and infants were still able to make correct inferences. Furthermore, preliminary results in our lab suggest that even when representativeness and probability are put in direct conflict, 11-month-olds are able to reason correctly based on probabilities (Denison & Xu, 2011). Similar experiments with 6-month-olds will help us more directly rule out the interpretation that young infants simply use the representativeness heuristic.

Although we now have evidence of intuitive probabilistic reasoning in 6-month-olds, it appears that 4.5-month-olds may not share similar intuitions. Two explanations may account for the negative findings from the 4.5-month-olds. The first possibility is that these findings demonstrate a true developmental difference between 4.5- and 6-month-old infants in probabilistic reasoning. This developmental progression parallels the one found in earlier studies on using conditional probabilities in causal learning (Sobel & Kirkham, 2006, 2007). In order to make accurate probabilistic inferences, infants must track where samples are drawn from, and this may be an ability that 4.5-month-old infants lack. When we coded where infants looked during the sampling phase of our experiment, we found that, surprisingly, 4.5-month-olds spent about 70% of their time attending to the unsampled population box. Six-month-olds, on the other hand, spent time scanning the entire scene. Perhaps the 6-month-old infants were able to scan the scene and extract the relevant information for making generalizations whereas the 4.5-month-olds were not yet able to hone in on the most pertinent components of a scene when making probabilistic inferences. Infants at this age may not realize that it is necessary to attend to the source from which a sample is drawn in order to make accurate generalizations. Future experiments are needed to explore this possibility more directly. The current experiment represents a much-needed attempt to fill the gap in the literature on visual statistical inference in infants below 8 months of age.

The second possible explanation of the 4.5-month-olds’ null results is that the experimental procedures were not suitable for use with infants of this age. The younger infants may have looked longer to the unsampled box not because they lack a conceptual understanding relevant to sampling but because they were confused or distracted by the methodology. The mimicking action on the unsampled box could have disrupted infants’ performance in a number of ways: They may have thought that balls had been drawn from both boxes or they may have been distracted by the mimicking, causing them to focus on the mimicking rather than the true sampling. Future work may use a manipulation such as drawing out balls from the ‘unsampled’ box and returning them each time without ever putting them in the small display container. This design would still allow us to equate the actions made with both boxes, and the act of sampling and returning balls may be more familiar to infants than pretend sampling. Another possibility is to consider measures other than looking-time when testing this ability in very young infants, for example, ERP experiments have found evidence for statistical learning in newborns (Teinonen et al., 2009).

Our findings here are consistent with recent research applying probabilistic models of human cognition to experimental findings in infancy and early childhood (Schulz et al., 2007;
Teglas et al, 2011; Xu & Tenenbaum, 2007). One of the key goals of this enterprise is to identify the rational, inferential, and statistical learning mechanisms that exist early in life, and have the power to support conceptual development (Xu & Griffiths, 2011). Indeed, if humans’ beliefs are represented probabilistically as this class of theories assume then at minimum, infants should be able to represent and compute rudimentary probabilities. Most of the current Bayesian models focus on ideal-observer analyses of human behavior at a computational level, but recently, several models have attempted to capture data from infants and young children by using algorithms that approximate full Bayesian inference (e.g., Bonawitz et al., 2011; Denison et al., 2011; Teglas et al., 2011). On the empirical side, there is also new evidence suggesting that we must take into account resource constraints such as working memory load when investigating young infants’ learning algorithms (e.g., Bulf et al., 2011). For the current study, a task reducing the information-processing demand may reveal earlier competence for probabilistic reasoning in infants younger than 6 months.

We presented the first experiment exploring probabilistic inference in young infants. The findings suggest that 6- but not 4.5-month-old infants can make generalizations from small samples to larger populations. Our results, in combination with recent evidence from similar experiments, provide convergent support for early competence in probabilistic reasoning in infancy (Denison & Xu, 2010a, 2010b; Teglas et al., 2007, 2011; Xu & Denison, 2009; Xu & Garcia, 2008). These findings, both from looking-time and action-based measures are particularly impressive given the extensive experimental results suggesting that adults often make faulty probabilistic inferences in a wide range of tasks. For example, Tversky and Kahneman (1974, 1981) found that adult judgments were often hindered by the incorrect application of reasoning heuristics when making probabilistic inferences. The infant findings suggest that humans do have an intuitive, implicit probabilistic reasoning mechanism. Starting at around 6 months of age, infants appear to understand something about the predictive relationship between samples and populations; by the end of the first year, infants can compute probabilities in looking-time studies and the output of these computations can guide their action.
Chapter 3

Integrating Physical Constraints in Probabilistic Inference: A Comparison of Infant and Adult Cognition

3.1 Introduction

Volumes have been written on how human adults use heuristics or shortcuts in reasoning and decision making (Ariely, 2008; Gilovich, Griffin & Kahneman, 2002; Tversky & Kahneman, 1974; 1981). Adults appear to behave irrationally in many situations that require rudimentary probabilistic reasoning; instead they apply heuristics such as representativeness or availability, resulting in biases such as base-rate neglect and the conjunction fallacy. Meanwhile, recent studies have shown that young children, and even preverbal infants, are tiny intuitive statisticians. In causal learning tasks, for instance, 4-year-olds are sensitive to conditional probabilities in deciding which object is causally efficacious (Kushnir & Gopnik, 2007). In studies on children’s social reasoning, preschoolers are shown to be sensitive to whether objects drawn from a box of toys had been the result of a random draw or a non-random draw that reflected an agent’s goal or preference (Kushnir, Xu & Wellman, 2010). Even preverbal infants have been shown, time and again, to be capable of rudimentary probabilistic reasoning: 6- to 12-month-old infants are able to compute which outcomes are more probable in a variety of tasks, and they are able to integrate both physical and psychological constraints in their computations (Denison et al., 2012; Teglas et al., 2007; Xu & Denison, 2009; Xu & Garcia, 2008).

How is it that infants are so competent at probabilistic reasoning while adults seem rather bad at it? Are young learners better at statistical reasoning than the more mature members of our species? This is a mystery that needs to be solved. It may be the case that verbal tasks are more complex and challenging; it may be the case that the infant tasks are simpler and closer to the scenarios that we encounter in real life. Here we explore the hypothesis that infants are better statisticians than adults because they lack real world knowledge and their primary tool for early learning is by being sensitive to the probabilistic evidence provided by the environment. In contrast, adults are stumbled by the factual knowledge that has accumulated over time and what they know about the real world may hinder their probabilistic reasoning abilities.

We report two series of experiments with infants and adults, using the same visual displays. We test to see which of two possible strategies they might use to make judgments in a probabilistic inference task: 1) Participants could apply the representativeness heuristic, defined by Kahneman & Tversky (1974) as using the relative similarity between the surface features of a
sample and population to make judgments (i.e., small samples should resemble larger populations from which they are drawn). 2) Participants might appropriately adjust the base rate of the population based on newly taught physical rules (e.g., most but all green balls are stuck inside the box and they are not moveable) and estimate the probabilities of two sampling outcomes. These experiments pit probability computations against representativeness by assessing whether infants and adults will make judgments either using probability calculations that require an adjustment to the base rate of a population, or by assessing the similarity in appearance between the population and samples.

In the first set of experiments, participants are taught that a physical constraint applies probabilistically to a set of otherwise identical balls (i.e., in a set of green balls most, but not all of the balls are immobile – they cannot be removed from boxes and containers). The entire set of red balls moves freely. After observing a number of demonstrations depicting this rule, they are shown a large covered box. An experimenter, removes a sample from the box, say 4 green and 1 red balls, places them in a container and then reveals the contents of the large box as consisting of a 3 green to 1 red ratio. On an alternating trial, infants are shown this same series of events, except that the experimenter removes a sample of 4 red and 1 green balls. In this case, if infants are sensitive to the physical constraint, and they integrate this constraint with the distribution of balls in the population box, they should look longer at the 4 green and 1 red balls sample, as this is an improbable event. If infants instead use the representativeness heuristic, they should look longer at the 4 red and 1 green balls sample, as this sample appears less probable, if the physical constraint is not taken into account.

In the second set of experiments, participants are shown a deterministic physical constraint – all green balls are immobile and therefore cannot be removed from boxes. They are then shown the exact same test trials as described in the probabilistic constraint experiments. In this case, if participants make judgments based on the physical constraint and probability, they should find both samples unexpected, as all green balls in the box should be unavailable for sampling. If they instead ignore the physical constraint and use representativeness, they should find the 4 red and 1 green balls sample more surprising, as this will appear less probable based on the box proportions (see the Predictions section for Experiments 1 and 2 for a more detailed explanation of the predictions).

For both sets of experiments, identical displays are shown to adults and infants. For the infants, the dependent measure is the duration of looking in to each of the two sampling outcomes. Adults are asked to rate how unexpected the outcomes are on a 1-7 scale.

### 3.2 Experiment 1a: Can Infants Integrate a Probabilistic Constraint?

#### 3.2.1 Methods

##### 3.2.1.1 Participants

Twelve 11-month-old infants participated in a violation of expectation looking-time paradigm. Infants were 9 males and 3 females with a Mean age of 11.33 months and a Range of
10.6 to 11.99 months. An additional two infants were tested and their data excluded, one for parental interference and one for fussiness.

3.2.1.2 Materials

**Ping-Pong balls.** A total of 168 (84 green and 84 red) balls were used. The green balls had six Velcro strips glued to them (approximately 0.8 cm · 1.5 cm); the red balls had no Velcro.

**Boxes and containers.** One (20 cm · 4.5 cm · 4 cm) Plexiglass container was placed at the front left-hand corner of the stage to display the five-ball samples pulled out of the box during test trials. Two additional Plexiglass containers (28.5 cm · 4.5 cm · 4 cm) were used during Demonstration Phase 1. One container held 8 red balls, the other 8 green. Six green balls were glued to the container (balls in positions 1,3,4,6,7,8).

A small white box (27 cm · 16 cm · 13 cm) was used in Demonstration Phase 2. The top of the box had a cutout for the experimenter to reach into to access the balls. This box contained 14 Ping-Pong balls, 7 green, 7 red. Six of the green balls were glued to the inside surfaces of the box.

A large, 39 cm · 34 cm · 22 cm box was used to display the population outcomes on the familiarization and test trials. This box was identical in design to the box used in Denison & Xu (2010). The box was a rectangular cube divided into three parts with two Plexiglass containers inserted into the front and back of the box, each containing 60 Ping-Pong balls, and a hidden center compartment used to hold the samples to be removed from the box during test trials. When viewed from the front or the back, the box appeared to be one large box filled with Ping-Pong balls. The front and back of the box were covered with curtains that could be lifted to reveal the contents. The “mostly red” side of the box contained 45 red and 15 green balls (red:green = 3:1), the “mostly green” side contained the opposite (red:green = 1:3); all green balls were covered with Velcro strips.

3.2.1.3 Apparatus

For a detailed description of the apparatus see Chapter 2. The experimenter sat behind a small stage to show infants the stimuli. An observer coded looking times using JHAB (R. Casstevens, 2007). Infants sat in a high chair approximately 70 cm from the center of the stage. The parent sat next to the infant facing the opposite direction and was instructed to avoid looking at the stage.

3.2.1.4 Procedure

The experimenter began by calibrating the observer to the infant’s looking behavior. The experimenter oriented the infant to the outer limits of the viewable area of the stage while the observer watched the infant’s eyes on the TV screen.

**Free play phase.** Infants were first shown a box with 4 green and 4 red Ping-Pong balls; 3 of 4 green balls were glued to the box. They were permitted to play with them for about 1 min. The experimenter shook the balls around and encouraged the infants to try to pick up some balls
of each color, giving them the opportunity to notice that the red balls were easily removed from the box and most of the green balls were stuck. The experimenter then went behind the stage.

**Demonstration Phase 1.** The experimenter brought out one of the long transparent containers, in counterbalanced order. With the container holding the red balls, she drew attention to each of the balls, one at a time. She did this by picking up each of the balls while saying, “Look at this one! See this one?” With the container holding the green balls, the experimenter grasped the balls one at a time, moving through the container from left to right. When she grasped one of the stuck balls she said, “Look at this one; it’s stuck!” and attempted to lift it, causing the entire container to move. When she grasped a movable ball, she said, “Look at this! See this?” The demonstration lasted for approximately 90s.

**Familiarization trials.** Each infant received four familiarization trials. On each trial, the experimenter placed the large box on the stage with the front curtain closed. She shook the box back and forth a few times, saying, “What’s in the box?” Then, with one hand, she lifted the front cover of the box and with the other, she simultaneously lowered the backdrop of the stage while saying “Look, [baby’s name], look!” The observer began timing upon hearing the second, “look,” as this was when the population of either mostly green (with Velcro) or mostly red (without Velcro) balls became visible. Once the backdrop was lowered, the experimenter was no longer visible to the infant. The trial ended when the infant looked away for 2 consecutive seconds. The four trials alternated between the mostly green and the mostly red population, with the first population appearing in counterbalanced order. The large box was removed after each trial and the black curtain was lowered between trials. The familiarization phase lasted approximately 4 min.

These trials were included to familiarize infants to the large box and the objects, as well as to the general procedure of the study. In addition, this phase was included to equate the amount of red and green balls that infants see throughout the entire procedure. Showing infants these complementary distributions might also prime them to attend to ratios.

**Demonstration Phase 2.** The experimenter brought the small white box onto the stage. She flipped the box upside down and then right side up, turned the box side to side and shook it, and placed it back on the stage. She reached into the box, her hand visible to the infant through the transparent front window. In counterbalanced order she drew the infant’s attention to balls of each color. For the red balls, she picked up 4 of the balls one at a time and lifted them to the top of the box while saying, “See this one? Look at this!” Then, in randomized order, she grasped 4 green balls, 3 of which were glued to the inside surface of the box, 1 was movable. When she grasped the immobile balls she said, “Look at this one. It’s stuck!” and when she grasped the movable ball, she lifted it just as she did with the red balls. She repeated the entire sequence once more, this time grasping 4 green balls that were all stuck. Following the second sequence, she removed the box from the stage. This phase lasted approximately 90s.

This phase was included so that infants could see again that most green balls were stuck inside the box and did not move even when the box was shaken vigorously.

**Test trials.** Each infant received 4 test trials (see Figure 3.1, for a schematic representation of the test trial procedure). On each trial, the experimenter placed the large box on the stage, with its front curtain closed. She shook the box a few times, closed her eyes, turned
her head away, and reached into the box. She pulled out 3 Ping-Pong balls and placed them into the small Plexiglas container to her right one at a time. She then repeated this action, pulling out 2 more balls. The experimenter lifted the front curtain of the box to reveal the population of 3 green to 1 red balls and lowered the backdrop while saying, “Look, [baby’s name], look!” Just as in the Familiarization trials, the observer began timing upon hearing the second “look”. The trial ended when the infant looked away for 2 consecutive seconds. The samples alternated between 4 green and 1 red balls and 4 red and 1 green balls on each trial, resulting in 2 sets of 2 trials. At the end of each trial, the stage was cleared. Test trials lasted approximately 7 min.

3.2.1.5 Design

The order of the familiarization trials (mostly red first or mostly green first), the order during the demonstration phases (red balls first or green balls first), and the order of the samples on the test trials (4 red, 1 green or 4 green, 1 red first) were counterbalanced across infants.

![Figure 3.1. Schematic representation of test trials.](image)

3.2.1.6 Predictions

See Figure 3.2 for an illustration of the predictions. If infants understand that, although more green balls are present in the population box, the majority of these balls are unavailable for sampling, then they should look longer at the 4 green and 1 red ball sample than the 4 red and 1 green ball sample. That is, seeing 4 green balls and 1 red ball being removed from a box that contains 75% green ball, most of which are physically constrained, should violate infants’ expectations. In mathematical terms, infants saw a 3:1 population (75% green balls, 25% red
According to the demonstrations, on average 79% of the green balls should be stuck in the box: 75% (free play), 75% (Demonstration 1), 86% (Demonstration 2); thus \(45 \times 0.79 = 35\) green balls are stuck, converting the population available for sampling to: 10 green to 15 red balls (2:3). The probability of drawing 4 red and 1 green balls = \(5! / 4! (3/5)^4 (2/5) = 0.259\); the probability of drawing 4 green 1 red balls = \(5! / 4! (2/5)^4 (3/5) = 0.0768\). Out of 100 draws, one is 4 times more likely to draw 4 red and 1 green balls than 4 green and 1 red.

If infants are unable to recognize and integrate this probabilistic physical constraint, the proportion of balls in the box should predict their looking behavior. That is, infants should look longer at the 4 red and 1 green ball sample, given the ratio of Ping-Pong balls in the box. Again, in mathematical terms, if infants do not learn that the majority of the green balls are stuck, the probability of drawing 4 red and 1 green balls = \(5! / 4! (1/4)^4 (3/4) = 0.0146\); the probability of drawing 4 green 1 red balls = \(5! / 4! (3/4)^4 (1/4) = 0.396\). Out of 100 draws, one is 45 times more likely to draw 4 green and 1 red balls.

### Figure 3.2. Illustration of predictions for Experiments 1 and 2.

#### 3.2.2 Results

Preliminary analyses found no effect of gender or order of test trials (expected vs. unexpected first); subsequent analyses collapsed across these variables. A second observer who was unaware of the order of the trials timed 50% of the familiarization and test trials. Inter-observer reliability averaged 94%.

An alpha level of .05 was used for all analyses. An analysis of variance (ANOVA) examined the effects of test trial order (4 red and 1 green ball sample drawn first vs. 4 green and 1 red ball sample drawn first), trial pair (1, 2), and outcome (4 red 1 green or 4 green 1 red balls). There was a main effect of outcome, \(F(1,11) = 9.585, p = .015\), effect size \((\eta^2) = .542\). Infants looked reliably longer at the 4 green and 1 red samples \((M = 13.719s, SD = 7.15)\) than the 4 red and 1 green samples \((M = 8.632s, SD = 5.067)\). There were no other main effects or interactions.
Eleven of twelve infants looked longer at the 4 green and 1 red ball outcomes, Wilcoxon signed-ranks test: $z = 2.748, p = .006$.

### 3.2.3 Discussion

Infants in this experiment were exposed to three instances indicating that most green balls in a set have the property of being immovable from boxes and containers. On test trials, they looked longer at events in which an experimenter drew samples of 4 green and 1 red balls rather than 4 red and 1 green balls from a box with a ratio of 3 green to 1 red balls. This suggests that infants a) comprehended that the majority of green balls (approximately 80%) in the demonstrations were immobile, b) applied this rule to a new box of balls during test trials, c) integrated the rule with the overall distribution of balls in the box by excluding roughly 80% of the green balls from the 3 green to 1 red balls ratio to accurately estimate the ratio of available balls for sampling, and d) judged that sampling 4 green and 1 red balls was relatively unlikely and sampling 4 red and 1 green balls was relatively likely from the remaining population. On the whole, this provides evidence that infants can integrate a probabilistic physical constraint in their statistical inferences. Furthermore, the fact that infants produced this pattern of looking becomes even more impressive given that the perceptual appearance of the test trial outcomes was in direct opposition with the probability of obtaining the two samples. These results indicate that when representativeness and probability are in direct conflict, infants use probability computations to make generalizations from samples to populations.

Although infants’ looking behavior is consistent with the above interpretation, an alternative interpretation of these data exists. It is possible that infants in this experiment looked longer at the 4 green and 1 red balls sample not because they understood and integrated the constraint but because it was less similar to the events they had seen in the demonstration phases than the 4 red and 1 green balls sample. That is, infants may have been surprised to see a relatively large number of green balls, but not surprised to see a relatively large number of red balls being removed from a box on test trials because they had seen very few instances of green balls being removed from boxes during the demonstration phases. The experimenter lifted 8 red balls and only 2 green balls from the container in demonstration phase 1, and she lifted 8 red balls and only 1 green ball in the box on demonstration phase 2 (because the remaining balls did not move when grasped). This design was necessary to indicate to infants that just a small proportion of green balls were movable. However, a stronger design would control for the number of times each color of ball was lifted from the containers on these demonstrations. We address this concern in Experiment 2.

A second minor flaw in the current design is that infants observed two familiarization trials with a population containing a ratio of 3 red to 1 green balls and two familiarization trials with a population containing a ratio of 3 green to 1 red balls. Recall that we showed infants these familiarization trials to familiarize them with the procedure, equate the amount of red and green balls visible to infants until test trials began, and to prime them to attend to distributional information. These familiarization boxes were identical in appearance when the front curtains were closed and the population contents were not visible (in fact, they were actually two sides of the same box). Thus, infants in this experiment may have made the inference based on familiarization trials that when a large box is present on stage, half of the time it will contain mostly red balls and half of the time it will contain mostly green balls. The infants may have
paid little attention to the contents of the box when revealed on test trials and simply assumed that the experimenter was sampling from the mostly red box on at least half of the trials. Although this is unlikely, the concern with this design is that if the infant assumes that the experimenter is sampling from the mostly red box on half of the test trials, the 4 red and 1 green balls sample should be expected under any and all of the following potential belief states of the infant: 1) infants did not understand the rule at all and thought all balls were movable, 2) infants understood the rule but failed to integrate it with the assumed 3 red to 1 green ball ratio of balls in the box, or 3) infants correctly integrated the rule with the assumed 3 red to 1 green ratio of balls. To definitely rule out this unlikely interpretation, in Experiment 1b we removed the mostly red box from familiarizations.

3.3 Experiment 1b: Replication

To address the design concerns from Experiment 1a and to attempt to replicate the findings, we collected data using a modified procedure with a second group of 11-month-old infants in Experiment 1b.

3.3.1 Method

3.3.1.1 Participants

Twenty-four 11-month-old infants participated in this experiment (17 females; Mean age = 11.07 months; Range = 10.5 to 11.6 months). An additional four infants were tested and their data excluded for parental interference (1) and fussiness (3).

3.3.1.2 Materials

Materials were the same as in Experiment 1a.

3.3.1.3 Procedure

The procedure was the same as in Experiment 1a, except for the following changes:

Demonstration Phase 1. The experimenter equated the number of times that infants saw red and green balls being lifted out of the container. To demonstrate the properties of the red balls, she picked up the 8 balls just as she did in Experiment 1a. To demonstrate the properties of the green balls, she grasped the 6 immobile green balls one at a time and then picked up the 2 movable balls one at a time and repeated her actions with the 2 movable balls 4 times each. She then grasped all 6 green immobile balls again to remind the infant that most of the balls were stuck.

Familiarization Trials. Each infant received two familiarization trials, both times observing the mostly green box.
Demonstration Phase 2. Just as in Demonstration Phase 1, the experimenter equated the number of times balls of each color were lifted. She demonstrated the properties of the red balls by lifting 4 red balls as in Experiment 1a. She demonstrated the properties of the green balls by lifting the 1 movable green ball 4 times and grasping 3 of the stuck green balls without being able to lift them.

Test Trials. Test trials were the same as in Experiment 1a.

3.3.1.4 Design and Predictions

These were the same as in Experiment 1a, including the mathematical predictions: If infants comprehend and integrate the rule, they should be more surprised to observe samples of 4 green and 1 red balls than samples of 4 red and 1 green balls and vice versa if they do not comprehend and integrate the rule.

3.3.2 Results

Preliminary analyses found no effect of gender or order of test trials (expected vs. unexpected first); subsequent analyses collapsed across these variables. A second observer timed 50% of the familiarization and test trials and inter-observer reliability averaged 95%.

An ANOVA examined the effects of test trial order (4 red and 1 green ball sample drawn first vs. 4 green and 1 red ball sample drawn first), trial pair (1, 2), and outcome (4 red 1 green or 4 green 1 red balls). There was a main effect of outcome, $F(1,23) = 6.129$, $p = .022$, effect size ($\eta^2$) = .235. Infants looked reliably longer at the 4 green and 1 red samples ($M = 11.628s$, $SD = 5.51$) than the 4 red and 1 green samples ($M = 9.120s$, $SD = 4.15$). There were no other main effects or interactions. Eighteen of twenty-four infants looked longer at the 4 green and 1 red balls outcomes, Wilcoxon signed-ranks test: $z = 2.432$, $p = .015$.

3.3.3 Discussion

As in Experiment 1a, infants looked longer at samples of 4 green and 1 red balls than at samples of 4 red and 1 green balls, indicating that they learned that the majority of green balls get stuck in boxes and integrated this with the 3 green to 1 red ratio in the population box. Infants in this experiment could not have looked longer at this sample simply because they had become familiar with observing more instances of red balls being removed from boxes in the demonstration phases, as the number of times balls were lifted was equated across colors in this experiment. Furthermore, infants in this experiment were led to assume that the box from which the experimenter was removing the sample was filled with a 3 green 1 red ratio of balls. Thus infants were required to integrate the physical constraint rule with the distributional information in the population to succeed at this task.

In the experiment that follows, we ask whether or not adults are capable of making similar inferences.
3.4 Experiment 1c: Can Adults Integrate a Probabilistic Constraint?

3.4.1 Method

3.4.1.1 Participants

Twelve undergraduates at the University of California, Berkeley participated in this experiment for course credit (7 males; Mean age= 21; Range= 18 to 24).

3.4.1.2 Procedure

Participants were shown video clips of the experimental procedure, and from those clips, asked to rate their reactions to various events using a 7-point Likert scale using pen and paper (with 7 being a “Very Expected” event and 1 being a “Very Unexpected” event). At the start of the experiment, students were given limited instructions, following a set script. Participants were informed that the experiment was generally run with infants, and because of this, it may seem repetitive and slow. In addition, they were told that the experiment would be run in two parts, with the first portion intended to teach them something about the behavior of the objects, and the second section intended for them to rate a variety of displays. The experimenter began by letting the participant see the balls from the Free Play Phase of the infant procedure. She then played the Demonstration and Familiarization clips. For the test trial clips, the experimenter paused the video after each sampling event to give the participant time to assign a rating.

3.4.2 Results

An ANOVA examined the effects of trial pair (1, 2), and outcome (4 red, 1 green vs. 4 green, 1 red). There was a marginal main effect of outcome, with participants rating the 4 red and 1 green sample as more unexpected ($M = 3.4, SD = 1.6$) than the sample of 4 green and 1 red ($M = 4.71, SD = 1.23$); $F(1,11) = 3.05, p = .1$, effect size ($\eta^2$) = 0.109.

3.4.3 Discussion

Participants rated the sample of 4 red and 1 green balls as more unexpected than the sample of 4 green and 1 red ball, contrary to our initial predictions. This suggests that they did not fully understand and integrate the physical constraint with their probabilistic inferences, as they rated the more probable event as more unexpected. Adults may have paid more attention to the perceptual features of the balls in the large box compared with the samples, rather than considering the constraints governing the green balls, combined with the distribution. This suggests that in this experiment, which pits probability calculations against representativeness, use of the heuristic won out in guiding adults’ judgments.
In contrast, after participating in the same experiment, infants correctly looked longer at the less probable sample, suggesting that they found the sample of 4 green and 1 red balls more surprising than the sample of 4 red and 1 green balls. Thus, adults and infants performed quite differently when faced with identical experimental materials, with infants outperforming adults. It is surprising that adults, unlike infants, were unable to learn a simple probabilistic physical rule and apply it in subsequent reasoning. One possibility is that adults’ factual knowledge may have hindered them, that is, they realized that the Velcro strips may temporarily render the green balls unmovable but it is just Velcro so enough force would make these balls movable. We return to this possibility in Experiment 3.

Now, we turn our attention to the question of whether or not infants can integrate a deterministic physical constraint in probabilistic inference.

3.5 Experiment 2a: Can Infants Integrate a Deterministic Constraint?

The purpose of Experiment 1 was to test whether infants can integrate a non-deterministic or probabilistic physical constraint in statistical inference. Therefore, it is important to demonstrate that infants will not draw the same conclusions when given evidence that the constraint applies to all balls in the set. In Experiment 2a, we examine infants’ inferences in cases where all green balls in the demonstrations are immobile, making both the 4 red and 1 green balls sample and the 4 green and 1 red balls sample impossible events. We ask whether infants erroneously believe that drawing 1 green ball and 4 red balls is somehow less impossible than drawing 4 green balls and 1 red when given no evidence that green balls can be moved? We predict that infants should look for approximately the same amount of time at each sample, given that removing any green balls should be interpreted as impossible. If, on the other hand, infants look longer at the 4 green and 1 red balls sample than the 4 red and 1 green balls sample, it will suggest that they do not understand what it means for a physical event to be impossible in our experimental setting. This looking pattern would also make the results of Experiments 1a and 1b more difficult to interpret, as it would be unclear as to whether infants in these experiments correctly integrated the probabilistic constraint or instead thought that the constraint applied deterministically but that the experimenter would still be more likely to draw just one green ball rather than four.

3.5.1 Method

The current experiment was identical to Experiment 1b, except that infants were shown that all red balls move and no green balls move.

3.5.1.1 Participants

Twenty-four 11-month-old infants participated in this experiment (16 females; Mean age = 11.36 months; Range = 10.5 months to 11.57 months). An additional four infants were tested and their data excluded, for parental interference (1), fussiness (2) and experimenter error (1).
3.5.1.2 Materials

Materials were the same as in Experiment 1b except that ALL green balls were stuck inside all boxes and containers.

3.5.1.3 Procedure

The procedure was the same as in Experiment 1b, except for the following changes:

Demonstration Phases 1 and 2. Due to the fact that all green balls were immobile, the experimenter picked up red balls as she did in Experiment 1b and grasped the green balls, lifting the entire container or box.

Familiarization and Test Trials. Identical to Experiment 1b.

3.5.1.4 Design and Predictions

The design was the same as Experiment 1b.

See Figure 3.2 for an illustration of the predictions. If infants understand that all green balls have the property of being physically immobile, they should look approximately equally to both samples on test trials, as they should find it just as unlikely to see 1 green and 4 red balls sampled as 4 green and 1 red balls. If infants do not understand the constraint, or assume that one ball could come loose but that it is unlikely for 4 balls to come loose, they should look longer at the 4 green and 1 red balls sample. If infants do not pick up on the constraint at all, they should look longer at the 4 red and 1 green balls sample, as the 4 green and 1 red balls sample is more probable based on the proportion of balls in the box.

3.5.2 Results and Discussion

Preliminary analyses found no effects of gender or order of the test trials. Subsequent analyses collapsed over these variables. A second observer timed 25% of the familiarization and test trials and inter-observer reliability averaged 95%.

An ANOVA examined the effects of test trial order (4 red and 1 green ball sample drawn first vs. 4 green and 1 red ball sample drawn first), trial pair (1, 2), and outcome (4 red 1 green or 4 green 1 red balls). There was no main effect of outcome, $F(1,23) = 0.01, p = .942$, effect size ($\eta^2_p$) = .000. Infants looked equally at the 4 green and 1 red samples ($M = 11.68s, SD = 3.856$) than the 4 red and 1 green samples ($M = 11.74s, SD = 5.121$). Twelve of twenty-four infants looked longer at the 4 green and 1 red balls sample (the unexpected sample in Experiments 1a and 1b), Wilcoxon signed-ranks test: $z = 0.69, p = .490$.

Infants in this experiment looked for approximately equal amounts of time at the test trial outcomes. This suggests that infants found the 4 green and 1 red balls sample and the 4 red and 1 green balls sample equally likely. That is, infants appear to understand that when an event is demonstrated to be physically impossible, as was the case with lifting the green balls that cohered to boxes in our experiment, it should not be violated some of the time (i.e., the rule of
cohesion states that objects do not spontaneously merge together or break apart and this rule applies to the green balls in this experiment).

3.5.2.1 Comparison of Experiments 1b and 2a

Infants in Experiments 1b and 2a observed identical displays and test trials, but they were either taught a probabilistic constraint (Experiment 1b) or a deterministic constraint (Experiment 3a). We conducted an ANOVA to examine whether or not infants looked for reliably different amounts of time on test trials, based on learning these different rules. We found a marginal interaction between Experiment and Outcome, $F(1,46) = 3.073, p = .086$, effect size ($\eta^2_p$) = .063. To examine this interaction, we performed four post hoc comparisons of infants’ looking times to each of the four test trial outcomes:

1) Comparison of the 4 red and 1 green outcome in the Probabilistic Constraint Experiment (expected sample) to the same outcome in the Deterministic Constraint Experiment ($M_{\text{Deterministic}} - M_{\text{Probabilistic}} =$2.56 sec., $t(46) = 2.22, p = 0.031$).

2) Comparison of the 4 red and 1 green outcome in the Probabilistic Constraint Experiment (i.e., expected sample) and the 4 green and 1 red outcome in the Deterministic Constraint Experiment ($M_{\text{Deterministic}} - M_{\text{Probabilistic}} =$2.62 sec., $t(46) = 1.96, p = 0.057$).

3) Comparison of the 4 green and 1 red outcome in the Probabilistic Constraint Experiment (i.e., unexpected sample) and the same outcome in the Deterministic Constraint Experiment ($M_{\text{Deterministic}} - M_{\text{Probabilistic}} =$0.113 sec., $t(46) = 0.08, p = 0.941$).

4) Comparison of the 4 green and 1 red sample in the Probabilistic Constraint Experiment (i.e., unexpected sample) and the 4 red and 1 green sample in the Deterministic Constraint Experiment ($M_{\text{Deterministic}} - M_{\text{Probabilistic}} =$0.113 sec., $t(46) = 0.07, p = 0.944$).

These results indicate that the mean looking times for the 4 green and 1 red balls outcome in the Probabilistic Constraint Experiment (the unexpected outcome), and the 4 red and 1 green balls outcome, and the 4 green and 1 red balls outcomes in the Deterministic Constraint Experiment were not reliably different from one another. The 4 red and 1 green balls outcome for the Probabilistic Constraint Experiment (the expected sample) was different from each of the other three means.

Infants’ differential performance in these two experiments suggests that they were able to quickly learn either a probabilistic or deterministic physical rule, use it to adjust the base rate of a population of balls, and make judgments accordingly. In doing so, they did not simply apply the representativeness heuristic by comparing the appearance of the small sample drawn and the proportions in the larger population in the box. Infants appear to be remarkably sophisticated, and flexible, when making these inferences.

However, when adults were presented with the same displays as infants in the Probabilistic Constraint Experiment (Experiment 1c), they did not appear to be capable of integrating the constraint in probabilistic inference. In Experiment 2b, we test whether or not adults have similar difficulties integrating a (potentially more straightforward) deterministic constraint.
3.6 Experiment 2b: Can Adults Integrate a Deterministic Constraint?

3.6.1 Methods

3.6.1.1 Participants

Sixteen undergraduates participated in Experiment 2b and received class credit for their participation (6 males; Mean Age: 21; Range= 18 to 27).

3.6.1.2 Procedure

As in Experiment 1c, adults were shown video clips of the experimental procedure, this time from Experiment 2a. They were asked to rate their reactions to the events using a 7-point Likert scale (with 7 being a “Very Expected” event and 1 being a “Very Unexpected” event). They were given the same instructions as in Experiment 1c.

3.6.2 Results

An ANOVA examined the effects of trial pair (1, 2), and outcome (4 red, 1 green trial or 4 green, 1 red trial), and found no main effect for outcome $F\left(1,15\right) = 0.55, p = .472$, effect size ($\eta_p^2$) = 0.109. No statistically reliable difference was found between ratings for the 4 green and 1 red balls outcome ($M = 2.75, SD = 1.81$) and the 4 red and 1 green balls outcome ($M = 3.25, SD = 1.52$).

3.6.3 Discussion

This finding suggests that adults understood the immobile properties of the green Ping-Pong balls and interpreted the events as equally unexpected. This is in contrast to the adults in Experiment 1c, who did not appear to comprehend and integrate the probabilistic physical constraint. Based on this set of results, it appears as though infants are more skilled than adults at interpreting and integrating a probabilistic physical constraint to adjust the base rate of a population to make probability computations, but adults do not have this difficulty with a deterministic constraint. Adults may have fallen prey to the representativeness heuristic in the probabilistic case, rating the sample that was more similar in perceptual appearance as more probable, when this sample was in fact less probable.

Given that adults are more cognitively sophisticated than infants, these divergent results are surprising. However, some recent arguments in the literature point to differences in child and adult cognition suggesting that adults are not always the most adaptive learners. For example, Thompson-Schill, Ramscar, and Chrysikou (2009) argue that children’s underdeveloped prefrontal cortex can confer an advantage for learning in some contexts, as it allows them to be more flexible during the learning process. Lucas et al., (in prep) pose a related argument. They
found that adults can be slower to learn causal rules that are infrequently encountered in the environment, whereas children learn these rules more rapidly. This is presumably because children have less preconceived notions about how causal systems tend to work and were thus more sensitive to the data given in the experiment. Therefore, in this case, children’s limited experience with the world may render them more likely to entertain less conventional hypotheses, causing them to learn a rule faster than adults in some scenarios.

In light of these arguments, we consider the hypothesis that adults’ poor performance in Experiment 1c might be attributable to limitations imposed by their prior experiences with Velcro. While both adults and infants saw Ping-Pong balls marked by Velcro, adults may have focused largely on the known cohesive properties of Velcro rather than the demonstrated properties of the set of green balls. Adults may have viewed samples involving 4 green and 1 red balls as likely, given the fact that their daily experiences dictate that Velcro is separable with relative ease, regardless of the fact that the videos clearly demonstrated that the immobile green balls in the experiment were not easily removed. Infants, on the other hand, quickly learned that most of the green balls were immobile, possibly because they do not have any prior knowledge about Velcro.

3.7 Experiment 3: Can Adults Integrate a Probabilistic Physical Constraint when Prior Knowledge is Mitigated?

In order to assuage the effects of prior knowledge on adults’ performance, Experiment 3 was run without using Velcro strips to mark the green Ping-Pong balls. We ask whether or not removing the prior knowledge conflict introduced by the Velcro might allow adults to focus on the rules introduced in this experiment. If the Velcro did not impact adults’ performance in Experiment 1c and they are simply unable to integrate a physical constraint that applies probabilistically to a set of objects, they should perform similarly to participants in Experiment 1c.

3.7.1 Methods

3.7.1.1 Participants

Twenty-four undergraduate students participated in this experiment for course credit (6 males; Mean age= 21; Range= 19 to 26).

3.7.1.2 Materials and Procedure

The procedure was identical to Experiment 1c, except that no balls had Velcro markings.

3.7.2 Results

An ANOVA examined the effects of trial pair (1, 2) and outcome (4 red, 1 green trial or 4 green, 1 red balls trial). There was a main effect of outcome; $F(1,23) = 6.877, p = .015$, effect
size ($\eta^2$) = 0.23. Participants rated the 4 green and 1 red sample as significantly more unexpected ($M = 2.88, SD = 1.85$) when compared with the 4 red and 1 green sample ($M = 4.67, SD = 1.95$).

### 3.7.2.1 Comparison of Experiment 1c and Experiment 3

An ANOVA examined the effects of Experiment (Experiment 1c vs. Experiment 3), trial pair (1, 2), and outcome (4 red 1 green – the expected trial or 4 green 1 red balls – the unexpected trial). There was a significant interaction between outcome (expected vs. unexpected) and Experiments (1c vs. 3); $F(1,34) = 8.092, p = .007$, effect size ($\eta^2$) = 0.192. In Experiment 1c, participants rated the expected outcome as more unexpected ($M = 3.4, SD = 1.6$) and the unexpected outcome as more expected ($M = 4.71, SD = 1.23$). The reverse was true for Experiment 3, as participants rated the expected outcome as more expected ($M = 4.18, SD = 1.85$) than the unexpected outcome ($M = 3.08, SD = 1.94$) (see Figure 3.3).

![Graph to compare adult results for Experiments 1 and 3 test trials](image)

**Figure 3.3.** Comparison of adult performance in Experiment 1c versus Experiment 3.

### 3.7.3 Discussion

In Experiment 3, participants correctly identified the expected sample of 4 red and 1 green Ping-Pong balls as significantly more expected than the sample of 4 green 1 red balls. These results indicate that adults were able to integrate the probabilistic physical constraint in statistical inference.

Adults’ performance in Experiment 3 closely resembles that of infants, but contrasts with the results found in Experiment 1c, where adult participants rated the expected sample of 4 red and 1 green balls as the more unexpected event. Recall that participants observed identical displays in Experiment 1c and Experiment 3, with the exception of the Velcro markings. This difference in ratings between the two experiments suggests that adults’ prior knowledge may have hindered their ability to comprehend and apply the physical rule when making judgments.
3.8 General Discussion

Here we provide the first evidence that preverbal human infants are able to acquire a probabilistic physical constraint on the spot and are able to adjust the base rate of a population to estimate the probabilities of two outcomes. In contrast, human adults are only able to do so when their real world knowledge is removed. Infants may be better statistical learners because they do not always draw on their scant real world knowledge, and their primary tool for learning is via statistical inference. Adults’ competence in statistical reasoning may sometimes be masked by their factual knowledge, which overrides correct statistical thinking (Thompson-Schill et al., 2009).

These results also call into question the dual-process model, where normatively correct statistical reasoning relies on explicit, effortful, and verbal means (Evans, 2003; Kokis et al., 2002; Reyna & Brainerd, 1995; Sloman, 1996). Instead, human infants may begin life with a set of powerful statistical, inferential mechanisms that guide learning, and later on in development, they acquire heuristics, shortcuts, and factual knowledge that sometimes override correct statistical thinking in exchange for expediency and automaticity.
Chapter 4

Single-Event Probabilistic Inference in Infancy

4.1 Introduction

Reasoning under uncertainty pervades nearly every discipline of study – from social and natural sciences such as psychology, economics, biology and physics – to law and medicine (Bell, Raiffa & Tversky, 1988; Gilovich et al., 2002; Koehler & Harvey, 2004; Pauker & Kassirer, 1980). Moreover, reasoning under uncertainty is the bread and butter of everyday life. For example, in a volatile stock market, economists calculate the odds of making a profit by assuming rational economic laws and making educated guesses about how people’s emotions may interfere with their judgments. In medicine, doctors are almost never certain of a patient’s diagnosis; all they have are symptoms that provide the basis of an estimate, e.g., the probability of a patient having a cold or lung cancer.

The current experiments ask where our probabilistic intuitions come from. Do untutored, preverbal infants use probabilities to make predictions that guide their actions, in order to fulfill their own desires and wishes? Traditional developmental theory suggests that children do not become proficient at making inferences on even the most basic problems requiring probability until ages 4 to 7 (Davies, 1965; Piaget & Inhelder, 1975; Schlottmann & Anderson, 1994; Yost et al., 1962; Zhu & Gigerenzer, 2006).

Meanwhile, the last decade has witnessed the steady accumulation of experimental evidence that young infants are quite competent at quantitative reasoning (Feigenson et al., 2004): they can estimate the number of dots in visual-spatial arrays or sound sequences when density, total area and duration are controlled (Lipton & Spelke, 2003; Xu & Spelke, 2000); they can track small numbers of objects over time and space (Feigenson et al., 2002); they recognize the ordinal relationship between numbers (Brannon, 2002); they can detect ratio differences between large numbers of objects (McCrink & Wynn, 2007); and they can even perform simple addition and subtraction (McCrink & Wynn, 2004; Wynn, 1992). These quantitative reasoning abilities, especially large number estimation and sensitivity to ratios provide some prerequisites for understanding probability.

Several recent experiments have explored whether preverbal infants may engage in rudimentary probabilistic reasoning (Denison et al., 2012; Denison & Xu, 2010; Teglas et al., 2007; 2011; Xu & Garcia, 2008). However, in all of these experiments infants could rely on a simple heuristic that compares absolute quantities: if the lottery machine has 3 blue objects and 1 yellow object, it is more probable to get a blue one on a single random draw because 3 is more than 1 (Teglas et al., 2007); if one gumball machine has 40 pink and 10 black candies and the
other has 10 pink and 40 black candies, then it is more probable to get a pink candy from the first gumball machine than the second on a random draw because 40 is more than 10. A true understanding of probability is based on proportions since probability is an intensive quantity (Bryant & Nunes, 2009): if one gumball machine contains 12 pink and many candies of other colors, and the other 12 pink and a few candies of other colors, even though the number of pink gumballs is the same in the two machines, it is the proportion of pink ones in each jar that determines which one is more likely to yield a pink one – your favorite! – when you insert your precious quarter into a machine. Furthermore, one gumball machine may contain fewer pink candies than the other, but depending on the proportion of pink ones relative to the total number of candies in each machine, the first machine may still have a better chance of delivering a pink one when your quarter is inserted.

Here we report three experiments investigating whether preverbal 10- to 12-month-old infants are sensitive to probabilities based on proportions. Our task is designed to answer two other important questions: First, can infants use their estimates of probabilities to guide their prediction and action? Previous work has used the standard looking-time methodology, which cannot address the issue of prediction and action. Second, can infants use their estimates of probabilities to fulfill their own desires and wishes (e.g., getting a highly desirable pink lollipop from a jar)? To our knowledge, the current experiments are the first to ask this question. If the answer is positive, it suggests that probabilistic reasoning in infancy is robust enough to provide a useful tool for navigating the world.

### 4.2 Experiment 1: Quantity Control of Target Objects

In Experiment 1 we ask whether or not infants can predict which of two populations is most likely to yield a lollipop of their preferred color on a single draw when quantity of target items is controlled. Thus, infants were shown one jar, which contained 12 preferred and 4 non-preferred objects (3:1) and a second jar, which contained 12 preferred and 36 non-preferred objects (1:3). If infants were able to use proportions to estimate the probability of getting a preferred lollipop, they should choose the cup that contained one lollipop that was drawn randomly out of the first jar.

### 4.2.1 Methods

#### 4.2.1.1 Participants
Twenty-four infants participated in the experiment (10 females; $M = 11.1$ months; $R = 10.03–12.99$ months). Six infants were excluded due to failure to complete the preference trial (2), or the test trial (3), or parental interference (1). Parents in all three experiments were recruited by phone from the San Francisco Bay Area and infants received a small gift for their participation.

#### 4.2.1.2 Materials
**Lollipops.** Lollipops (132 total) were used as stimuli for the experiment. Half of the lollipop tops were covered in black construction paper and half in pink construction paper. Each pink lollipop had 3 small gold stars on each side.

**Population Jars.** Four cylindrical, glass jars were used to hold the populations of lollipops (approx. 1320 cm$^3$ in volume). Four cylindrical covers made from orange construction paper were used as covers for the jars.

**Cups.** Two opaque cups (10 cm in diameter, 9 cm in height) were used to hold the samples removed from the population jars. Each cup had a white cover with a small hole to allow the sticks to remain visible while the lollipops were in the cups and the covers were closed.

### 4.2.1.3 Procedure and Design

All infants were tested individually in a forced-choice paradigm. Each infant sat with her parent across from an experimenter. Parents were asked to hold their infant in front of them and to refrain from talking, pointing, or influencing their child in any way.

**Preference trial.** The experimenter brought out one pink lollipop and one black lollipop. She drew the infant’s attention to each one. She then drew the infant’s attention to both lollipops at the same time again and said, “Do you want to come pick one?” while placing the lollipops approximately 1 meter apart on the floor, equal distance from the infant. The parent was instructed to let go of the infant. Infants were encouraged to crawl or walk to a lollipop of their choice (“Come get one!”) — the color chosen was considered their preference. The experimenter clapped for the infant and said, “Good job, you found the one you like!”

**Test Trial.** If the infant preferred the pink lollipop on the preference trial, the experimenter presented them with two transparent jars: one was filled with 12 pink and 4 black lollipops (3:1) and the other was filled with 12 pink and 36 black lollipops (1:3). Conversely, if the infant preferred the black lollipop, she saw one jar with 12 black and 4 pink lollipops and a second jar with 12 black and 36 pink lollipops. Thus the number of preferred lollipops was equated across the jars but the probability of drawing a preferred lollipop differed between them.

The general procedure of the test trial was the same for each infant (see Figure 4.1). Half of the infants completed a standard test trial during which the experimenter brought out one set of the large covered glass jars and placed them 1 meter apart on the floor. She also brought out two opaque cups and placed them next to the jars and towards the center of the jars. She began the trial by removing the covers from both jars simultaneously to reveal the populations to the infant. Next, the experimenter drew the infant’s attention to each jar by lifting the jar off the floor, shaking it, and rotating it around, always beginning with the jar on the right side. After placing each jar back on the floor, she simultaneously replaced the covers on the jars, closed her eyes and reached into the jar on the right. She pulled out one lollipop such that the infant could see the stick but could not see the color of the lollipop, as it was occluded by the experimenter’s hand. She placed the lollipop in the cup next to the jar and closed the cover on the cup. She repeated this action with the jar on the left and placed the lollipop in the cup next to the jar. Finally, the experimenter lifted both cups simultaneously and said, “Come choose one!” She
placed them back down and instructed the parent to let go of their infant. Once the infant crawled or walked to a cup the experiment was over.

The other half of the infants completed a test trial identical to the one described above except that when the experimenter removed the lollipops from the jars, instead of placing each lollipop into the cup next to the jar from which the object was removed, she placed them in the opposite cups.

**Design.** Approximately half of the infants saw the pink lollipop on the right during preference trials (N=13) and half the infants (N=11) saw the mostly pink jar on the right during test trials.

---

**Figure 4.1.** Schematic representation of a test trial.

---

5 While infants were making choices on both the preference and test trial, the experimenter looked directly at him/her, not directing her own attention to either lollipop or cup. If the infant was hesitant or looked to the experimenter, she simply smiled, nodded her head up and down and said, “Yeah, you can have one!” For some infants, the experimenter showed the lollipops twice on the preference trial, or moved them closer to the infant.
4.2.2 Results and Discussion

An ANOVA revealed no effects of gender, age, test trial type (standard or switched), whether or not the correct object was on the same on the test trial as it was on the preference trial, or whether or not the infant preferred the pink or black lollipop. Eighteen out of 24 (75%, $SD = .44$) infants selected the correct cup (i.e., the cup containing a sample from the 12:4 population), reliably different from chance, binomial test, $p = .024$ (2-tailed; .95C.I.=56.40, 93.58).

Thus, when absolute quantity of target objects is equated across populations, infants make a judgment based on proportions, suggesting that when they are prevented from making an inference based on quantity, they correctly use proportional information to make a judgment. However, this does not inform us as to how infants would make judgments if absolute quantity and proportional information were in conflict.

4.3 Experiment 2: Pitting Quantity Against Probability

In Experiment 2, quantity and probability are pitted against each other. One jar contained 16 pink and 4 black lollipops (4:1) and the other jar contained 24 pink and 96 black lollipops (1:4). The first jar now has fewer pink lollipops than the second, but a larger proportion of pink ones. Therefore there is a higher probability of yielding a pink one on a single random draw from the first jar than the second.

4.3.1 Methods

4.3.1.1 Participants

Twenty-four 10- to 13-month-old infants participated (17 females; mean age = 11.5 months; range = 10.58–13.24 months). Nine infants were tested and excluded due to failure to complete the preference trial (4), or failure to complete the test trial (5). Four infants were excluded for having a black lollipop preference. We only included infants with a pink preference, as we did not have enough lollipops to construct the additional populations.

4.3.1.2 Materials

Materials were identical to Experiment 1, except that the lollipops were altered to make the pink even more appealing. This was done to potentially induce an even stronger preference for them, and to increase the likelihood that all infants would prefer the same color (pink). We used brighter colored paper, silver stars and glitter and the lollipop for the preference trial had a small LED light in it that the experimenter switched on and off twice while showing infants the lollipops.

The population jars consisted of 16 pink and 4 black lollipops (4:1) versus 24 pink and 96 black lollipops (1:4).
4.3.1.3 Procedure, Design and Predictions

The design and procedure were the same as Experiment 1 except that the population jars consisted of 16 pink and 4 black lollipops (4:1) versus 24 pink and 96 black lollipops (1:4). These proportions were chosen because it has been demonstrated that 10-month-old infants discriminate the difference in 2:3 ratios, thus they should recognize that there are more pink objects in the lower probability population (i.e., they should discriminate the difference between 16 and 24).

If infants rely on a quantity heuristic in this experiment, they should search in the location with a sample from the 24:96 population, as 24 target items is greater than 16 target items. If they instead correctly make judgments based on comparisons of proportions of pink to black lollipops, they should search in the location with a sample from the 16:4 population, as this is more likely to yield a pink lollipop.

4.3.2 Results and Discussion

An ANOVA revealed no effects of age, test trial type (standard or switched), whether or not the correct object was on the same or the opposite side on preference and test trials, or whether or not they preferred the pink or black lollipop. Nineteen out of 24 (79%; SD=.41) infants selected the correct cup (i.e., the cup containing a sample from the 16:4 population), reliably different from chance, binomial test, \( p = .007 (.95C.I.=61.26, 96.73). \)

This result suggests that infants used comparisons of proportional information and not absolute quantity to guide their search. Thus, even when infants could distinguish the two populations by using a simple quantity heuristic, they instead used proportional information to make an accurate probabilistic inference.

4.4 Experiment 3: Controlling Quantity of All Objects

Experiments 1 and 2 controlled for quantity of the preferred lollipops across the two jars, and we showed that infants were able to use the proportions of the preferred lollipop to guide their prediction and action, in order to fulfill their own desires and wishes. However, infants could have used a different quantity heuristic in our task: they may have avoided the jar with more non-preferred lollipops. To rule out this alternative, we tested another group of infants in Experiment 3, where three different colored lollipops were used.

4.4.1 Methods

4.4.1.1 Participants

Twenty-four 10- to 12-month-old infants participated (14 females; mean age = 11.28 months; range = 10.36–12.39 months). Four infants were tested and excluded due to parental reports of non-typical development (2) and fussiness (2).
4.4.1.2 Materials

Materials were identical to Experiment 2, except that an additional set of lollipops was included. These lollipops were covered in green construction paper and were undecorated.

4.4.1.3 Procedure, Design and Predictions

The experimental procedure was the same as Experiments 1 and 2, except for the following changes: Infants were familiarized to one pink, one black and one green lollipop in the reception room prior to the experiment. This familiarization consisted of the experimenter showing the infants the lollipops one at a time for approximately 1 minute.

During the experiment, infants completed a preference trial with one pink and one black lollipop, just as in Experiment 1. Green lollipops were the neutral ones. As in Experiment 1, the populations that infants saw on test trials were dependent on the results of their individual preference trial. If infants preferred the pink lollipop, they were shown one jar containing 8 pink, 12 black and 2 green lollipops and a second jar containing 8 pink, 8 black and 64 green lollipops. If they preferred black, they saw one jar with 8 black, 12 pink and 2 green lollipops and a second jar with 8 black, 8 pink and 64 green lollipops. On the test trial, infants were not permitted to see which color of lollipop was in the cup that they chose, as infants completed a preference post-test immediately following the test trial and we did not want to bias this final trial. This trial was identical to the original preference trial, with the side of the pink and black lollipops counterbalanced across infants. The post-test was added to make sure that infants still had the same preference as in the initial preference trial.

In this design, infants must choose a sample from one of two populations, each with three types of lollipops: preferred, non-preferred, and neutral (e.g., pink, black, and green lollipops). Infants see one population containing 8 preferred, 12 non-preferred and 2 neutral lollipops and a second population containing 8 preferred, 8 non-preferred and 64 neutral lollipops. In this case, if infants are motivated by increasing their likelihood to obtain a preferred lollipop, and they achieve this based on comparisons of proportions, they should choose the one lollipop drawn from the 8:12:2 population, as this population is more likely to yield a preferred lollipop than the 8:8:64 population (8/22 vs. 8/80). If infants are instead motivated by avoiding non-preferred lollipops, either based on comparisons of absolute quantity (8 vs. 12) or proportions (12/22 vs. 8/80) they should choose the lollipop drawn from the 8:8:64 population. Finally, if infants make choices based on comparisons of absolute quantity of preferred lollipops across populations (8 vs. 8), they should perform at chance. Thus, infants should only make the inference to search in the location containing a sample from the 8:12:2 location if they are a) motivated by increasing the probability of obtaining a preferred lollipop and b) reasoning based on comparisons of proportions of preferred lollipops to all other lollipops across the populations (8/22 = .36 vs. 8/80 = .10).

4.4.2 Results and Discussion

An ANOVA revealed no effects of gender, age, test trial type (standard or switched), whether or not the correct object was on the same or the opposite side on preference and test trials, or whether or not they preferred the pink or black lollipop. On the test trial, 18 out of 24
(75%, SD=.44) infants selected the correct cup, i.e., the cup containing a sample from the 8:12:2 population, reliably different from chance, binomial test, \( p=.024 \) (2-tailed; .95C.I.=56.40, 93.58). Furthermore, 21 out of 23 infants maintained a consistent preference between the preference trial and preference post-test, binomial test, \( p < .001 \). These results suggest that infants rely on proportional information and not comparisons of the absolute number of either preferred or non-preferred objects to make probabilistic inferences.

### 4.4.2.1 Results of Experiments 1 – 3

We computed a variety of additional statistics, collapsed across all three experiments, to reveal whether or not any additional factors impacted infants’ performance (see Table 4.1). Infants chose to search in the correct cup at greater than chance levels regardless of the sampling procedure (switch or no switch), gender, age, consistency of side (whether the correct cup on Test was located on the same or opposite side that the infant crawled to in the Preference trial), or the infant’s color preference (black or pink).

<table>
<thead>
<tr>
<th>Factor</th>
<th># of infants correct</th>
<th>Percent correct</th>
<th>Binomial probability</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sampling procedure</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Switch (( n=36 ))</td>
<td>29</td>
<td>81%</td>
<td>( p &lt; 0.001 )</td>
</tr>
<tr>
<td>Switch (( n=36 ))</td>
<td>26</td>
<td>72%</td>
<td>( p = 0.011 )</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male (( n=31 ))</td>
<td>24</td>
<td>77%</td>
<td>( p = 0.004 )</td>
</tr>
<tr>
<td>Female (( n=41 ))</td>
<td>31</td>
<td>76%</td>
<td>( p = 0.002 )</td>
</tr>
<tr>
<td><strong>Age split</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Younger (( n=36, M=10.76mo.s ))</td>
<td>29</td>
<td>81%</td>
<td>( p &lt; 0.001 )</td>
</tr>
<tr>
<td>Older (( n=36, M=11.81mo.s ))</td>
<td>26</td>
<td>72%</td>
<td>( p = .011 )</td>
</tr>
<tr>
<td><strong>Consistent side</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same side (( n=35 ))</td>
<td>25</td>
<td>71%</td>
<td>( p = 0.018 )</td>
</tr>
<tr>
<td>Different side (( n=37 ))</td>
<td>30</td>
<td>81%</td>
<td>( p &lt; 0.001 )</td>
</tr>
<tr>
<td><strong>Color preference</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pink preference (( n=57 ))</td>
<td>43</td>
<td>75%</td>
<td>( p &lt; 0.001 )</td>
</tr>
<tr>
<td>Black preference (( n=15 ))</td>
<td>12</td>
<td>80%</td>
<td>( p = 0.038 )</td>
</tr>
</tbody>
</table>

### 4.5 General Discussion

These findings provide strong evidence that preverbal infants are capable of rudimentary probabilistic inference – infants as young as ten months use proportions to predict the outcome.
of a single draw, reflecting a true understanding of probability. Furthermore, unlike previous studies that used the standard looking-time methodology, where it was ambiguous whether infants could make predictions or just postdictions after being given possible outcomes (Aslin, 2007), our task provides clear evidence that the format of infants’ probabilistic computations is strong enough to support prediction and action. We also provide the first evidence from preverbal infants that they can use their sensitivity to probability to fulfill their desires and wishes, making it a useful tool for navigating the world. In our case, infants were able to estimate the probability of getting a preferred lollipop and fulfill their own desires by choosing to go to the correct location to find it. To our knowledge, this is the first demonstration of this sort with preverbal infants.

These findings are important for understanding the origins of probabilistic reasoning and probabilistic inference. Our empirical results indicate that probabilistic inference may be an innate learning mechanism that is the foundation for later learning, and such domain-general learning mechanisms may give rise to domain-specific knowledge (Tenenbaum et al., 2011; Xu & Griffiths, 2011). They also add to a growing body of research suggesting that human reasoning may not be as irrational as once thought (Griffiths et al., 2010; Tversky & Kahneman, 1974); in fact, it may be the case that human learners, along with other animals (Behrend & Bitterman, 1961) have an innate sensitivity for probabilities and the capacity for probabilistic inference. As some have argued, use of heuristics may be a later-developing phenomenon – the accumulation of factual knowledge may be the source of these heuristics and they may indeed provide useful shortcuts in real life situations (Kokis et al., 2002). Finally, the recent surge of Bayesian models of human cognition suggests that the probabilistic inference framework captures many key components of human reasoning in a variety of domains throughout the lifespan (Griffiths et al., 2010; Tenenbaum et al., 2011; Xu & Griffiths, 2011; Chater & Oaksford, 2008). These ideal observer models aim to characterize human reasoning at the abstract computational level. Our studies provide direct evidence for probabilistic inference in preverbal human infants, suggesting that probabilistic models may characterize human reasoning at the algorithmic and neural levels as well (Marr, 1982).

In studies on absolute number, researchers are very careful to distinguish whether infants or non-human animals use discrete (e.g., number of elements in a visual-spatial array or number of sounds in a sequence) or continuous quantities (e.g., the total area covered by all the elements in a visual-spatial array or the total duration of a sound sequence) in their computations. This, of course, is because only evidence for the former would constitute evidence for representations of number (Brannon, Abbott & Lutz, 2004; Lipton & Spelke, 2003; Xu & Spelke, 2000).

Estimating probabilities is interestingly different: one can estimate proportions using either discrete or continuous quantities. For example, if there are about 4 pink lollipops in a jar of about 20 total, then the proportion of pink ones is 0.2; thus if I make a single random draw from the jar, then the probability of drawing a pink one is also 0.2. Similarly, if the length of a straight line is about the length of my hand and a pink portion of the line is about the length of my middle finger, then the proportion of the pink segment is about 0.4; thus if I drop a small object on the line, the probability of it landing on the pink segment is .04. In our experiments, we have not teased apart whether infants computed proportions based on discrete or continuous variables, since all the lollipops were the same size. However, this does not detract from our finding that infants were sensitive to proportions and they used this information to estimate the probability of getting a desirable lollipop from the jar. Future studies may investigate whether infants are able to compute probabilities using both discrete and continuous variables.
Our work also has implications for mathematics education. Much research in education shows that proportional reasoning and the probability concept are very difficult for 9- and 10-year-old school-age children (Shaunnessy, 2003; 2007). For example, students think that 1/75 is larger than 1/60 because 75 is bigger than 60. Now imagine putting the students in a task like ours: they see 1 pink ball in a jar of 75 balls or 1 pink ball in a jar of 60 balls. The question is which one has a greater probability of delivering a pink ball if I shake the jar and one ball falls out. Students should succeed on this type of task easily and perhaps this intuitive understanding of probability can be translated into understanding the notation of fraction as one integer divided by another integer. Recent studies with human adults and non-human animals provide some support for this line of thinking: the brain may encode proportions holistically (using both discrete and continuous variables) and de-emphasizing whole numbers may be a way to improve students’ understanding of proportions, fractions, and probability (Jacob, Vallentin & Nieder, 2012).

The overall picture of early quantitative development has changed drastically in the last decade. Here we show that sensitivity to probability and the ability to perform probabilistic inference are well within the repertoire of preverbal infants. These early competences lay the foundation for later development of mathematical thinking and scientific reasoning. As Laplace put it, correctly and insightfully, “The theory of probabilities is at bottom nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which oftentimes they are unable to account.” (Laplace, 1814)
Chapter 5

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

5.1 Introduction

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

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

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

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

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

5.1.1 Belief Revision and Sampling

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

One strategy for implementing Bayesian inference is Monte Carlo approximation, which is based on the idea of sampling from a probability distribution. Using sophisticated Monte Carlo algorithms, it is possible to generate samples from the posterior distribution without having to evaluate all of the hypotheses assigned probability by that distribution (Robert, & Casella, 1999). Following this approach, people might be approximating Bayesian inference by evaluating a small sample of the many possible hypotheses that could account for observed data. Formally, this sample should be drawn from the posterior distribution, $p(h|d)$, which indicates the degree of belief assigned to each hypothesis $h$ given the observed data $d$. Recent work has
shown how Monte Carlo methods that approximate this posterior distribution can account for human behavior in a range of tasks (Levy, Reali, & Griffiths, 2009; Sanborn, Griffiths, & Navarro, 2010; Shi, Feldman, & Griffiths, 2008). Other results suggest that people might be basing their decisions on just a few samples from appropriate probability distributions (Goodman 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 a hypothesis from a distribution necessarily involves a degree of randomness. However, the process is not entirely random in the conventional sense of giving equal probability to each alternative, as when we flip a coin or roll a die, since hypotheses with high probability under the distribution will be sampled more often than those with lower probability. This strategy allows the learner to entertain a variety of hypotheses and in the long run, ensures that they will give more consideration to likely hypotheses but will not overlook a lower probability hypothesis that could turn out to be correct. The Sampling Hypothesis thus suggests that at least some of the variability that appears in children’s responses should be systematic—determined by the posterior distribution over hypotheses.

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

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

### 5.1.2 The wisdom of crowds

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

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

5.1.3 Probability matching

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

Probability matching has also been shown in children in similar reinforcement paradigms. For example, if there are two levers, one that generates a reward when depressed 70% of the time and another that generates the reward 30% of the time, young children learn (over a series of 100 trials) to favor the lever which generates the reward more frequently. However, young preschoolers (i.e., three-year-olds) actually tend more towards maximization when making probabilistic inferences, while four- and five-year-olds, like non-human animals, show probability matching in reinforcement learning (e.g. Jones & Liverant, 1960).
There has been much less work exploring probability matching beyond simple reinforcement learning. Will children probability match when they are formulating hypotheses rather than simply learning reinforced responses? In language learning paradigms, when children are inferring more abstract linguistic hypotheses, they do not probability match but rather maximize, in fact they trend more towards maximizing than adults do (Hudson Kam & Newport, 2005, 2009). In the case of causal inference, there are some suggestive results in which the variability of children’s guesses does seem to be related to the probability of different hypotheses (e.g., Bonawitz & Lombrozo, 2012; Kushnir & Gopnik, 2007; Kushnir, Wellman, & Gelman, 2008; Sobel et al., 2004). However, this possibility has not been systematically tested—these patterns of responding may reflect matching, or they may reflect a noisy maximization process.

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

5.1.4 Testing the predictions of the Sampling Hypothesis

Our experiments test the predictions of the Sampling Hypothesis using a causal learning task that does not involve reinforcement. In particular, children in our task had to learn about the probability of different hypotheses by considering the distribution of different colored blocks in a bag. When a bag has twice as many red blocks in it as blue ones, it is twice as likely that a random block that falls out of the bag will be red rather than blue. Other studies show that even infants are sensitive to this sort of distributional information and can use it to make probability judgments (Teglas et al., 2007; Denison & Xu, 2010b; Xu & Garcia, 2008). This technique also allows us to fine-tune the probability of different hypotheses quite precisely by manipulating the number of blocks in the bag, and it means that children are never differentially reinforced for their responses. Instead, the children had to use the distribution information to inform their guesses about which block had fallen from the bag and caused an effect.

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

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

5.2.1 Methods

5.2.1.1 Participants

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

5.2.1.2 Stimuli

A large box (12 in. (30.48 cm) × 12 in. (30.48 cm) × 18 in. (45.72 cm)) constructed out of cardboard and covered in yellow felt was used. A toy consisting of a transparent sphere connected to a cylindrical shaft was inserted in a hole in the top of the box on the front right corner such that only the sphere (which had a spinner and lights) was visible to the children. The toy was activated by pressing a button on the shaft, causing the sphere portion to light up and play music. An opaque activator bin, made of a plastic container and construction paper, was placed on the back left corner of the box. Additional stimuli included red, blue, and green domino sized wooden blocks; a rigid green bag; and a transparent container (see Figure 5.1).
5.2.1.3 Procedure

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

In both the long wait and short wait conditions, the experimental session began with the child and experimenter sitting across from one another at a table with the large yellow box in between them—the front side facing the child and the back side facing the experimenter. The experimenter introduced children to the large yellow box saying, “This is my big toy and I’m going to show you how it works.” The experimenter then took two blocks of each color (red, blue, and green) and placed them on the table. One block at a time, the experimenter picked up a block of each of the three colors and dropped it into the activator bin. She showed the children that when a red block or a blue block is placed in the activator bin, the toy lights up and plays music, and when a green block is placed in the bin, the toy does not activate. In reality, the experimenter was surreptitiously activating the toy by pressing a button hidden from view. Previous work using this causal scenario suggests that children (and even adults) find this manipulation compelling and that use of the ineffective green block helps to establish that the red and blue blocks did cause the effect (Bonawitz & Lombrozo, 2012).
In a comprehension check, children were asked whether each of the three colors would make the machine go. The experimenter picked up a block and asked, “What will happen if I put a [red, blue, green] block into the machine?” In order to be included in analyses, children had to remember that red and blue blocks make the toy go and green blocks do not. Order of colors was randomized across children for the initial demonstrations and the comprehension check, except that the green block was never demonstrated first in the initial demonstrations.

On Test Trial 1, the experimenter and child counted out 20 red blocks and 5 blue blocks (i.e., an 80:20 distribution) one at a time and placed them into a transparent container. Which block color was counted first was counterbalanced across children. After counting the blocks, the experimenter asked, in the same order as she counted, “So how many red ones did we count? And how many blue ones?” and corrected the child if they were incorrect. Then she shook the blocks in the container to mix them and poured them into the rigid opaque bag. She placed the container upside down in front of the activator bin on the yellow box and placed the bag on top

**Figure 5.1.** Stimuli and procedure used for testing the Sampling Hypothesis in children.
of the container. She then ‘accidentally’ knocked the bag over toward the activator bin. Just after the bag fell over, the experimenter activated the toy and said, “Oh, I think one of the blocks must have fallen into the toy and made it go! Can you tell me which color it was?” Once the child answered the question, the experimenter pretended to remove the block while turning off the toy. Finally she asked, “And why do you think it was a [red, blue] block?” Occasionally children initially responded “both” when asked which color fell in. The experimenter would then prompt the child by saying, “The toy only works when just one block falls in. What color do you think it was?”

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

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

5.2.2 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 for 75% of the trials. All responses uniquely and unambiguously were either “red” or “blue” and agreement was 100%. There was no effect of gender or which color was counted first in either of the two conditions; we collapsed across these variables for subsequent analyses.

5.2.2.1 Probability matching on initial trial

As should be expected, there were no differences between conditions for children’s first predictions, \( \chi^2(1, N = 40) = 1.9, p = ns \). To assess whether or not children probability matched, we averaged the first response of children in both the long wait and short wait conditions. Overall, children’s responses reflected probability matching (28/40 trials, 70% providing the more probable chip response and 30% providing the less probable chip response). Though there was some noise not accounted for by probability matching, the children were not simply randomly guessing, as responses were significantly different from chance (binomial test, \( p < .05 \)) but not significantly different from the predicted distribution of .8 (binomial test, \( p = ns \)). Similarly, children appear not to have “maximized” by always providing the most probable response (i.e. always choosing the red block), or responses would have approached ceiling.

5.2.2.2 Probability matching across all trials

This result suggested that there was probability matching across children – a kind of “wisdom of crowds” effect. Was there evidence of probability matching within individual
children’s responses, as in Vul and Pashler (2008)? We first computed the predictions of independent sampling; that is, given probability $\theta$ of sampling a particular block, what should the distribution of three responses look like? Because there are two possible responses (red (r) or blue (b)) and there are three trials, there are simply $2 \times 2 \times 2$ or 8 possible hypotheses (rrr, rrb, rbr, rbb, ..., bbb). Thus, assuming independence between trials, the probability of any particular hypothesis (e.g., rrb) is simply the probability of sampling each block (i.e. $(.8) \times (.8) \times (.2)$). In this way, we can compute probabilities for all eight hypotheses. We compared this expected distribution to the observed distribution given by children in the short wait and long wait conditions (see Table 5.1). Both the long wait condition and short wait condition were significantly different from the expected distribution (long wait: $\chi^2 (7, N = 20) = 33.91, p < .01$; short wait: $\chi^2 (7, N = 20) = 77.75, p < .0001$). However, there was also a significant difference between children’s responses in the short wait condition and the long wait condition, $\chi^2 (7, N = 40) = 22.3, p < .05$, suggesting that the manipulation had an effect on children’s pattern of responding.

Closer examination showed that the difference from the expected distribution in the long wait condition was due to the extremely low predicted probability of a “blue, blue, blue” response, which was nevertheless chosen by two children. Indeed, when analyses exclude the two children choosing all blue in the long wait condition, the pattern of the remaining 18 children’s responses is not significantly different from the predicted distribution, $\chi^2 (6, N = 18) = 12.7, p = .09$. In contrast, the children’s responses in the short wait condition were much further removed from the expected distribution. By far the most frequent response was for children to alternate responses across trials in spite of the relatively low probability of that hypothesis.

Table 5.1. Pattern of responses expected under independent sampling compared with frequencies in the long wait and short wait conditions.

<table>
<thead>
<tr>
<th>Responses</th>
<th>Expectation</th>
<th>Long Wait</th>
<th>Short Wait</th>
</tr>
</thead>
<tbody>
<tr>
<td>red,red,red</td>
<td>.512</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>red,red,blue</td>
<td>.128</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>red,blue,red</td>
<td>.128</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>red,blue,blue</td>
<td>.032</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>blue,red,red</td>
<td>.128</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>blue,red,blue</td>
<td>.032</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>blue,blue,red</td>
<td>.032</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>blue,blue,blue</td>
<td>.008</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

5.2.2.3 Dependency Measures

We investigated the dependency of children’s responses in two ways. A quick examination of Table 1 suggests that children in the short wait condition were alternating guesses, a strategy that demonstrates dependencies among those responses. To directly compare the two conditions, we coded children’s responses in terms of whether they repeated a guess (e.g. “red” then “red” again) or alternated (e.g. “red” then “blue”), both patterns that would reflect

---

7 Because the approximation to the $\chi^2$ distribution is unreliable with small cell entries, we computed the null distribution numerically. We generated 10,000 contingency tables with these frequencies, computed $\chi^2$ for each, and then computed p values by examining the quantile of the observed $\chi^2$ value.
dependencies among the responses. Comparing condition by repetition/alternation revealed significant differences both when we coded for repetition/alternation over all three responses, Fisher Exact \((N = 33), p < .0001\), and when we coded for repetition/alternation over two responses, Fisher Exact \((N = 80), p < .0001^8\). Children were more likely to repeat or alternate guesses in the short wait than in the long wait condition.

Another way to think about dependency is to model children’s responses as a Markov process and consider the transition matrix. We computed the empirical frequencies with which children moved from a “red block” response to a “blue block” response, and so forth (see Table 5.2). 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 \((N = 20), p < .05\). These results suggest that although children’s pattern of responses in the long wait condition was close to the predicted distribution, there were still some dependencies between a single child’s guesses. Indeed, this is particularly suggested by the anomalous frequency of the blue, blue, blue responses in the long wait condition, responses that might well have reflected a pattern of dependency even in the long wait condition; that is, these children may simply have repeated the response they made on the previous trial.

### Table 5.2. Experiment 1 transition matrices in the two conditions.

<table>
<thead>
<tr>
<th></th>
<th>Long Wait</th>
<th>Short Wait</th>
</tr>
</thead>
<tbody>
<tr>
<td>Next r</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Next b</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current r</td>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td>Current b</td>
<td>4</td>
<td>18</td>
</tr>
</tbody>
</table>

#### 5.2.3 Discussion

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

The Sampling Hypothesis suggests that in both short and long wait conditions children respond in a way that reflects sampling after each new query, and because responses are sampled close together, there are likely to be greater dependencies between guesses in the short wait condition. These dependencies could arise for a number of reasons; for example, recent research suggests that children’s sensitivity to the knowledgeable and helpfulness of an interviewer can

---

8 We use Fisher Exact tests for consistency due to small sample sizes on some tests.
explain children’s tendency to switch guesses on repeated questioning (Gonzolaz, Shafto, Bonawitz, & Gopnik, 2012). Regardless of the specific factors that cause greater dependency when samples are generated over shorter intervals, the overall response pattern of the preschoolers is consistent with the results of Vul and Pashler (2008) with adults. There is some evidence for a “crowd within” effect and the effect is weaker when there is more dependency between responses.

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

5.3 Experiment 2: Varying Proportions

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

5.3.1.1 Participants

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

5.3.1.2 Stimuli

The stimuli were the same as in Experiment 1.

5.3.1.3 Procedure

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

5.3.2 Results

All children generated a unique and unambiguous response of either “red” or “blue;” an assistant blind to condition and hypotheses coded 40% of the trials in each condition and agreement was 100%. There were no differences in the ages of children across conditions ($F(2,72) = 2.73, p = .07$).\footnote{Although this difference is marginally significant, this is mainly due to the children in the 50:50 condition being marginally older than the other two conditions. Children in the 95:5 condition and the 75:25 condition were no different in age, $t(48) = .03, p = .8$.} There were no effects of gender, which color block (red or blue) was used as the majority color, or which color was counted first in any of the three conditions; we collapsed across these variables for subsequent analyses.
5.3.2.1 Probability matching

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

5.3.2.2 Comparing the probability matching and the noisy-max model

To directly compare the probability matching and the noisy-max strategy, we performed three additional analyses. Recall that the sampling prediction is that the proportion of blocks of a particular color in the sample would have a linear effect on the children’s responses—as the proportion of blue blocks goes up, children should be proportionately more likely to guess that the causal block was blue. In contrast, noisy-max predicts no difference between those groups. We performed a logistic regression to test whether or not assignment to a particular condition significantly increased the log odds ratio of guessing the majority color block. Because the method for the initial predictions (i.e., Trial 1) in the 80:20 condition of Experiment 1 are identical to the 95:5 and 75:25 conditions here, we included these data in our analyses, providing yet another distribution to test. We dummy coded children’s responses into 1’s and 0’s—children received a 1 for guessing the majority color block and a 0 for guessing the minority color block (the 50:50 condition was arbitrarily coded such that the color block children saw first when the distributions were counted was given a score of 1). We entered the data from all four conditions into the model and found that the odds ratio for choosing the majority color block was significant in the 95:5 and 80:20 conditions but not in the 75:25 condition (see Table 5.3 for significance tests for all conditions).

In our second analysis, we conducted a logistic regression with distribution of blocks in the bag (i.e., condition) as an ordered predictor variable (95:5; 80:20; 75:25; 50:50). We scaled the Condition variable to more accurately reflect the magnitude of the differences between conditions: the 95:5 Condition was scaled to log(19); the 80:20 Condition = log(3); the 75:25 Condition = log(2); and the 50:50 Condition = log(1). The regression found evidence for a linear increase in the proportion of choices of the more numerous block based on condition (Wald test: df = 3; $z = 2.99$, $S.E. = .199$, $p = .003$). This is consistent with the hypothesis that the children’s responses are sensitive to the distribution of blocks, with a linear relationship being what we should expect if children are probability matching. The analysis also confirms that this probability matching is imperfect, as the coefficient for the linear model is .595, 95% Confidence Interval = (0.205, 0.985).
Finally, we compared the likelihoods of observing the data under a model of probability matching and a model of the noisy-max account. Both models predict random responding for the 50:50 condition, so we did not include responses for this condition in either model. We did include the responses in the 80:20 condition in Experiment 1. Both models are a mixture of chance responding and a variable $\theta$, mediated by a free parameter $\alpha$. The probability matching model is: $\alpha*\text{chance} + (1-\alpha)*\theta$, where chance given two chips is .5 and $\theta$ reflects the probability of that block by condition (i.e. $\theta = .75$ in the 75:25 condition). Because the noisy max model predicts always selecting the maximally likely hypothesis, $\theta$ is 1 for all conditions such that the noisy max model is simply: $\alpha*\text{chance} + (1-\alpha)$. The single free parameter $\alpha$ can be thought of as the parameter that varies how many children are chance responding and how many children perfectly match to theta. Thus, when $\alpha = 1$ chance responding is predicted, and when $\alpha = 0$ all responses are driven by $\theta$. We selected the $\alpha$ that best fit the data for each model. The $\alpha$ that best accounted for the noisy-max model was .58, indicating that the best fit for this model assumes that more than half the children guess at chance and predictions should always fall at around 71% of the more probable block. The $\alpha$ that best fit the probability matching model was .3, indicating that the majority of children probability match but a few children guess at chance and draw response distributions towards .5. The probability matching model was a better fit to the data (log-likelihood = -46.7) than the noisy-max model (log-likelihood = -48.0); see Figure 5.2.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Odds Estimate</th>
<th>Std. Error</th>
<th>z-value</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>95:5</td>
<td>1.899</td>
<td>.678</td>
<td>2.801</td>
<td>.005</td>
</tr>
<tr>
<td>80:20</td>
<td>1.089</td>
<td>.531</td>
<td>2.052</td>
<td>.04</td>
</tr>
<tr>
<td>75:25</td>
<td>0.647</td>
<td>.574</td>
<td>1.127</td>
<td>.26</td>
</tr>
</tbody>
</table>
5.3.3 Discussion

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

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

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

5.4 Experiment 3: Varying Proportions with Three Alternatives

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

5.4.1 Method

5.4.1.1 Participants

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

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

5.4.1.3 Procedure

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

5.4.2 Results

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

5.4.2.1 Probability matching

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

5.4.2.2 Comparing the probability matching and the noisy-max model

As with Experiment 2, we compared the likelihoods of observing the data under a model of probability matching and a model of the noisy-max and compared the majority color chip choices to the combined minority color chip choices. Whether using the $\alpha$ that best fit the data
for each model in Experiment 2 or choosing a new \( \alpha \) for each model that best fit only the data from Experiment 3, the probability matching model was a better fit to the data (log-likelihood = -65.7 when \( \alpha \) set from Experiment 2 (\( \alpha \) = .30); log-likelihood = -65.4 when \( \alpha \) fit to data (\( \alpha \) = .43)) than the noisy-max model (log-likelihood = -70.1 when \( \alpha \) set from Experiment 2; log-likelihood = -67.3 when \( \alpha \) fit to data (\( \alpha \) = .80)). Note also that while the best value for \( \alpha \) in the noisy max model varies greatly for the results of Experiments 2 and 3 (from .58 in Experiment 2 to .80 in Experiment 3) the best \( \alpha \) in the probability matching model is relatively constant across the two experiments (.30 and .43 respectively) indicating that the probability matching model is a more robust model.

5.4.3 Discussion

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

5.5 Experiment 4: Beyond Frequency Matching

In Experiment 4, we designed a task that directly pits naïve frequency matching against a more sophisticated sampling strategy. The design consists of a set of events in which the more numerous color block was actually less likely to have made the machine go than the less numerous color block. For example, there might be more red blocks overall, but it is more likely that a blue block fell into the machine. We asked whether children correctly reasoned about how a sample could be generated by integrating the distributional information overall with information about the physical separation of the population of objects into two distinct distributions. Previous research using looking-time with infants suggests that they can compute probabilities in situations where overall numerosity and probability conflict, based on a physical constraint on the sampling process (Denison & Xu, 2010; Teglas et al., 2007, 2011).

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

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

5.5.1 Method

5.5.1.1 Participants

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

5.5.1.2 Stimuli

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

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

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

5.5.1.3 Procedure

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

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

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

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

5.5.2 Results

Responses fell unambiguously in one of the two color categories. An assistant blind to hypotheses coded 48% of the trials with 100% agreement. There were no differences in performance based on gender or which color objects were counted first in either condition. In the separate distributions condition, there were no differences in performance between children who completed just the first trial ($N = 10$) or all three trials ($N = 10$), $z = 0, p > .05$. We collapsed across these variables for the remainder of the analyses.

In the separate distributions condition, children chose the correct color (blue, orange, or white—i.e., the overall less numerous color) on 26/38 (68%) of trials. This was not different from the predicted distribution of 65% for the rational sampling strategy predicted by the Sampling Hypothesis (binomial test, $p > .5$), and it is higher than chance (50%) performance (binomial test, $p = .03$) and also higher than the naïve frequency matching prediction of 36% (binomial test, $p < .001$). This suggests that children were in fact able to combine the 50% probability of choosing a particular distribution with the 30% and 100% probability of obtaining the correct colored object within each of these containers (dual color vs. uniform color). In the merged distributions condition, children chose the overall more numerous object color (red, purple, or black) on 24/36 (67%) of trials. This is not different from the predicted distribution of 64% (binomial test, $p > .5$) and is marginally different from chance (50%) (binomial test, $p = .065$). It is also significantly different from children’s choices in the separate distributions condition (24/36 trials vs. 12/38 trials), $t(72) = 3.18, p = .002$.

---

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


5.5.3 Discussion

The results of Experiment 4 suggest that children are using a sophisticated sampling strategy. Preschoolers in the two conditions provided different patterns of responses based on the distributional information and how data were generated. In the separate distributions condition, children integrated their prior knowledge about how the blocks were selected with their knowledge about the frequencies of different colors. In the merged distributions condition, children guessed the more numerous color at a rate equivalent to the expected distribution when summed across the entire population, as in Experiments 1 and 2.

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

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

We cannot say with certainty why four- and five-year-olds in our experiment were able to combine probability in this sophisticated way. One possibility is that the physical separation of the distributions into two sets assisted young children in making accurate inferences in our task. Girotto and Gonzalez used a single distribution, separable only on the basis of object features or categories (e.g., shape), and this may have made their task more challenging for very young children. A second possibility is that use of a causal inference task helped children reveal earlier competence. Evidence from previous experiments suggests that adults are better at making judgments that require probability computations when causal variables are made clearer, as they are less likely to engage in base-rate neglect in these circumstances (Krynski & Tenenbaum, 2007). Additionally, evidence suggests that children perform better in probability tasks when
they are encouraged to use intuitive estimation strategies, rather than reason explicitly about numbers or likelihood (Bonawitz & Lombrzo, 2012; Boyer et al., 2008; Ahl et al., 1992; Jeong et al., 2007). Our paradigm may lead children to think less explicitly about the proportions of objects in the distributions and rely on a more intuitive probability sense by asking them to make a causal inference about an accidentally falling block, rather than asking what item was “mostly likely” to be drawn from a distribution.

5.6 General Discussion

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

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

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

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

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

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

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

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

5.7 Conclusion

We proposed that the Sampling Hypothesis might help to explain some competing findings on children’s hypothesis testing and theory building strategies. If children are, in fact, approximating rational inference by sampling hypotheses as our results suggest, this provides an account of the variability that is often observed in patterns of responding and connects that variability to computational level accounts. More generally, the Sampling Hypothesis also suggests that while children's responses can appear irrational when examined individually, they may actually reflect a rational strategy overall.
Chapter 6

Conclusions

6.1 Conclusions and implications of the empirical work

In Chapter 2 I presented data from the first experiment exploring probabilistic inference in infants younger than 8-months. The findings suggest that 6-month-old but not 4-month-old infants can make generalizations from small samples to larger populations. These results, in combination with evidence from similar experiments with 8- and 12-month-olds (Teglas et al., 2007, 2011; Xu & Garcia, 2008), provide convergent support for early competence in rudimentary probabilistic reasoning in infancy.

The results of the experiments reported in Chapter 3 suggest that, in some cases, infants can outperform adults in making probabilistic inferences. Infants were able to flexibly learn and integrate a physical constraint that applied either probabilistically or deterministically to a set of balls. More specifically, the main focus of Chapter 3 was to assess whether or not infants use either the representativeness heuristic or probability computations to make probabilistic inferences. We found that when representativeness and analytical computations were pitted against one another, infants made judgments consistent with a full analytical computation. A second motivation was to compare infant performance to adults, whom we know often use heuristics when making inferences rather than making full analytical computations. The adults’ performance was less straightforward to interpret than the infants’ – adults did integrate the constraints and compute the probabilities accurately, but only when the opportunity to apply their prior knowledge regarding physical objects was removed.

Additionally, recall that a major motivation for examining probabilistic inference in infancy was to reveal whether or not probabilistic inference is a viable domain-general inductive inference mechanism that can be used to acquire domain knowledge. Thus another motivation for the experiments in Chapter 3 was to uncover whether or not infants’ probability computations are integrated with domain knowledge early on. The findings suggest that when infants make probabilistic computations, they do not do so in a purely bottom-up, data driven way, making automatic inferences about populations based on samples. Rather, this appears to be a top-down process, in which human learners can integrate their substantive domain knowledge in order to influence the output of their probabilistic computations. When faced with a situation in which infants must make generalizations from samples to populations, they make these computations by considering potential constraints placed on the generative process. This integration with domain knowledge suggests that probabilistic inference is a good candidate for use in a domain-general inductive inference mechanism that could help build early-emerging domain knowledge.
In Chapter 4, I addressed the question of whether or not infants can use probabilistic computations to make predictions and guide their actions. Before this work, nearly all of the experiments examining probabilistic inference in infancy employed the VOE looking-time method. One ambiguity in using looking time methods is that it is difficult to assess whether or not infants can make predictions about the likely outcome of events or if they can only judge the events post hoc. A second ambiguity is that the interpretation of looking-time data can be controversial, as lower level interpretations of the findings are often plausible, as opposed to rich interpretations regarding conceptual knowledge (Aslin, 2007). The three experiments in Chapter 4 provide a converging measure of probabilistic inference in infancy, and suggest that infants can use probabilistic computations to guide prediction and action. Furthermore, that series of experiments teased apart one final important confound in all of the previously discussed infant probabilistic inference experiments. The experiments were designed such that infants could not rely on a simple heuristic that compares absolute quantities, as increasingly complex quantity controls were introduced across the three experiments. Together, the results of the three experiments demonstrated that infants can accurately navigate the world by using probabilistic computations based on proportional reasoning to guide their search for desired objects.

Finally, Chapter 5 examined the Sampling Hypothesis in 4- and 5-year-old children. In particular, the purpose of the experiments was to examine how it is that children’s responses in experimental settings show evidence of both variability and rationality. Together, the experiments provide support for the conjecture that at least some of the variability we see in children’s responses is the result of a rational strategy for inductive inference. Children appear to sample hypotheses from the distribution of possible hypotheses, and this sampling behavior brings with it two consequences observed in the experiments. First, as a group, children’s responses reflected probability matching to the posterior. Second, children demonstrated dependencies between their responses, and as expected when undergoing a process of sampling, this dependency decreased as a function of time between guesses. The fourth experiment in the series provided evidence that children were truly engaged in a process of sampling hypotheses from the posterior, and not simply matching the frequencies of the colors observed during the experimental session.

### 6.2 Remaining questions and future directions

The evidence reviewed above in support of probabilistic inference in infants and young children suggests that human learners may have an intuitive notion of probability. This ability may provide the foundation for inductive inference across a variety of domains beginning early in development. In fact, evidence from the preschool-aged children in Chapter 5 suggests that 4- and 5-year-old children can track the probability of up to three hypotheses to inform their guesses about which of a number of objects made a machine activate. Thus, the evidence reviewed here suggests that, beginning at least by 6-months of age, infants possess one of the key pre-requisite abilities for an inductive inference mechanism based on the principles of rational Bayesian inference. Children can later use this ability to reason using probabilistic data to inform their hypotheses in a causal inference task.
6.2.1 Remaining questions and future directions from the infant work

One important question left open in the infant work is whether infants are computing probabilities over discrete, individual objects or over continuous regions of color. As noted in Chapter 4, in studies on numerical reasoning, researchers carefully disentangle whether infants or non-human animals use discrete (e.g., number of elements in a visual-spatial array or number of sounds in a sequence) or continuous quantities (e.g., the total area covered by all the elements in a visual-spatial array or the total duration of a sound sequence) in their computations. Estimating probabilities is quite different from estimating number, as one can estimate proportions using either discrete or continuous quantities. Even though the lollipop experiments cannot disentangle whether infants use discrete or continuous quantities to estimate probabilities, it does not detract from the overall finding that infants can compute probabilities. Additionally, it seems more likely that infants in most of these experiments were reasoning about the proportions of individual objects and not continuous quantities, given much evidence suggesting that young infants are better at reasoning about discrete objects than continuous quantities (e.g., Feigenson et al., 2002; Huntley-Fenner, Carey, & Solimando, 2002; Rosenberg & Carey, 2009, though see Hespos, Ferry, & Rips, 2009). Take for example the experiments in Chapter 3, suggesting that infants can integrate both probabilistic and deterministic physical constraints in probabilistic inference. Much of infants’ knowledge of naïve physics, including the understanding of cohesion, does not translate to continuous arrays like piles of sand (Spelke & Born, 1993). Thus it seems unlikely that infants would be able to integrate such a rule in their probability computations if they were not computing over objects. Regardless, future studies should investigate whether infants are able to compute probabilities using both discrete and continuous variables. In fact, some evidence suggests that older children perform better at proportional reasoning when dealing with continuous rather than discrete quantities (Boyer et al., 2008; Jeong et al., 2007; Spinillo and Bryant, 1999). It is an open question as to whether or not infants would perform better with one stimuli type or the other.

The work on infant probabilistic reasoning reported here has implications for the large body of literature investigating rational inference in human adults, adding to a growing body of research suggesting that human reasoning may not be as irrational as was once thought (Gigerenzer, 2000; Gigerenzer & Gaissmaier, 2011; Griffiths & Tenenbaum, 2006; Tenenbaum et al., 2011). In Tversky and Kahneman’s classic experiments examining adult reasoning, they found that people often use mental shortcuts such as the availability heuristic or the representativeness heuristic to make judgments under uncertainty, and that use of these heuristics often result in biased, incorrect inferences (1974,1981). The findings reported in the infant experiments here suggest that infants do not begin by relying on representativeness or similarity heuristics in making probabilistic inferences in cases where accurate analytical reasoning and heuristic reasoning should produce conflicting results. According to these findings, the use of heuristics may be a later-developing phenomenon – the accumulation of factual knowledge may be the source of these heuristics and they may indeed provide useful shortcuts in real life situations (Kokis et al., 2002). Further exploration of heuristic reasoning in infancy and early childhood is needed to investigate whether or not young learners can rely on heuristics under appropriate circumstances, to increase efficiency when a full analysis of probabilities is not required.

This dissertation presents a strong foundation of evidence suggesting that sophisticated probabilistic inference abilities are present early in infancy. Future work with infants should
continue to focus on a number of additional questions left open from the reported experiments. First, future studies should continue to explore the age at which this ability comes online. Are infants younger than 6-months really incapable of making inferences between samples and populations, or might a different task reveal competence at earlier ages? If infants younger than 6-months cannot compute or estimate probabilities, this will surely have implications for the kind of learning that could take place within the first 6-months, and for how the ability to make probabilistic inferences comes to be.

Second, future work should explore whether or not infants can use sampling information to acquire new knowledge early in infancy in domains such as naïve physics and psychology. Recall that part of the motivation for examining probabilistic inference in infancy was to reveal whether or not probabilistic inference is a viable domain-general inductive inference mechanism that can be used to acquire domain knowledge. The work reported here suggests that the ability to compute probabilities is integrated with substantive domain knowledge early on in infancy. One motivation for conducting the experiments reported in Chapter 3 was to uncover whether or not infants’ probability computations are integrated with domain knowledge early on. The experiments in Chapter 3 revealed that infants can integrate knowledge from the domain of naïve physics with probabilistic inference. If making generalizations from samples to populations is truly a mechanism for learning or a pre-requisite to inductive inference in infancy, the next step is to examine whether or not young infants can use sampling information to learn a new rule in a specific domain. For example, infants should be able to use the information gained from observing a small number of instances of physical events to make generalizations that allow them to establish rules in naïve physics that they have not yet acquired. Recent work suggests that older infants and preschoolers are capable of using sampling information to infer that an agent has a preference for a particular object (Kushnir et al., 2010; Ma & Xu, 2011). It remains to be seen if younger infants can use sampling information to make similar inferences.

Finally, probabilistic inference may be linked to the two known systems used for quantitative reasoning throughout the lifespan. That is, in the experiments that examine probabilistic inference with large populations and multi-object samples (e.g., Xu & Garcia, 2008), approximate numerosities are likely provided by the analog magnitude system to establish ratio information. In the lottery machine experiments, object files, which can be used to track quantity, are likely involved in allowing infants to recognize which of two object types is more numerous in each set, allowing an estimate of which object type is most likely to be sampled on a random draw. If these two systems for reasoning about quantity provide the input to probabilistic computations, predictable limits should be encountered in probabilistic reasoning, mirroring those found in numerical reasoning (see Feigenson et al., 2004). It will be interesting to examine how numerical reasoning interacts with probabilistic inference in future studies.

### 6.2.2 Remaining questions and future directions from the preschool experiments

Our findings with preschoolers raise a number of new and interesting questions. First, as discussed in the chapter, it will be interesting to continue to explore how children generate samples for hypothesis testing; that is, what specific algorithms do learners use to generate samples? Current work suggests one algorithm consistent with children’s and adults’ behavior is the Win-Stay, Lose-Shift strategy. Use of this algorithm requires that the learner only
occasionally resample a hypothesis from the full posterior distribution, which minimizes the amount of computation required but still results in accurate belief revision.

Another future direction will involve investigating the sampling hypothesis in even younger children. The methodology used in Chapter 4 can be modified to examine the Sampling Hypothesis with 10- to 12-month-old infants. Future work will attempt to tease apart whether young infants succeed at the tasks in Chapter 4 by using a maximization strategy or a probability matching strategy. Previous research suggests that younger children (i.e., 3-year-olds) tend to maximize in probability learning experiments using variable reinforcement (Vulkan, 2000). However, no previous research investigates probability matching versus maximizing in children under 2-years of age, and the tasks that revealed maximizing behavior in 3-year-olds are very different from the tasks used in the experiments that explore the Sampling Hypothesis. The strategy that infants use in making probabilistic inferences such as the ones examined in this dissertation is an entirely open question.

Finally, we measured one additional aspect of children’s performance that we did not discuss in the chapter but that warrants discussion here. Following each child’s response to the critical question: “What color block do you think fell into the machine?” the experimenter elicited an explanation by simply asking the child why they guessed that particular color. Of the 247 children who were asked this question, only 14 provided an explanation that seemed reasonable given the data they observed (~5% of children). Recall also that a large number of these children were given multiple opportunities to provide explanations, as children were asked the question three times in two of the experiments. These 14 “sensible” explanations were counted liberally, as any child who produced an explanation that mentioned something about the distribution having more of a particular color was counted as providing a reasonable explanation. Most of these children said something along the lines of: “It was red because red’s the most” or “I think blue because there were lots”. No children appealed to the relative numbers of the colored chip or made any attempts at describing the random nature of the process. The most common explanations, even after an additional prompt of “Can you tell me anything else?” were “Just because” or “I don’t know”. Additionally, children were more likely to give a fanciful response than a reasonable responses such as, “Green because I have a spaceship that plays that green song” or “Red because red is the color of hot lava.” Despite this inability to explain why they guessed a particular color, children guessed that the more probable chip was sampled more often than would be expected by chance in every experiment (except in one condition of Experiment 2, which was only marginally significantly different from chance). This finding is in agreement with other experiments examining probabilistic inference in early childhood. Denison, Konopczynski, Garcia & Xu (2006) elicited explanations from 4-year-olds in an experiment examining rudimentary probabilistic inference, which was similar in design to the 8-month-old experiments in Xu & Garcia (2008). They found that only 1 out of 18 children (again, ~5%) provided an explanation that appealed to quantity. In future work, it will be interesting to examine children’s explanations in these and related tasks more systematically. For example, future experiments could explore the various reasons why children struggle with explanations in these tasks, in contrast to findings that children as young as 3-years of age can produce sensible explanations in tasks of a different nature (e.g., Bartsch & Wellman, 1989 find that 3-year-olds can explain why a character in a story took a particular action). Future experiments could explore which aspects of our tasks make explanation so difficult: Do children struggle specifically with probabilistic explanations? Or, more generally, do they struggle to explain how they come to know particular facts from observed evidence?
6.3 Conclusions

In conclusion, the empirical work reviewed in this dissertation reveals that notions of probability and randomness are available to children far earlier in development than was previously posited. The picture of the young child’s probability concept has changed dramatically over the past 60 years, as researchers have revealed competences in infants and preschool-aged children that were not initially credited to children until they reached the first or second grade.

The broad finding that infants can make generalizations from samples to populations provides important insight into classic debates concerning the initial cognitive state of the human infant and the learning mechanisms that are available to support conceptual change. The empirical results reviewed in this dissertation indicate that probabilistic inference may be an innate learning mechanism that is the foundation for later learning, as illustrated by the causal learning experiments with preschool-aged children. Future work will not only explore further the nature and limits of early probabilistic reasoning, but also how this kind of learning mechanism may be used to construct new concepts and new knowledge.
Bibliography


Lucas, C. G., Gopnik, A., Griffiths, T. L., (In Prep). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. Unpublished manuscript.


